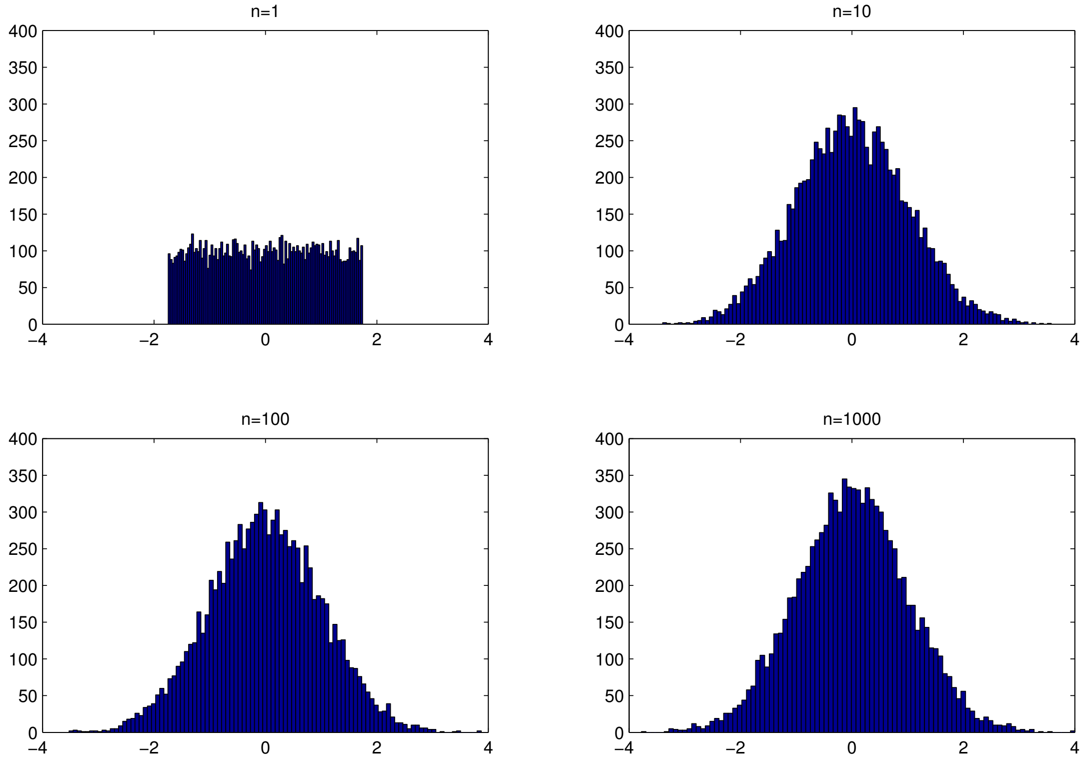
ADVANCED PROBABILITY AND STATISTICAL INFERENCE I

Lecture Notes of BIOS 760



Distribution of Normalized Summation of n i.i.d Uniform Random Variables

PREFACE

These course notes have been revised based on my past teaching experience at the department of Biostatistics in the University of North Carolina in Fall 2004 and Fall 2005. The context includes distribution theory, probability and measure theory, large sample theory, theory of point estimation and efficiency theory. The last chapter specially focuses on maximum likelihood approach. Knowledge of fundamental real analysis and statistical inference will be helpful for reading these notes.

Most parts of the notes are compiled with moderate changes based on two valuable textbooks: *Theory of Point Estimation* (second edition, Lehmann and Casella, 1998) and *A Course in Large Sample Theory* (Ferguson, 2002). Some notes are also borrowed from a similar course taught in the University of Washington, Seattle, by Professor Jon Wellner. The revision has incorporated valuable comments from my colleagues and students sitting in my previous classes. However, there are inevitably numerous errors in the notes and I take all the responsibilities for these errors.

Donglin Zeng
August, 2006

PREFACE TO REVISED VERSION

The revised version involves several minor corrections and additions, but has not been changed much compared to the original.

Michael R. Kosorok
August, 2010

CHAPTER 1 A REVIEW OF DISTRIBUTION THEORY

This chapter reviews some basic concepts of discrete and continuous random variables. Distribution results on algebras and transformations of random variables (vectors) are given. Part of the chapter pays special attention to the properties of Gaussian distributions. The final part of the chapter introduces some commonly-used distribution families.

1.1 Basic Concepts

Random variables are often classified into *discrete random variables* and *continuous random variables*. By name, discrete random variables are variables which take on discrete values with an associated *probability mass function*; while, continuous random variables are variables taking non-discrete values (usually R) with an associated *probability density function.* A probability mass function consists of countable non-negative values with their total sum being one and a probability density function is a non-negative function on the real line with its entire integral being one.

However, the above definitions are not rigorous. What is the precise definition of a random variable? Why shall we distinguish between mass functions or density functions? Can some random variable be both discrete and continuous? The answers to these questions will become clear in next chapter on probability measure theory. However, a brief glimpse is given below:

(a) Random variables are essentially *measurable functions* from a *probability measure space* to a real space. Especially, discrete random variables map into a discrete set and continuous random variables map into the whole real line.

(b) The probability (probability measure) is a function assigning non-negative values to sets of a σ*-field* and it satisfies the property of *countable additivity.*

(c) The probability mass function for a discrete random variable is the *Radon-Nykodym derivative* of a *random variable-induced measure* with respect to a *counting measure.* The probability density function for continuous random variable is the Radon-Nykodym derivative of the random variable-induced measure with respect to the *Lebesgue measure.*

For this chapter, we do not need to worry about these abstract definitions.

Some quantities to describe the distribution of a random variable include *cumulative distribution function*, *mean*, *variance*, *quantile*, *mode*, *moments*, *centralized moments*, *kurtosis* and *skewness.* For instance, if X is a discrete random variable taking values $x_1, x_2, ...$ with probabilities $m_1, m_2, ...$ The cumulative distribution function of X is defined as $F_X(x) = \sum_{x_i \leq x} m_i$. The

kth moment of X is given as $E[X^k] = \sum_i m_i x_i^k$ and the kth centralized moment of X is given as $E[(X-\mu)^k]$ where μ is the expectation of X. If X is a continuous random variable with probability density function $f_X(x)$, then the cumulative distribution function $F_X(x) = \int_{-\infty}^x f_X(t)dt$ and the kth moment of X is given as $E[X^k] = \int_{-\infty}^{\infty} x^k f_X(x)dx$ if the integration is finite. The skewness of X is given by $E[(X-\mu)^3]/Var(X)^{3/2}$ and the kurtosis of X is given by $E[(X-\mu)^4]/Var(X)^2$. The last two quantities describe the shape of the density function: negative values for the skewness indicate distributions that are skewed to the left and positive values indicate distributions skewed to the right. By skewed to the left, we mean that the left tail is heavier than the right tail. Similarly, skewed to the right means that the right tail is heavier than the left tail. A large kurtosis indicates a "peaked" distribution and a small kurtosis indicates a "flat" distribution. Note that we have already used $E[g(X)]$ to denote the expectation of $g(X)$. Sometimes, we use $\int g(x)dF_X(x)$ to represent this whether or not X is continuous or discrete. This notation will be clear after we introduce probability measures.

Next we review an important quantity in distribution theory, namely the *characteristic function* of X. By definition, the characteristic function for X is defined as $\phi_X(t) = E[\exp\{itX\}] = \int \exp\{itx\}dF_X(x)$, where i is the imaginary unit, the square-root of -1. Equivalently, $\phi_X(t)$ is equal to $\int \exp\{itx\}f_X(x)dx$ for continuous X and is $\sum_j m_j \exp\{itx_j\}$ for discrete X. The characteristic function is important since it uniquely determines the distribution function of X, the fact implied in the following theorem:

Theorem 1.1 (Uniqueness Theorem) If a random variable X with distribution function F_X has a characteristic function $\phi_X(t)$ and if a and b are continuous points of F_X, then

$$F_X(b) - F_X(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t)dt.$$

Moreover, if F_X has a density function f_X (for continuous random variable X) , then

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t)dt.$$

†

We defer the proof to Chapter 3. Similar to the characteristic function, we can define the *moment generating function* for X as $M_X(t) = E[\exp\{tX\}]$. However, we note that $M_X(t)$ may not exist for some t but $\phi_X(t)$ always exists.

Another important and distinct aspect in distribution theory is the independence of two random variables. For two random variables X and Y, we say X and Y are *independent* if $P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$; i.e., the joint distribution function of (X,Y) is the product of the two marginal distributions. If (X,Y) has a joint density, then an equivalent definition is that the joint density of (X,Y) is the product of two marginal densities. Independence introduces many useful properties, among which one important property is that $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for any sensible functions g and h. In the more general case when X and Y may not be independent, we can calculate the *conditional density* of X given Y, denoted by $f_{X|Y}(x|y)$, as the ratio between the joint density of (X,Y) and the marginal density of Y. Thus, the conditional expectation of X given $Y = y$ is equal to

$E[X|Y=y]=\int x f_{X|Y}(x|y)dx$. Clearly, when X and Y are independent, $f_{X|Y}(x|y)=f_X(x)$ and $E[X|Y=y]=E[X]$. For conditional expectations, two formulae are useful:

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

So far, we have reviewed some basic concepts for a single random variable. All the above definitions can be generalized to a multivariate random vector $X=(X_1,...,X_k)'$ with a joint probability mass function or a joint density function. For example, we can define the mean vector of X as $E[X]=(E[X_1],...,E[X_k])'$ and define the covariance matrix for X as $E[XX']-E[X]E[X]'$. The cumulative distribution function for X is a k-variate function $F_X(x_1,...,x_k)=P(X_1\le x_1,...,X_k\le x_k)$ and the characteristic function of X is a k-variate function, defined as

$$\phi_X(t_1,...,t_k)=E[e^{i(t_1X_1+...+t_kX_k)}]=\int_{R^k} e^{i(t_1x_1+...+t_kx_k)}dF_X(x_1,...,x_k).$$

Similar to Theorem 1.1, an inversion formula holds: Let $A=\{(x_1,..,x_k): a_1<x_1\le b_1,\dots,a_k<x_k\le b_k\}$ be a rectangle in R^k and assume $P(X\in\partial A)=0$, where ∂A is the boundary of A. Then

$$F_X(b_1,...,b_k)-F_X(a_1,...,a_k)=P(X\in A)$$

$$=\lim_{T\to\infty}\frac{1}{(2\pi)^k}\int_{-T}^{T}\cdots\int_{-T}^{T}\prod_{j=1}^{k}\frac{e^{-it_ja_j}-e^{-it_jb_j}}{it_j}\phi_X(t_1,...,t_k)dt_1\cdots dt_k.$$

Finally, we can define the conditional density, the conditional expectation, and independence of two random vectors similarly to the univariate case.

1.2 Examples of Special Distributions

We list some commonly-used distributions in the following examples.

Example 1.1 Bernoulli Distribution and Binomial Distribution A random variable X is said to be Bernoulli(p) if $P(X=1)=p=1-P(X=0)$. If $X_1,...,X_n$ are independent, identically distributed (i.i.d) Bernoulli(p), then $S_n=X_1+...+X_n$ has a binomial distribution, denoted by $S_n\sim Binomial(n,p)$, with

$$P(S_n=k)=\binom{n}{k}p^k(1-p)^{n-k}.$$

The mean of S_n is equal to np and the variance of S_n is equal to $np(1-p)$. The characteristic function for S_n is given by

$$E[e^{itS_n}]=(1-p+pe^{it})^n.$$

Clearly, if $S_1\sim Binomial(n_1,p)$ and $S_2\sim Binomial(n_2,p)$ and S_1,S_2 are independent, then $S_1+S_2\sim Binomial(n_1+n_2,p)$.

Example 1.2 Geometric Distribution and Negative Binomial Distribution Let $X_1,X_2,...$ be i.i.d Bernoulli(p). Define $W_1=\min\{n: X_1+...+X_n=1\}$. Then it is easy to see

$$P(W_1=k)=(1-p)^{k-1}p,\quad k=1,2,...$$

We say W_1 has a geometric distribution: $W_1 \sim Geometric(p)$. To be general, define $W_m = \min\{n : X_1 + ... + X_n = m\}$ to be the first time that m successes are obtained. Then

$$P(W_m = k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, ...$$

W_m is said to have negative binomial distribution: $W_m \sim$ Negative Binomial(m, p). The mean of W_m is equal to m/p and the variance of W_m is $m/p^2 - m/p$. If $Z_1 \sim$ Negative Binomial(m_1, p) and $Z_2 \sim$ Negative Binomial(m_2, p) and Z_1, Z_2 are independent, then

$$Z_1 + Z_2 \sim \text{ Negative Binomial}(m_1 + m_2, p).$$

Example 1.3 Hypergeometric Distribution A hypergeometric distribution can be obtained using the following urn model: suppose that an urn contains N balls with M bearing the number 1 and $N - M$ bearing the number 0. We randomly draw a ball and denote its number as X_1. Clearly, $X_1 \sim Bernoulli(p)$ where $p = M/N$. Now replace the ball back in the urn and randomly draw a second ball with number X_2 and so forth. Let $S_n = X_1 + ... + X_n$ be the sum of all the numbers in n draws. Clearly, $S_n \sim Binomial(n, p)$. However, if each time we draw a ball but do not replace it back, then $X_1, ..., X_n$ are dependent random variable. It is known that S_n has a hypergeometric distribution:

$$P(S_n = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, .., n.$$

Or, we write $S_n \sim Hypergeometric(N, M, n)$.

Example 1.4 Poisson Distribution A random variable X is said to have a Poisson distribution with rate λ, denoted $X \sim Poisson(\lambda)$, if

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, ...$$

It is known that $E[X] = Var(X) = \lambda$ and the characteristic function for X is equal $\exp\{-\lambda(1 - e^{it})\}$. Thus, if $X_1 \sim Poisson(\lambda_1)$ and $X_2 \sim Poisson(\lambda_2)$ are independent, then $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$. It is also straightforward to check that conditional on $X_1 + X_2 = n$, X_1 is Binomial$(n, \lambda_1/(\lambda_1 + \lambda_2))$. In fact, a Poisson distribution can be considered as the summation of a sequence of Bernoulli trials each with small success probability: suppose that $X_{n1}, ..., X_{nn}$ are i.i.d Bernoulli(p_n) and $np_n \to \lambda$. Then $S_n = X_{n1} + ... + X_{nn}$ has a Binomial(n, p_n). We note that for fixed k, when n is large,

$$P(S_n = k) = \frac{n!}{k!(n-k)!} p_n^k (1 - p_n)^{n-k} \to \frac{\lambda^k}{k!} e^{-\lambda}.$$

Example 1.5 Multinomial Distribution Suppose that $\{B_1, ..., B_k\}$ is a partition of R. Let $Y_1, ..., Y_n$ be i.i.d random variables. Let $\underline{X}_i = (X_{i1}, ..., X_{ik}) \equiv (I_{B_1}(Y_i), ..., I_{B_k}(Y_i))$ for $i = 1, ..., n$

and set $\underline{N} = (N_1, ..., N_k) = \sum_{i=1}^n \underline{X_i}$. That is, $N_l, 1 \leq l \leq k$ counts the number of times that $\{Y_1, ..., Y_n\}$ fall into B_l. It is easy to calculate

$$P(N_1 = n_1, ..., N_k = n_k) = \binom{n}{n_1, ..., n_k} p_1^{n_1} \cdots p_k^{n_k}, \quad n_1 + ... + n_k = n,$$

where $p_1 = P(Y_1 \in B_1), ..., p_k = P(Y_1 \in B_k)$. Such a distribution is called the Multinomial distribution, denoted Multinomial$(n, (p_1, .., p_k))$. We note that each N_l is a binomial distribution with mean np_l. Moreover, the covariance matrix for $(N_1, ..., N_k)$ is given by

$$n \begin{pmatrix} p_1(1-p_1) & \dots & -p_1 p_k \\ \vdots & \ddots & \vdots \\ -p_1 p_k & \dots & p_k(1-p_k) \end{pmatrix}.$$

Example 1.6 Uniform Distribution A random variable X has a uniform distribution in an interval $[a, b]$ if X's density function is given by $I_{[a,b]}(x)/(b-a)$, denoted by $X \sim Uniform(a, b)$. Moreover, $E[X] = (a+b)/2$ and $Var(X) = (b-a)^2/12$.

Example 1.7 Normal Distribution The normal distribution is the most commonly used distribution and a random variable X with $N(\mu, \sigma^2)$ has a probability density function

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}.$$

Moreover, $E[X] = \mu$ and $var(X) = \sigma^2$. The characteristic function for X is given by $\exp\{it\mu - \sigma^2 t^2/2\}$. We will discuss this distribution in detail later.

Example 1.8 Gamma Distribution A Gamma distribution has a probability density

$$\frac{1}{\beta^\theta \Gamma(\theta)} x^{\theta-1} \exp\{-\frac{x}{\beta}\}, \quad x > 0$$

denoted by $\Gamma(\theta, \beta)$. It has mean $\theta\beta$ and variance $\theta\beta^2$. Specially, when $\theta = 1$, the distribution is called the exponential distribution, $Exp(\beta)$. When $\theta = n/2$ and $\beta = 2$, the distribution is called the Chi-square distribution with degrees of freedom n, denoted by χ_n^2.

Example 1.9 Cauchy Distribution The density for a random variable $X \sim Cauchy(a, b)$ has the form

$$\frac{1}{b\pi \{1 + (x-a)^2/b^2\}}.$$

Note $E[X] = \infty$. Such a distribution is often used as a counterexample in distribution theory.

Many other distributions can be constructed using some elementary algebra such as summations, products, and quotients of the above special distributions. We will discuss these in the next section.

1.3 Algebra and Transformation of Random Variables (Vectors)

In many applications, one wishes to calculate the distribution of some algebraic expression of independent random variables. For example, suppose that X and Y are two independent random variables. We wish to find the distributions of $X+Y$, XY and X/Y (we assume $Y > 0$ for the last two cases).

The calculation of these algebraic distributions is often done using the conditional expectation. To see how this works, we denote $F_Z(\cdot)$ as the cumulative distribution function of any random variable Z. Then for $X+Y$,

$$F_{X+Y}(z) = E[I(X+Y \le z)] = E_Y[E_X[I(X \le z-Y)|Y]] = E_Y[F_X(z-Y)] = \int F_X(z-y)dF_Y(y);$$

symmetrically,

$$F_{X+Y}(z) = \int F_Y(z-x)dF_X(x).$$

The above formula is called the *convolution formula*, sometimes denoted by $F_X * F_Y(z)$. If both X and Y have densities functions f_X and f_Y respectively, then the density function for $X+Y$ is equal to

$$f_X * f_Y(z) \equiv \int f_X(z-y)f_Y(y)dy = \int f_Y(z-x)f_X(x)dx.$$

Similarly, we can obtain the formulae for XY and X/Y as follows:

$$F_{XY}(z) = E[E[I(XY \le z)|Y]] = \int F_X(z/y)dF_Y(y), \quad f_{XY}(z) = \int f_X(z/y)/y f_Y(y)dy,$$

$$F_{X/Y}(z) = E[E[I(X/Y \le z)|Y]] = \int F_X(yz)dF_Y(y), \quad f_{X/Y}(z) = \int f_X(yz)y f_Y(y)dy.$$

These formulae can be used to construct some familiar distributions from simple random variables. We assume X and Y are independent in the following examples.

Example 1.10 (i) $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.
(ii) $X \sim Cauchy(0, \sigma_1)$ and $Y \sim Cauchy(0, \sigma_2)$ implies $X+Y \sim Cauchy(0, \sigma_1+\sigma_2)$.
(iii) $X \sim Gamma(r_1, \theta)$ and $Y \sim Gamma(r_2, \theta)$ implies that $X+Y \sim Gamma(r_1+r_2, \theta)$.
(iv) $X \sim Poisson(\lambda_1)$ and $Y \sim Poisson(\lambda_2)$ implies $X+Y \sim Poisson(\lambda_1+\lambda_2)$.
(v) $X \sim$ Negative Binomial(m_1, p) and $Y \sim$ Negative Binomial(m_2, p). Then $X+Y \sim$ Negative Binomial(m_1+m_2, p).

The results in Example 1.10 can be verified using the convolution formula. However, these results can also be obtained using characteristic functions, as stated in the following theorem.

Theorem 1.2 Let $\phi_X(t)$ denote the characteristic function for X. Suppose X and Y are independent. Then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$. †

The proof is direct. We can use Theorem 1.2 to find the distribution of $X+Y$. For example, in (i) of Example 1.10, we know $\phi_X(t) = \exp\{\mu_1 t - \sigma_1^2 t^2/2\}$ and $\phi_Y(t) = \exp\{\mu_2 t - \sigma_2^2 t^2/2\}$. Thus,

$$\phi_{X+Y}(t) = \exp\{(\mu_1+\mu_2)t - (\sigma_1^2+\sigma_2^2)t^2/\};$$

while the latter is the characteristic function of a normal distribution with mean $(\mu_1 + \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$.

Example 1.11 Let $X \sim N(0,1)$, $Y \sim \chi_m^2$ and $Z \sim \chi_n^2$ be independent. Then

$$\frac{X}{\sqrt{Y/m}} \sim \text{ Student's } \; t(m),$$

$$\frac{Y/m}{Z/n} \sim \text{Snedecor's } \; F_{m,n},$$

$$\frac{Y}{Y+Z} \sim \text{ Beta}(m/2, n/2),$$

where

$$f_{t(m)}(x) = \frac{\Gamma((m+1)/2)}{\sqrt{\pi m}\Gamma(m/2)} \frac{1}{(1+x^2/m)^{(m+1)/2}} I_{(-\infty,\infty)}(x),$$

$$f_{F_{m,n}}(x) = \frac{\Gamma(m+n)/2}{\Gamma(m/2)\Gamma(n/2)} \frac{(m/n)^{m/2} x^{m/2-1}}{(1+mx/n)^{(m+n)/2}} I_{(0,\infty)}(x),$$

$$f_{\text{Beta}(a,b)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} I(0 < x < 1).$$

Example 1.12 If $Y_1, ..., Y_{n+1}$ are i.i.d Exp(θ), then

$$Z_i = \frac{Y_1 + \ldots + Y_i}{Y_1 + \ldots + Y_{n+1}} \sim \text{Beta}(i, n-i+1).$$

Particularly, $(Z_1, \ldots, Z_n)$ has the same joint distribution as that of the order statistics $(\xi_{n:1}, ..., \xi_{n:n})$ of n Uniform(0,1) random variables.

Both the results in Example 1.11 and 1.12 can be derived using the formulae at the beginning of this section. We now start to examine the transformation of random variables (vectors). Especially, the following theorem holds.

Theorem 1.3 Suppose that X is k-dimension random vector with density function $f_X(x_1, ..., x_k)$. Let g be a one-to-one and continuously differentiable map from R^k to R^k. Then $Y = g(X)$ is a random vector with density function

$$f_X(g^{-1}(y_1, ..., y_k)) |J_{g^{-1}}(y_1, ..., y_k)|,$$

where g^{-1} is the inverse of g and $J_{g^{-1}}$ is the Jacobian of g^{-1}. †

The proof is simply based on the variable-transformation in integration. One application of this result is given in the following example.

Example 1.13 Let X and Y be two independent standard normal random variables. Consider the polar coordinate of (X, Y), i.e., $X = R\cos\Theta$ and $Y = R\sin\Theta$. Then Theorem 1.3 gives that R^2 and Θ are independent and moreover, $R^2 \sim Exp\{2\}$ and $\Theta \sim Uniform(0, 2\pi)$. As an application, if one can simulate variables from a uniform distribution (Θ) and an exponential distribution (R^2), then using $X = R\cos\Theta$ and $Y = R\sin\Theta$ produces variables from a standard normal distribution. This is exactly the way normally distributed numbers are generated in most statistical packages.

1.4 Multivariate Normal Distribution

One particular distribution we will encounter in larger-sample theory is the multivariate normal distribution. A random vector $Y = (Y_1, ..., Y_n)'$ is said to have a multivariate normal distribution with mean vector $\mu = (\mu_1, ..., \mu_n)'$ and non-degenerate covariance matrix $\Sigma_{n\times n}$, denoted as $N(\mu, \Sigma)$ or $N_n(\mu, \Sigma)$ to emphasize Y's dimension, if Y has a joint density as

$$f_Y(y_1, ..., y_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\}.$$

We can derive the characteristic function of Y using the following ad hoc way:

$$\begin{aligned}
\phi_Y(t) &= E[e^{it'Y}] \\
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \int \exp\{it'y - \frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\}dy \\
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \int \exp\{-\frac{1}{2}y'\Sigma^{-1}y + (it + \Sigma^{-1}\mu)'y - \frac{\mu'\Sigma^{-1}\mu}{2}\}dy \\
&= \frac{\exp\{-\mu'\Sigma^{-1}\mu/2\}}{(2\pi)^{n/2}|\Sigma|^{1/2}} \int \exp\left\{-\frac{1}{2}(y - \Sigma it - \mu)'\Sigma^{-1}(y - \Sigma it - \mu)\right. \\
&\qquad \left. + \frac{1}{2}(\Sigma it + \mu)'\Sigma^{-1}(\Sigma it + \mu)\right\} dy \\
&= \exp\{it'\mu - \frac{1}{2}t'\Sigma t\}.
\end{aligned}$$

Particularly, if Y has standard multivariate normal distribution with mean zero and covariance $I_{n\times n}$, $\phi_Y(t) = \exp\{-t't/2\}$.

The following theorem describes the properties of a multivariate normal distribution.

Theorem 1.4 If $Y = A_{n\times k}X_{k\times 1}$ where $X \sim N_k(0, I)$ (standard multivariate normal distribution), then Y's characteristic function is given by

$$\phi_Y(t) = \exp\{-t'\Sigma t/2\}, \quad t = (t_1, ..., t_n) \in R^k,$$

where $\Sigma = AA'$ and $rank(\Sigma) = rank(A)$. Conversely, if $\phi_Y(t) = \exp\{-t'\Sigma t/2\}$ with $\Sigma_{n\times n} \geq 0$ of rank k, then, for some $n \times k$ matrix A for which $AA' = \Sigma$,

$$Y = A_{n\times k}X_{k\times 1} \text{ with } rank(A) = k \text{ and } X \sim N_k(0, I).$$

†

Proof

$$\phi_Y(t) = E[\exp\{it'(AX)\}] = E[\exp\{i(A't)'X\}] = \exp\{-(A't)'(A't)/2\} = \exp\{-t'AA't/2\}.$$

Thus, $\Sigma = AA'$ and $rank(\Sigma) = rank(A)$. Conversely, if $\phi_Y(t) = \exp\{-t'\Sigma t/2\}$, then from matrix theory, there exist an orthogonal matrix O such that $\Sigma = O'DO$, where D is a diagonal matrix with first k diagonal elements positive and the rest $(n-k)$ elements being zero. Denote these positive diagonal elements as $d_1, ..., d_k$. Define $Z = OY$. Then the characteristic function for Z is given by

$$\phi_Z(t) = E[\exp\{it'(OY)\}] = E[\exp\{i(O't)'Y\}] = \exp\{-(O't)'\Sigma(O't)/2\}$$

$$= \exp\{-d_1t_1^2/2 - ... - d_kt_k^2/2\}.$$

This implies that $Z_1, ..., Z_k$ are independent $N(0, d_1), ..., N(0, d_k)$ and $Z_{k+1} = ... = Z_n = 0$. Let $X_i = Z_i/\sqrt{d_i}$ for $i = 1, ..., k$ and write $O' = (B_{n\times k}, C_{n\times(n-k)})$. Then

$$Y = O'Z = B_{n\times k}\begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix} = B_{n\times k}\text{diag}\{(\sqrt{d_1}, ..., \sqrt{d_k})\}\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \equiv AX.$$

Clearly, $rank(A) = k$. †

Theorem 1.5 Suppose that $Y = (Y_1, ..., Y_k, Y_{k+1}, ..., Y_n)'$ has a multivariate normal distribution with mean $\mu = (\mu^{(1)'}, \mu^{(2)'})'$ and a non-degenerate covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then
(i) $(Y_1, ..., Y_k)' \sim N_k(\mu^{(1)}, \Sigma_{11})$.
(ii) $(Y_1, ..., Y_k)'$ and $(Y_{k+1}, ..., Y_n)'$ are independent if and only if $\Sigma_{12} = \Sigma_{21} = 0$.
(iii) For any matrix $A_{m\times n}$, AY has a multivariate normal distribution with mean $A\mu$ and covariance $A\Sigma A'$.
(iv) The conditional distribution of $Y^{(1)} = (Y_1, ..., Y_k)'$ given $Y^{(2)} = (Y_{k+1}, ..., Y_n)'$ is a multivariate normal distribution given as

$$Y^{(1)}|Y^{(2)} \sim N_k(\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(Y^{(2)} - \mu^{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

†

Proof (i) From Theorem 1.4, we obtain that the characteristic function for $(Y_1, ..., Y_k) - \mu^{(1)}$ is given by $\exp\{-t'(D\Sigma)(D\Sigma)'t/2\}$, where $D = (I_{k\times k} \quad 0_{k\times(n-k)})$. Thus, the characteristic function is equal to

$$\exp\left\{-(t_1, ..., t_k)\Sigma_{11}(t_1, ..., t_k)'/2\right\},$$

which is the same as the characteristic function from $N_k(0, \Sigma_{11})$.
(ii) The characteristics function for Y can be written as

$$\exp\left[it^{(1)'}\mu^{(1)} + it^{(2)'}\mu^{(2)} - \frac{1}{2}\left\{t^{(1)'}\Sigma_{11}t^{(1)} + 2t^{(1)'}\Sigma_{12}t^{(2)} + t^{(2)'}\Sigma_{22}t^{(2)}\right\}\right].$$

If $\Sigma_{12} = 0$, the characteristics function can be factorized as the product of the separate functions for $t^{(1)}$ and $t^{(2)}$. Thus, $Y^{(1)}$ and $Y^{(2)}$ are independent. The converse is obviously true.
(iii) The result follows from Theorem 1.4.

(iv) Consider $Z^{(1)} = Y^{(1)} - \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}(Y^{(2)} - \mu^{(2)})$. From (iii), $Z^{(1)}$ has a multivariate normal distribution with mean zero and covariance calculated by

$$Cov(Z^{(1)}, Z^{(1)}) = Cov(Y^{(1)}, Y^{(1)}) - 2\Sigma_{12}\Sigma_{22}^{-1}Cov(Y^{(2)}, Y^{(1)}) + \Sigma_{12}\Sigma_{22}^{-1}Cov(Y^{(2)}, Y^{(2)})\Sigma_{22}^{-1}\Sigma_{21}$$
$$= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

On the other hand,

$$Cov(Z^{(1)}, Y^{(2)}) = Cov(Y^{(1)}, Y^{(2)}) - \Sigma_{12}\Sigma_{22}^{-1}Cov(Y^{(2)}, Y^{(2)}) = 0.$$

From (ii), $Z^{(1)}$ is independent of $Y^{(2)}$. Then the conditional distribution $Z^{(1)}$ given $Y^{(2)}$ is the same as the unconditional distribution of $Z^{(1)}$; i.e.,

$$Z^{(1)}|Y^{(2)} \sim N(0, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

The result follows. †

With normal random variables, we can use algebra of random variables to construct a number of useful distributions. The first one is the Chi-square distribution. Suppose $X \sim N_n(0, I)$, then $\|X\|^2 = \sum_{i=1}^n X_i^2 \sim \chi_n^2$, the chi-square distribution with n degrees of freedom. One can use the convolution formula to obtain that the density function for χ_n^2 is equal to the density for the $Gamma(n/2, 2)$, denoted by $g(y; n/2, 1/2)$.

Corollary 1.1 If $Y \sim N_n(0, \Sigma)$ with $\Sigma > 0$, then $Y'\Sigma^{-1}Y \sim \chi_n^2$. †

Proof Since $\Sigma > 0$, there exists a positive definite matrix A such that $AA' = \Sigma$. Then $X = A^{-1}Y \sim N_n(0, I)$. Thus

$$Y'\Sigma^{-1}Y = X'X \sim \chi_n^2.$$

†

Suppose $X \sim N(\mu, 1)$. Define $Y = X^2, \delta = \mu^2$. Then Y has density

$$f_Y(y) = \sum_{k=0}^{\infty} p_k(\delta/2)g(y; (2k+1)/2, 1/2),$$

where $p_k(\delta/2) = \exp(-\delta/2)(\delta/2)^k/k!$. Another ways to obtain this is: $Y|K = k \sim \chi_{2k+1}^2$ where $K \sim Poisson(\delta/2)$. We say Y has the noncentral chi-square distribution with 1 degree of freedom and noncentrality parameter δ and write $Y \sim \chi_1^2(\delta)$. More generally, if $X = (X_1, ..., X_n)' \sim N_n(\mu, I)$ and let $Y = X'X$, then Y has a density $f_Y(y) = \sum_{k=0}^{\infty} p_k(\delta/2)g(y; (2k+n)/2, 1/2)$ where $\delta = \mu'\mu$. We write $Y \sim \chi_n^2(\delta)$ and say Y has a noncentral chi-square distribution with n degrees of freedom and noncentrality parameter δ. It is then easy to show that if $X \sim N(\mu, \Sigma)$, then $Y = X'\Sigma^{-1}X \sim \chi_n^2(\delta)$.

If $X \sim N(0, 1), Y \sim \chi_n^2$ and they are independent, then $X/\sqrt{Y/n}$ is called the t-distribution with n degrees of freedom. If $Y_1 \sim \chi_m^2, Y_2 \sim \chi_n^2$ and Y_1 and Y_2 are independent, then $(Y_1/m)/(Y_2/m)$ is called an F-distribution with degrees freedom of m and n. These distributions have already been introduced in Example 1.11.

1.5 Families of Distributions

In Examples 1.1-1.12, we have listed a number of different distributions. Interestingly, a number of them can be unified into a family of general distribution form. One advantage of this unification is that in order to study the properties of each distribution within the family, we can examine this family as a whole.

The first family of distributions is called the *location-scale* family. Suppose that X has a density function $f_X(x)$. Then the location-scale family based on X consists of all the distributions generated by $aX + b$ where a is a positive constant (scale parameter) and b is a constant called the location parameter. We notice that distributions such as $N(\mu, \sigma^2)$, $Uniform(a, b)$, $Cauchy(\mu, \sigma)$ belong to the location-scale family. For a location-scale family, we can easily see that $aX + b$ has a density $f_X((y - b)/a)/a$ and it has mean $aE[X] + b$ and variance $a^2 var(X)$.

The second important family, which we will discuss in more detail, is called the *exponential family*. In fact, many examples of either univariate or multivariate distributions, including binomial, poisson distributions for discrete variables and normal distribution, gamma distribution, and beta distribution for continuous variables, belong to some exponential family. Especially, a family of distributions, $\{P_\theta\}$, is said to form an s-parameter exponential family if the distributions P_θ have densities (with respect to some common dominating measure μ) of the form

$$p_\theta(x) = \exp\left\{\sum_{k=1}^{s} \eta_k(\theta)T_k(x) - B(\theta)\right\} h(x).$$

Here η_i and B are real-valued functions of θ and T_i are real-value functions of x. When $\{\eta_k(\theta)\} = \theta$, the above form is called the canonical form of the exponential family. Clearly, it stipulates that

$$\exp\{B(\theta)\} = \int \exp\{\sum_{k=1}^{s} \eta_k(\theta)T_k(x)\}h(x)d\mu(x) < \infty.$$

Example 1.14 $X_1, ..., X_n$ are i.i.d according to $N(\mu, \sigma^2)$. Then the joint density of $(X_1, ..., X_n)$ is given by

$$\exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 - \frac{n}{2\sigma^2}\mu^2\right\} \frac{1}{(\sqrt{2\pi}\sigma)^n}.$$

Then $\eta_1(\theta) = \mu/\sigma^2$, $\eta_2(\theta) = -1/2\sigma^2$, $T_1(x_1, ..., x_n) = \sum_{i=1}^n x_i$, and $T_2(x_1, ..., x_n) = \sum_{i=1}^n x_i^2$.

Example 1.15 X has binomial distribution $Binomial(n, p)$. The distribution of $X = x$ can written as

$$\exp\{x \log \frac{p}{1-p} + n\log(1-p)\}\binom{n}{x}.$$

Clearly, $\eta(\theta) = \log(p/(1 - p))$ and $T(x) = x$.

Example 1.16 X has poisson distribution with poisson rate λ. Then

$$P(X = x) = \exp\{x \log \lambda - \lambda\}/x!.$$

Thus, $\eta(\theta) = \log \lambda$ and $T(x) = x$.

Since the exponential family covers a number of familiar distributions, one can study the exponential family as a whole to obtain some general results applicable to all the members within the family. One result is to derive the moment generation function for $(T_1, ..., T_s)$, which is defined as

$$M_T(t_1, ..., t_s) = E\left[\exp\{t_1T_1 + ... + t_sT_s\}\right].$$

Note that the coefficients in the Taylor expansion of M_T correspond to the moments of $(T_1, ..., T_s)$.

Theorem 1.6 Suppose the densities of an exponential family can be written as the canonical form

$$\exp\{\sum_{k=1}^{s} \eta_k T_k(x) - A(\eta)\}h(x),$$

where $\eta = (\eta_1, ..., \eta_s)'$. Then for $t = (t_1, ..., t_s)'$,

$$M_T(t) = \exp\{A(\eta + t) - A(\eta)\}.$$

†

Proof It follows from that

$$M_T(t) = E\left[\exp\{t_1T_1 + ... + t_sT_s\}\right] = \int \exp\{\sum_{k=1}^{s} (\eta_i + t_i)T_i(x) - A(\eta)\}h(x)d\mu(x)$$

and

$$\exp\{A(\eta)\} = \int \exp\{\sum_{k=1}^{s} \eta_i T_i(x)\}h(x)d\mu(x).$$

†

Therefore, for an exponential family with canonical form, we can apply Theorem 1.6 to calculate moments of some statistics. Another generating function is called the cumulant generating function defined as

$$K_T(t_1, ..., t_s) = \log M_T(t_1, ..., t_s) = A(\eta + t) - A(\eta).$$

Its coefficients in the Taylor expansion are called the cumulants for $(T_1, ..., T_s)$.

Example 1.17 In normal distribution of Example 1.14 with $n = 1$ and σ^2 fixed, $\eta = \mu/\sigma^2$ and

$$A(\eta) = \frac{1}{2\sigma^2}\mu^2 = \eta^2\sigma^2/2.$$

Thus, the moment generating function for $T = X$ is equal to

$$M_T(t) = \exp\{\frac{\sigma^2}{2}((\eta + t)^2 - \eta^2)\} = \exp\{\mu t + t^2\sigma^2/2\}.$$

From the Taylor expansion, we can obtain that the moments of X, whose mean is zero ($\mu = 0$), is given by

$$E[X^{2r+1}] = 0, E[X^{2r}] = 1 \cdot 2 \cdots (2r - 1)\sigma^{2r}, r = 1, 2, ...$$

Example 1.18 X has a gamma distribution with density

$$\frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b}, x > 0.$$

For fixed a, it has a canonical form

$$\exp\{-x/b + (a-1)\log x - \log(\Gamma(a)b^a)\} I(x > 0).$$

Correspondingly, $\eta = -1/b, T = X, A(\eta) = \log(\Gamma(a)b^a) = a\log(-1/\eta) + \log\Gamma(a)$. Then the moment generating function for $T = X$ is given by

$$M_X(t) = \exp\{a \log \frac{\eta}{\eta + t}\} = (1 - bt)^{-a}.$$

After Taylor expansion around zero, we obtain

$$E[X] = ab, E[X^2] = ab^2 + (ab)^2, ...$$

As a further note, the exponential family has an important role in classical statistical inference since it possesses many nice statistical properties. We will revisit this in Chapter 4.

READING MATERIALS: You should read Lehmann and Casella, Sections 1.4 and 1.5.

PROBLEMS

1. Verify the densities of $t(m)$ and $F_{m,n}$ in Example 1.11.

2. Verify the two results in Example 1.12.

3. Suppose $X \sim N(\nu, 1)$. Show that $Y = X^2$ has a density

$$f_Y(y) = \sum_{k=0}^{\infty} p_k(\mu^2/2) g(y; (2k+1)/2, 1/2),$$

where $p_k(\mu^2/2) = \exp(-\mu^2/2)(\mu^2/2)^k/k!$ and $g(y; n/2, 1/2)$ is the density of $Gamma(n/2, 2)$.

4. Suppose $X = (X_1, ..., X_n) \sim N(\mu, I)$ and let $Y = X'X$. Show that Y has a density

$$f_Y(y) = \sum_{k=0}^{\infty} p_k(\mu'\mu/2) g(y; (2k+n)/2, 1/2).$$

5. Let $X \sim Gamma(\alpha_1, \beta)$ and $Y \sim Gamma(\alpha_2, \beta)$ be independent random variables. Derive the distribution of $X/(X+Y)$.

6. Show that for any random variables X, Y and Z,

$$Cov(X,Y) = E[Cov(X,Y|Z)] + Cov(E[X|Z], E[Y|Z]),$$

where $Cov(X,Y|Z)$ is the conditional covariance of X and Y given Z.

7. Let X and Y be i.i.d Uniform(0,1) random variables. Define $U = X-Y$, $V = \max(X,Y) = X \vee Y$.

   (a) What is the range of (U,V)?

   (b) find the joint density function $f_{U,V}(u,v)$ of the pair (U,V). Are U and V independent?

8. Suppose that for $\theta \in R$,

$$f_\theta(u,v) = \{1 + \theta(1-2u)(1-2v)\}\, I(0 \le u \le 1, 0 \le v \le 1).$$

   (a) For what values of θ is f_θ a density function in $[0,1]^2$?

   (b) For the set of θ's identified in (a), find the corresponding distribution function F_θ and show that it has Uniform(0,1) marginal distributions.

   (c) If $(U,V) \sim f_\theta$, compute the correlation $\rho(U,V) \equiv \rho$ as a function of θ.

9. Suppose that F is the distribution function of random variables X and Y with $X \sim$ Uniform$(0,1)$ marginally and $Y \sim$ Uniform$(0,1)$ marginally. Thus, $F(x,y)$ satisfies

$$F(x,1) = x, \quad 0 \le x \le 1, \quad \text{and} \quad F(1,y) = y, \quad 0 \le y \le 1.$$

   (a) Show that

$$F(x,y) \le x \wedge y$$

   for all $0 \le x \le 1, 0 \le y \le 1$. Here $x \wedge y = \min(x,y)$ and we denote it as $F_U(x,y)$.

   (b) Show that

$$F(x,y) \ge (x+y-1)^+$$

   for all $0 \le x \le 1, 0 \le y \le 1$. Here $(x+y-1)^+ = \max(x+y-1, 0)$ and we denote it as $F_L(x,y)$.

   (c) Show that F_U is the distribution function of (X,X) and F_L is the distribution function of $(X, 1-X)$.

10. (a) If $W \sim \chi^2_2 = Gamma(1,2)$, find the density of W, the distribution function W and the inverse distribution function explicitly.

   (b) Suppose that $(X,Y) \sim N(0, I_{2\times 2})$. In two-dimensional plane, let R be the distance of (X,Y) from $(0,0)$ and θ be the angle between the line from (0,0) to (X,Y) and the right-half line of x-axis. Then $X = R\cos\Theta$ and $Y = R\sin\Theta$. Show that R and Θ are independent random variables with $R^2 \sim \chi^2_2$ and $\Theta \sim$ Uniform$(0, 2\pi)$.

   (c) Use the above two results to show how to use two independent Uniform(0,1) random variables U and V to generate two standard normal random variables. *Hint*: use one result that if X has a distribution function F then $F(X)$ has a uniform distribution in $[0,1]$.

11. Suppose that $X \sim F$ on $[0, \infty)$, $Y \sim G$ on $[0, \infty)$, and X and Y are independent random variables. Let $Z = \min\{X, Y\} = X \wedge Y$ and $\Delta = I(X \leq Y)$.

    (a) Find the joint distribution of (Z, Δ).

    (b) If $X \sim$ Exponential(λ) and $Y \sim$ Exponential(μ), show that Z and Δ are independent.

12. Let $X_1, ..., X_n$ be i.i.d $N(0, \sigma^2)$. $(w_1, ..., w_n)$ is a constant vector such that $w_1, ..., w_n > 0$ and $w_1 + ... + w_n = 1$. Define $\bar{X}_{nw} = \sqrt{w_1}X_1 + ... + \sqrt{w_n}X_n$. Show that

    (a) $Y_n = \bar{X}_{nw}/\sigma \sim N(0, 1)$.

    (b) $(n-1)S_n^2/\sigma^2 = (\sum_{i=1}^n X_i^2 - \bar{X}_{nw}^2)/\sigma^2 \sim \chi_{n-1}^2$.

    (c) Y_n and S_n^2 are independent so $T_n = Y_n/\sqrt{S_n^2} \sim t_{n-1}/\sigma$.

    (d) when $w_1 = ... = w_n = 1/n$, show that Y_n is the standardized sample mean and S_n^2 is the sample variance.

    Hint: Consider an orthogonal matrix Σ such that the first row is $(\sqrt{w_1}, ..., \sqrt{w_n})$. Let

    $$\begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = \Sigma \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}.$$

    Then $Y_n = Z_1/\sigma$ and $(n-1)S_n^2/\sigma^2 = (Z_2^2 + ... + Z_n^2)/\sigma^2$.

13. Let $X_{n\times 1} \sim N(0, I_{n\times n})$. Suppose that A is a symmetric matrix with rank r. Then $X'AX \sim \chi_r^2$ if and only if A is a projection matrix (that is, $A^2 = A$). *Hint*: use the following result from linear algebra: for any symmetric matrix, there exits an orthogonal matrix O such that $A = O' \text{ diag}((d_1, ..., d_n))O$; A is a projection matrix if and only if $d_1, ..., d_n$ take values of 0 or 1's.

14. Let $W_m \sim$ Negative Binomial(m, p). Consider p as a parameter.

    (a) Write the distribution as an exponential family.

    (b) Use the result for the exponential family to derive the moment generating function of W_m, denoted by $M(t)$.

    (c) Calculate the first and the second cumulants of W_m. By definition, in the expansion of the cumulant generating function,

    $$\log M(t) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} t^k,$$

    μ_k is the kth cumulant of W_m. Note that these two cumulants are exactly the mean and the variance of W_m.

15. For the density $C \exp\left\{-|x|^{1/2}\right\}, -\infty < x < \infty$, where C is the normalizing constant, show that moments of all orders exist but the moment generating function exists only at $t = 0$.

16. Lehmann and Casella, page 64, problem 4.2.
17. Lehmann and Casella, page 66, problem 5.6.
18. Lehmann and Casella, page 66, problem 5.7.
19. Lehmann and Casella, page 66, problem 5.8.
20. Lehmann and Casella, page 66, problem 5.9.
21. Lehmann and Casella, page 66, problem 5.10.
22. Lehmann and Casella, page 67, problem 5.12.
23. Lehmann and Casella, page 67, problem 5.14.

CHAPTER 2 MEASURE, INTEGRATION AND PROBABILITY

This chapter is an introduction to (probability) measure theories, a foundation for all the probabilistic and statistical framework. We first give the definition of a measure space. Then we introduce measurable functions in a measure space and the integration and convergence of measurable functions. Further generalization including the product of two measures and the Radon-Nikodym derivatives of two measures is introduced. As a special case, we describe how the concepts and the properties in measure space are used in parallel in a probability measure space.

2.1 A Review of Set Theory and Topology in Real Space

We review some basic concepts in set theory. A *set* is a collection of elements, which can be a collection of real numbers, a group of abstract subjects and etc. In most of cases, we consider that these elements come from one largest set, called a *whole space*. By custom, a whole space is denoted by Ω so any set is simply a *subset* of Ω. We can exhaust all possible subsets of Ω then the collection of all these subsets is denoted as 2^{Ω}, called the *power set* of Ω. We also include the empty set, which has no element at all and is denoted by $\emptyset$, in this power set.

For any two subsets A and B of the whole space Ω, A is said to be a *subset* of B if B contains all the elements of A, denoted as $A \subseteq B$. For arbitrary number of sets $\{A_\alpha : \alpha \text{ is some index}\}$, where the index of α can be finite, countable or uncountable, we define the *intersection* of these sets as the set which contains all the elements common to A_α for any α. The intersection of these sets is denoted as $\cap_\alpha A_\alpha$. A_α's are *disjoint* if any two sets have empty intersection. We can also define the *union* of these sets as the set which contains all the elements belonging to at least one of these sets, denoted as $\cup_\alpha A_\alpha$. Finally, we introduce the *complement* of a set A, denoted by A^c, to be the set which contains all the elements not in A. Among the definitions of set intersection, union and complement, the following relationships are clear: for any B and $\{A_\alpha\}$,

$$B \cap \{\cup_\alpha A_\alpha\} = \cup_\alpha \{B \cap A_\alpha\}, \quad B \cup \{\cap_\alpha A_\alpha\} = \cap_\alpha \{B \cup A_\alpha\},$$

$$\{\cup_\alpha A_\alpha\}^c = \cap_\alpha A_\alpha^c, \quad \{\cap_\alpha A_\alpha\}^c = \cup_\alpha A_\alpha^c. \quad \text{(de Morgan law)}$$

Sometimes, we use $(A - B)$ to denote a subset of A excluding any elements in B. Thus $(A - B) = A \cap B^c$. Using this notation, we can always partition the union of any countable sets $A_1, A_2, ...$ into a union of countable disjoint sets:

$$A_1 \cup A_2 \cup A_3 \cup ... = A_1 \cup (A_2 - A_1) \cup (A_3 - A_1 \cup A_2) \cup ...$$

For a sequence of sets $A_1, A_2, A_3, ...$, we now define the limit sets of the sequence. The *upper limit set* of the sequence is the set which contains the elements belonging to infinite number of the sets in this sequence; the *lower limit set* of the sequence is the set which contains the elements belonging to all the sets except a finite number of them in this sequence. The former is denoted by $\overline{\lim}_n A_n$ or $\limsup_n A_n$ and the latter is written as $\underline{\lim}_n A_n$ or $\liminf_n A_n$. We can show

$$\limsup_n A_n = \cap_{n=1}^{\infty} \left\{\cup_{m=n}^{\infty} A_m\right\}, \quad \liminf_n A_n = \cup_{n=1}^{\infty} \left\{\cap_{m=n}^{\infty} A_m\right\}.$$

When both limit sets agree, we say that the sequence has a limit set. In the calculus, we know that for any sequence of real numbers $x_1, x_2, ...$, it has a upper limit, $\limsup_n x_n$, and a lower limit, $\liminf_n x_n$, where the former refers to the upper bound of the limits for any convergent subsequences and the latter is the lower bound. It should be cautious that such upper limit or lower limit is different from the upper limit or lower limit of sets.

The second part of this section reviews some basic topology in a real line. Because the distance between any two points is well defined in a real line, we can define a *topology* in a real line. A set A of the real line is called an *open set* if for any point $x \in A$, there exists an open interval $(x - \epsilon, x + \epsilon)$ contained in A. Clearly, any open interval (a, b) where a could be $-\infty$ and b could be ∞, is an open set. Moreover, for any number of open sets A_α where α is an index, it is easy to show that $\cup_\alpha A_\alpha$ is open. A *closed set* is defined as the complement of an open set. It can also be show that A is closed if and only if for any sequence $\{x_n\}$ in A such that $x_n \to x$, x must belong to A. By the de Morgan law, we also see that the intersection of any number of closed sets is still closed. Only $\emptyset$ and the whole real line are both open set and closed set; there are many sets neither open or closed, for example, the set of all the rational numbers. If a closed set A is bounded, A is also called a *compact set.* These basic topological concepts will be used later. Note that the concepts of open set or closed set can be easily generalized to any finite dimensional real space.

2.2 Measure Space

2.2.1 Introduction

Before we introduce a formal definition of measure space, let us examine the following examples.

Example 2.1 Suppose that a whole space Ω contains countable number of distinct points $\{x_1, x_2, ...\}$. For any subset A of Ω, we define a set function $\mu^{\#}(A)$ as the number of points in A. Therefore, if A has n distinct points, $\mu^{\#}(A) = n$; if A has infinite many number of points, then $\mu^{\#}(A) = \infty$. We can easily show that (a) $\mu^{\#}(\emptyset) = 0$; (b) if $A_1, A_2, ...$ are disjoint sets of Ω, then $\mu^{\#}(\cup_n A_n) = \sum_n \mu^{\#}(A_n)$. We will see later that $\mu^{\#}$ is a measure called the *counting measure* in Ω.

Example 2.2 Suppose that the whole space $\Omega = R$, the real line. We wish to measure the sizes of any possible subsets in R. Equivalently, we wish to define a set function λ which assigns some non-negative values to the sets of R. Since λ measures the size of a set, it is clear that λ should satisfy (a) $\lambda(\emptyset) = 0$; (b) for any disjoint sets $A_1, A_2, ...$ whose sizes are measurable, $\lambda(\cup_n A_n) = \sum_n \lambda(A_n)$. Then the question is how to define such a λ. Intuitively, for any interval $(a, b]$, such a value can be given as the length of the interval, i.e., $(b - a)$. We can further define λ-value of any set in $\mathcal{B}_0$, which consists of $\emptyset$ together with all finite unions of disjoint intervals with the form $\cup_{i=1}^n (a_i, b_i]$, or $\cup_{i=1}^n (a_i, b_i] \cup (a_{n+1}, \infty)$, $(-\infty, b_{n+1}] \cup \cup_{i=1}^n (a_i, b_i]$, with $a_i, b_i \in R$, as the total length of the intervals. But can we go beyond it, as the real line has far far many sets which are not intervals, for example, the set of rational numbers? In other words, is it possible to extend the definition of λ to more sets beyond intervals while preserving the values for intervals? The answer is yes and will be given shortly. Moreover, such an extension is unique. Such set function λ is called the *Lebesgue measure* in the real line.

Example 2.3 This example simply asks the same question as in Example 2.2, but now on k-dimensional real space. Still, we define a set function which assigns any hypercube its volume and wish to extend its definition to more sets beyond hypercubes. Such a set function is called the *Lebesgue measure in* R^k, denoted as λ^k.

From the above examples, we can see that three pivotal components are necessary in defining a measure space:

(i) the whole space, Ω, for example, $\{x_1, x_2, ...\}$ in Example 2.1, R and R^k in the last two examples,

(ii) a collection of subsets whose sizes are measurable, for example, all the subsets in Example 2.1, the unknown collection of subsets including all the intervals in Example 2.2,

(iii) a set function which assigns negative values (sizes) to each set of (ii) and satisfies properties (a) and (b) in the above examples.

For notation, we use $(\Omega, \mathcal{A}, \mu)$ to denote each of them; i.e., Ω denotes the whole space, $\mathcal{A}$ denotes the collection of all the measurable sets, and μ denotes the set function which assigns non-negative values to all the sets in $\mathcal{A}$.

2.2.2 Definition of a measure space

Obviously, Ω should be a fixed non-void set. The main difficulty is the characterization of $\mathcal{A}$. However, let us understand intuitively what kinds of sets should be in $\mathcal{A}$: as a reminder, $\mathcal{A}$ contains the sets whose sizes are measurable. Now suppose that a set A in $\mathcal{A}$ is measurable then we would think that its complement is also measurable, intuitively, the size of the whole space minus the size of A. Additionally, if $A_1, A_2, ...$ are in $\mathcal{A}$ so are measurable, then we should be able to measure the total size of $A_1, A_2, ...$, i.e, the union of these sets. Hence, as expected, $\mathcal{A}$ should include the complement of a set which is in $\mathcal{A}$ and the union of any countable number of sets which are in $\mathcal{A}$. This turns out that $\mathcal{A}$ must be a σ-field, whose definition is given below.

Definition 2.1 (fields, σ-fields) A non-void class $\mathcal{A}$ of subsets of Ω is called a:
(i) *field* or *algebra* if $A, B \in \mathcal{A}$ implies that $A \cup B \in \mathcal{A}$ and $A^c \in \mathcal{A}$; equivalently, $\mathcal{A}$ is closed under complements and finite unions.
(ii) *σ-field* or *σ-algebra* if $\mathcal{A}$ is a field and $A_1, A_2, ... \in \mathcal{A}$ implies $\cup_{i=1}^\infty A_i \in \mathcal{A}$; equivalently, $\mathcal{A}$ is closed under complements and countable unions. †

In fact, a σ-field is not only closed under complement and countable union but also closed under countable intersection, as shown in the following proposition.

Proposition 2.1. (i) For a field $\mathcal{A}$, $\emptyset, \Omega \in \mathcal{A}$ and if $A_1, ..., A_n \in \mathcal{A}$, $\cap_{i=1}^n A_i \in \mathcal{A}$.
(ii) For a σ-field $\mathcal{A}$, if $A_1, A_2, ... \in \mathcal{A}$, then $\cap_{i=1}^\infty A_i \in \mathcal{A}$. †

Proof (i) For any $A \in \mathcal{A}$, $\Omega = A \cup A^c \in \mathcal{A}$. Thus, $\emptyset = \Omega^c \in \mathcal{A}$. If $A_1, ..., A_n \in \mathcal{A}$ then $\cap_{i=1}^n A_i = (\cup_{i=1}^n A_i^c)^c \in \mathcal{A}$.
(ii) can be shown using the definition of a (σ-)field and the de Morgan law. †

We now give a few examples of σ-field or field.

Example 2.4 The class $\mathcal{A} = \{\emptyset, \Omega\}$ is the smallest σ-field and $2^\Omega = \{A : A \subset \Omega\}$ is the largest σ-field. Note that in Example 2.1, we choose $\mathcal{A} = 2^\Omega$ since each set of $\mathcal{A}$ is measurable.

Example 2.5 Recall $\mathcal{B}_0$ in Example 2.2. It can be checked that $\mathcal{B}_0$ is a field but not a σ-field, since $(a, b) = \cup_{n=1}^\infty (a, b - \frac{1}{n}]$ does not belong to $\mathcal{B}_0$.

After defining a σ-field $\mathcal{A}$ on Ω, we can start to introduce the definition of a measure. As implicated before, a measure can be understood as a set-function which assigns non-negative value to each set in $\mathcal{A}$. However, the values assigned to the sets of $\mathcal{A}$ are not arbitrary and they should be compatible in the following sense.

Definition 2.2 (measure, probability measure) (i) A *measure* μ is a function from a σ-field $\mathcal{A}$ to $[0, \infty)$ satisfying: $\mu(\emptyset) = 0$; $\mu(\cup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \mu(A_n)$ for any countable (finite) disjoint sets $A_1, A_2, ... \in \mathcal{A}$. The latter is called the *countable additivity.*
(ii) Additionally, if $\mu(\Omega) = 1$, μ is a *probability measure* and we usually use P instead of μ to indicate a probability measure. †

The following proposition gives some properties of a measure.

Proposition 2.2 (i) If $\{A_n\} \subset \mathcal{A}$ and $A_n \subset A_{n+1}$ for all n, then $\mu(\cup_{n=1}^\infty A_n) = \lim_{n\to\infty} \mu(A_n)$.
(ii) If $\{A_n\} \subset \mathcal{A}$, $\mu(A_1) < \infty$ and $A_n \supset A_{n+1}$ for all n, then $\mu(\cap_{n=1}^\infty A_n) = \lim_{n\to\infty} \mu(A_n)$.
(iii) For any $\{A_n\} \subset \mathcal{A}$, $\mu(\cup_n A_n) \le \sum_n \mu(A_n)$ (*countable sub-additivity*). †

Proof (i) It follows from

$$\mu(\cup_{n=1}^\infty A_n) = \mu(A_1 \cup (A_2 - A_1) \cup ...) = \mu(A_1) + \mu(A_2 - A_1) + ...$$
$$= \lim_n \{\mu(A_1) + \mu(A_2 - A_1) + ... + \mu(A_n - A_{n-1})\} = \lim_n \mu(A_n).$$

(ii) First,

$$\mu(\cap_{n=1}^\infty A_n) = \mu(A_1) - \mu(A_1 - \cap_{n=1}^\infty A_n) = \mu(A_1) - \mu(\cup_{n=1}^\infty (A_1 \cap A_n^c)).$$

Then since $A_1 \cap A_n^c$ is increasing, from (i), the second term is equal to $\lim_n \mu(A_1 \cap A_n^c) = \mu(A_1) - \lim_n \mu(A_n)$. (ii) thus holds.
(iii) From (i), we have

$$\mu(\cup_n A_n) = \lim_n \mu(A_1 \cup ... \cup A_n) = \lim_n \left\{ \sum_{i=1}^n \mu(A_i - \cup_{j<i} A_j) \right\}$$

$$\leq \lim_n \sum_{i=1}^n \mu(A_i) = \sum_n \mu(A_n).$$

The result holds. † .

If a class of sets $\{A_n\}$ is increasing or decreasing, we can treat $\cup_n A_n$ or $\cap_n A_n$ as its limit set. Then Proportion 2.2 says that such a limit can be taken out of the measure for increasing sets and it can be taken out of the measure for decreasing set if the measure of some A_n is finite. For an arbitrary sequence of sets $\{A_n\}$, in fact, similar to Proposition 2.2, we can show

$$\mu(\liminf_n A_n) = \lim_n \mu(\cap_{k=n}^\infty A_n) \leq \liminf_n \mu(A_n).$$

The triplet $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. Any set in $\mathcal{A}$ is called a *measurable set*. Particularly, if $\mu = P$ is a probability measure, $(\Omega, \mathcal{A}, P)$ is called a *probability measure space*, abbreviated as probability space; an element in Ω is called a *probability sample* and a set in $\mathcal{A}$ is called a *probability event*. As an additional note, a measure μ is called *σ-finite* if there exists a countable sets $\{F_n\} \subset \mathcal{A}$ such that $\Omega = \cup_n F_n$ and for each F_n, $\mu(F_n) < \infty$.

Example 2.6 (i) A measure μ on $(\Omega, \mathcal{A})$ is discrete if there are finitely or countably many points $\omega_i \in \Omega$ and masses $m_i \in [0, \infty)$ such that

$$\mu(A) = \sum_{\omega_i \in A} m_i, \quad A \in \mathcal{A}.$$

Some examples include probability measures in discrete distributions.
(ii) in Example 2.1, we define a counting measure $\mu^\#$ in a countable space. This definition can be generalized to any space. Especially, a counting measure in the space R is not σ-finite.

2.2.3 Construction of a measure space

Even though $(\Omega, \mathcal{A}, \mu)$ is well defined, a practical question is how to construct such a measure space. In the specific Example 2.2, one asks whether we can find a σ-field including all the intervals of $\mathcal{B}_0$ and on this σ-field, whether we can define a measure λ such that λ assigns any interval its length. Even more general, suppose that we have a class of sets $\mathcal{C}$ and a set function μ satisfying property (i) of Definition 2.2. Can we find a σ-field which contains all the sets of $\mathcal{C}$ and moreover, can we obtain a measure defined for any set of this σ-field such that the measure agrees with μ in $\mathcal{C}$? The answer is positive for the first question and is positive for the second question when $\mathcal{C}$ is a field. Indeed, such a σ-field is the smallest σ-field containing all the sets of $\mathcal{C}$, called *σ-field generated by* $\mathcal{C}$, and such a measure can be obtained using the measure extension result as given below.

First, we show that the σ-field generated by $\mathcal{C}$ exists and is unique.

Proposition 2.3 (i) Arbitrary intersections of fields (σ-fields) are fields (σ-fields).
(ii) For any class $\mathcal{C}$ of subsets of Ω, there exists a minimal σ-field containing $\mathcal{C}$ and we denote it as $\sigma(\mathcal{C})$. †

Proof (i) can be shown using the definitions of a (σ-)field. For (ii), we define

$$\sigma(\mathcal{C}) = \cap_{\mathcal{C} \subset \mathcal{A}, \mathcal{A} \text{ is } \sigma\text{-field}} \mathcal{A},$$

i.e., the intersection of all the σ-fields containing $\mathcal{C}$. From (i), this class is also σ-field. Obviously, it is the minimal one among all the σ-fields containing $\mathcal{C}$. †

Then the following result shows that an extension of μ to $\sigma(\mathcal{C})$ is possible and unique if $\mathcal{C}$ is a field.

Theorem 2.1 (Caratheodory Extension Theorem) A measure μ on a field $\mathcal{C}$ can be extended to a measure on the minimal σ-field $\sigma(\mathcal{C})$. If μ is σ-finite on $\mathcal{C}$, then the extension is unique and also σ-finite. †

Proof The proof is skipped. Essentially, we define an extension of μ using the following outer measure definition: for any set A,

$$\mu^*(A) = \inf \left\{ \sum_{i=1}^{\infty} \mu(A_i) : A_i \in \mathcal{C}, A \subset \cup_{i=1}^{\infty} A_i \right\}.$$

This is also the way of calculating the measure of any set in $\sigma(\mathcal{C})$. †

Using the above results, we can construct many measure spaces. In Example 2.2, we first generate a σ-field containing all the intervals of $\mathcal{B}_0$. Such a σ-field is called the *Borel σ-field*, denoted by $\mathcal{B}$, and any set in $\mathcal{B}$ is called a *Borel set*. Then we can extend λ to $\mathcal{B}$ and the obtained measure is called the Lebesgue measure. The triplet $(R, \mathcal{B}, \lambda)$ is named the *Borel measure space*. Similarly, in Example 2.3, we can obtain the Borel measure space in R^k, denoted by $(R^k, \mathcal{B}^k, \lambda^k)$.

We can also obtain many different measures in the Borel σ-field. To do that, let F be a fixed generalized distribution function: F is non-decreasing and right-continuous. Then starting from any interval $(a, b]$, we define a set function $\lambda_F((a, b]) = F(b) - F(a)$ thus λ_F can be easily defined for any set of $\mathcal{B}_0$. Using the σ-field generation and measure extension, we thus obtain a different measure λ_F in $\mathcal{B}$. Such a measure is called the *Lebesgue-Stieltjes measure generated by F*. Note that the Lebesuge measure is a special case with $F(x) = x$. Particularly, if F is a distribution function, i.e., $F(\infty) = 1$ and $F(-\infty) = 0$, this measure is a probability measure in R.

In a measure space $(\Omega, \mathcal{A}, \mu)$, it is intuitive to assume that any subsets of a set with measure zero should be given measure zero. However, these subsets may not be included in $\mathcal{A}$. Therefore, a final stage of constructing a measure space is to perform the completion by including such nuisance sets in the σ-field. Especially, a general definition of the *completion of a measure* is given as follows: for a measure space $(\Omega, \mathcal{A}, \mu)$, a completion is another measure space $(\Omega, \bar{\mathcal{A}}, \bar{\mu})$ where

$$\bar{\mathcal{A}} = \{A \cup N : A \in \mathcal{A},\ N \subset B \text{ for some } B \in \mathcal{A} \text{ such that } \mu(B) = 0\}$$

and let $\bar{\mu}(A \cup N) = \mu(A)$. Particularly, the completion of the Borel measure space is called the *Lebesgue measure space* and the completed Borel σ-field is called the σ-field of *Lebesgue sets*. From now on, we always assume that a measure space is completed.

2.3 Measurable Function and Integration

2.3.1 Measurable function

In measure theory, functions defined on a measure space are more interesting and important, as compared to measure space itself. Specially, only so-called measurable functions are useful.

Definition 2.3 (measurable function) Let $X : \Omega \mapsto R$ be a function defined on Ω. X is *measurable* if for $x \in R$, the set $\{\omega \in \Omega : X(\omega) \leq x\}$ is measurable, equivalently, belongs to $\mathcal{A}$. Especially, if the measure space is a probability measure space, X is called a *random variable*. †

Hence, for a measurable function, we can evaluate the size of the set such like $X^{-1}((-\infty, x])$. In fact, the following proposition concludes that for any Borel set $B \in \mathcal{B}$, $X^{-1}(B)$ is a measurable set in $\mathcal{A}$.

Proposition 2.4 If X is measurable, then for any $B \in \mathcal{B}$, $X^{-1}(B) = \{\omega : X(\omega) \in B\}$ is measurable. †

Proof We defined a class as below:

$$\mathcal{B}^* = \left\{B : B \subset R, X^{-1}(B) \text{ is measurable in } \mathcal{A}\right\}.$$

Clearly, $(-\infty, x] \in \mathcal{B}^*$. Furthermore, if $B \in \mathcal{B}^*$, then $X^{-1}(B) \in \mathcal{A}$. Thus, $X^{-1}(B^c) = \Omega - X^{-1}(B) \in \mathcal{A}$ then $B^c \in \mathcal{B}^*$. Moreover, if $B_1, B_2, ... \in \mathcal{B}^*$, then $X^{-1}(B_1), X^{-1}(B_2), ... \in \mathcal{A}$. Thus, $X^{-1}(B_1 \cup B_2 \cup ...) = X^{-1}(B_1) \cup X^{-1}(B_2) \cup ... \in \mathcal{A}$. So $B_1 \cup B_2 \cup ... \in \mathcal{B}^*$. We conclude that $\mathcal{B}^*$ is a σ-field. However, the Borel set $\mathcal{B}$ is the minimal σ-filed containing all intervals of the type $(-\infty, x]$. So $\mathcal{B} \subset \mathcal{B}^*$. Then for any Borel set B, $X^{-1}(B)$ is measurable in $\mathcal{A}$. †

One special example of a measurable function is a *simple function* defined as $\sum_{i=1}^n x_i I_{A_i}(\omega)$, where $A_i, i = 1, ..., n$ are disjoint measurable sets in $\mathcal{A}$. Here, $I_A(\omega)$ is the indicator function of A such that $I_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise. Note that the summation and maximum of a finite number of simple functions are still simple functions. More examples of measurable functions can be constructed from elementary algebra.

Proposition 2.5 Suppose that $\{X_n\}$ are measurable. Then so are $X_1 + X_2, X_1X_2, X_1^2$ and $\sup_n X_n$, $\inf_n X_n$, $\limsup_n X_n$ and $\liminf_n X_n$. †

Proof All can be verified using the following relationship:

$$\{X_1 + X_2 \leq x\} = \Omega - \{X_1 + X_2 > x\} = \Omega - \cup_{r \in Q} \{X_1 > r\} \cap \{X_2 > x - r\},$$

where Q is the set of all rational numbers. $\{X_1^2 \leq x\}$ is empty if $x < 0$ and is equal to $\{X_1 \leq \sqrt{x}\} - \{X_1 < -\sqrt{x}\}$. $X_1X_2 = \{(X_1 + X_2)^2 - X_1^2 - X_2^2\}/2$ so it is measurable. The remaining proofs can be seen from the following:

$$\left\{\sup_n X_n \leq x\right\} = \cap_n \{X_n \leq x\}.$$

$$\left\{\inf_n X_n \leq x\right\} = \left\{\sup_n(-X_n) \geq -x\right\}.$$

$$\left\{\limsup_n X_n \leq x\right\} = \cap_{r\in Q, r>0} \cup_{n=1}^{\infty} \cap_{k\geq n} \left\{X_k < x+r\right\}.$$

$$\liminf_n X_n = -\limsup_n(-X_n).$$

†

One important and fundamental fact for measurable function is given in the following proposition.

Proposition 2.6 For any measurable function $X \geq 0$, there exists an increasing sequence of simple functions $\{X_n\}$ such that $X_n(\omega)$ increases to $X(\omega)$ as n goes to infinity. †

Proof Define

$$X_n(\omega) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} I\{\frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n}\} + nI\left\{X(\omega) \geq n\right\}.$$

That is, we simply partition the range of X and assign the smallest value within each partition. Clearly, X_n is increasing over n. Moreover, if $X(\omega) < n$, then $|X_n(\omega) - X(\omega)| < \frac{1}{2^n}$. Thus, $X_n(\omega)$ converges to $X(\omega)$. †

This fact can be used to verify the measurability of many functions, for example, if g is a continuous function from R to R, then $g(X)$ is also measurable.

2.3.2 Integration of measurable function

Now we are ready to define the integration of a measurable function.

Definition 2.4 (i) For any simple function $X(\omega) = \sum_{i=1}^n x_i I_{A_i}(\omega)$, we define $\sum_{i=1}^n x_i \mu(A_i)$ as the *integral* of X with respect to measure μ, denoted as $\int X d\mu$.
(ii) For any $X \geq 0$, we define $\int X d\mu$ as

$$\int X d\mu = \sup_{Y \text{ is simple function, } 0\leq Y \leq X} \int Y d\mu.$$

(iii) For general X, let $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$. Then $X = X^+ - X^-$. If one of $\int X^+ d\mu$, $\int X^- d\mu$ is finite, we define $\int X d\mu = \int X^+ d\mu - \int X^- d\mu$. †

Particularly, we call X is *integrable* if $\int |X| d\mu = \int X^+ d\mu + \int X^- d\mu$ is finite. Note the definition (ii) is consistent with (i) when X itself is a simple function. When the measure space is a probability measure space and X is a random variable, $\int X d\mu$ is also called the *expectation* of X, denoted by $E[X]$.

Proposition 2.7 (i) For two measurable functions $X_1 \geq 0$ and $X_2 \geq 0$, if $X_1 \leq X_2$, then $\int X_1 d\mu \leq \int X_2 d\mu$.
(ii) For $X \geq 0$ and any sequence of simple functions Y_n increasing to X, $\int Y_n d\mu \to \int X d\mu$. †

Proof (i) For any simple function $0 \leq Y \leq X_1$, $Y \leq X_2$. Thus, $\int Y d\mu \leq \int X_2 d\mu$ by the definition of $\int X_2 d\mu$. We take the supreme over all the simple functions less than X_1 and obtain $\int X_1 d\mu \leq \int X_2 d\mu$.
(ii) From (i), $\int Y_n d\mu$ is increasing and bounded by $\int X d\mu$. It suffices to show that for any simple function $Z = \sum_{i=1}^m x_i I_{A_i}(\omega)$, where $\{A_i, 1 \leq i \leq m\}$ are disjoint measurable sets and $x_i > 0$, such that $0 \leq Z \leq X$, it holds

$$\lim_n \int Y_n d\mu \geq \sum_{i=1}^m x_i \mu(A_i).$$

We consider two cases. First, suppose $\int Z d\mu = \sum_{i=1}^m x_i \mu(A_i)$ is finite thus both x_i and $\mu(A_i)$ are finite. Fix an $\epsilon > 0$, let $A_{in} = A_i \cap \{\omega : Y_n(\omega) > x_i - \epsilon\}$. Since Y_n increases to X who is larger than or equal to x_i in A_i, A_{in} increases to A_i. Thus $\mu(A_{in})$ increases to $\mu(A_i)$ by Proposition 2.2. It yields that when n is large,

$$\int Y_n d\mu \geq \sum_{i=1}^m (x_i - \epsilon)\mu(A_i).$$

We conclude $\lim_n \int Y_n d\mu \geq \int Z d\mu - \epsilon \sum_{i=1}^m \mu(A_i)$. Then $\lim_n \int Y_n d\mu \geq \int Z d\mu$ by letting ϵ approach 0. Second, suppose $\int Z d\mu = \infty$ then there exists some i from $\{1, ..., m\}$, say 1, so that $\mu(A_1) = \infty$ or $x_1 = \infty$. Choose any $0 < x < x_1$ and $0 < y < \mu(A_1)$. Then the set $A_{1n} = A_1 \cap \{\omega : Y_n(\omega) > x\}$ increases to A_1. Thus when n large enough, $\mu(A_{1n}) > y$. We thus obtain $\lim_n \int Y_n d\mu \geq xy$. By letting $x \to x_1$ and $y \to \mu(A_1)$, we conclude $\lim_n \int Y_n d\mu = \infty$. Therefore, in either case, $\lim_n \int Y_n d\mu \geq \int Z d\mu$. †

Proposition 2.7 implies that, to calculate the integral of a non-negative measurable function X, we can choose any increasing sequence of simple functions $\{Y_n\}$ and the limit of $\int Y_n d\mu$ is the same as $\int X d\mu$. Particularly, such a sequence can chosen as constructed as Proposition 2.6; then

$$\int X d\mu = \lim_n \left\{ \sum_{k=1}^{n2^n-1} \frac{k}{2^n} \mu(\frac{k}{2^n} \leq X < \frac{k+1}{2^n}) + n\mu(X \geq n) \right\}.$$

Proposition 2.8 (Elementary Properties) Suppose $\int X d\mu$, $\int Y d\mu$ and $\int X d\mu + \int Y d\mu$ exit. Then
(i)

$$\int (X+Y) d\mu = \int X d\mu + \int Y d\mu, \quad \int cX d\mu = c \int X d\mu;$$

(ii) $X \geq 0$ implies $\int X d\mu \geq 0$; $X \geq Y$ implies $\int X d\mu \geq \int Y d\mu$; and $X = Y$ a.e., that is, $\mu(\{\omega : X(\omega) \neq Y(\omega)\}) = 0$, implies that $\int X d\mu = \int Y d\mu$;
(iii) $|X| \leq Y$ with Y integrable implies that X is integrable; X and Y are integrable implies that $X + Y$ is integrable.†

Proposition 2.8 can be proved using the definition. Finally, we give a few facts of computing integration without proof.

(a) Suppose $\mu^{\#}$ is a counting measure in $\Omega = \{x_1, x_2, ...\}$. Then for any measurable function g,

$$\int g d\mu^{\#} = \sum_i g(x_i).$$

(b) For any continuous function $g(x)$, which is also measurable in the Lebsgue measure space $(R, \mathcal{B}, \lambda)$, $\int g d\lambda$ is equal to the usual Riemann integral $\int g(x)dx$, whenever g is integrable.

(c) In a Lebsgue-stieljes measure space $(\Omega, \mathcal{B}, \lambda_F)$, where F is differentiable except discontinuous points $\{x_1, x_2, ...\}$, the integration of a continuous function $g(x)$ is given by

$$\int g d\lambda_F = \sum_i g(x_i)\left\{F(x_i) - F(x_i-)\right\} + \int g(x)f(x)dx,$$

where $f(x)$ is the derivative of $F(x)$.

2.3.3 Convergence of measurable functions

In this section, we provide some important theorems on how to take limits in the integration.

Theorem 2.2 (Monotone Convergence Theorem) If $X_n \geq 0$ and X_n increases to X, then $\int X_n d\mu \to \int X d\mu$. †

Proof Choose non-negative simple function X_{km} increasing to X_k as $m \to \infty$. Define $Y_n = \max_{k \leq n} X_{kn}$. $\{Y_n\}$ is an increasing series of simple functions and it satisfies

$$X_{kn} \leq Y_n \leq X_n, \quad \text{so} \int X_{kn} d\mu \leq \int Y_n d\mu \leq \int X_n d\mu.$$

By letting $n \to \infty$, we obtain

$$X_k \leq \lim_n Y_n \leq X, \quad \int X_k d\mu \leq \int \lim_n Y_n d\mu = \lim_n \int Y_n d\mu \leq \lim_n \int X_n d\mu,$$

where the equality holds since Y_n is simple function. By letting $k \to \infty$, we obtain

$$X \leq \lim_n Y_n \leq X, \quad \lim_k \int X_k d\mu \leq \int \lim_n Y_n d\mu \leq \lim_n \int X_n d\mu.$$

The result holds. †

Example 2.7 This example shows that the non-negative condition in the above theorem is necessary: let $X_n(x) = -I(x > n)/n$ be measurable function in the Lebesgue measure space. Clearly, X_n increases to zero but $\int X_n d\lambda = -\infty$.

Theorem 2.3 (Fatou's Lemma) If $X_n \geq 0$ then

$$\int \liminf_n X_n d\mu \leq \liminf_n \int X_n d\mu.$$

†

Proof Note

$$\liminf_n X_n = \sup_{n=1}^{\infty} \inf_{m \geq n} X_m.$$

Thus, the sequence $\{\inf_{m \geq n} X_m\}$ increases to $\liminf_n X_n$. By the Monotone Convergence Theorem,

$$\int \liminf_n X_n d\mu = \lim_n \int \inf_{m \geq n} X_m d\mu \leq \int X_n d\mu.$$

Take the lim inf on both sides and the theorem holds. †

The next theorem requires two more definitions.

Definition 2.5 A sequence X_n *converges almost everywhere* (a.e.) to X, denoted $X_n \to_{a.e.} X$, if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega - N$ where $\mu(N) = 0$. If μ is a probability, we write a.e. as a.s. (almost surely). A sequence X_n *converges in measure* to a measurable function X, denoted $X_n \to_\mu X$, if $\mu(|X_n - X| \geq \epsilon) \to 0$ for all $\epsilon > 0$. If μ is a probability measure, we say X_n *converges in probability* to X. †

The following proposition further justifies the convergence almost everywhere.

Proposition 2.9 Let $\{X_n\}$, X be finite measurable functions. Then $X_n \to_{a.e.} X$ if and only if for any $\epsilon > 0$,

$$\mu(\cap_{n=1}^{\infty} \cup_{m \geq n} \{|X_m - X| \geq \epsilon\}) = 0.$$

If $\mu(\Omega) < \infty$, then $X_n \to_{a.e.} X$ if and only if for any $\epsilon > 0$,

$$\mu(\cup_{m \geq n} \{|X_m - X| \geq \epsilon\}) \to 0.$$

†

Proof Note that

$$\{\omega : X_n(\Omega) \to X(\omega)\}^c = \cup_{k=1}^{\infty} \cap_{n=1}^{\infty} \cup_{m \geq n} \left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\}.$$

Thus, if $X_n \to_{a.e} X$, the measure of the left-hand side is zero. However, the right-hand side contains $\cap_{n=1}^{\infty} \cup_{m \geq n} \{|X_m - X| \geq \epsilon\}$ for any $\epsilon > 0$. The direction $\Rightarrow$ is proved. For the other direction, we choose $\epsilon = 1/k$ for any k, then by countable sub-additivity,

$$\mu(\cup_{k=1}^{\infty} \cap_{n=1}^{\infty} \cup_{m \geq n} \left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\})$$

$$\leq \sum_k \mu(\cap_{n=1}^{\infty} \cup_{m\geq n} \left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\}) = 0.$$

Thus, $X_n \to_{a.e.} X$. When $\mu(\Omega) = 1$, the latter holds by Proposition 2.2. †

The following proposition describes the relationship between the convergence almost everywhere and the convergence in measure.

Proposition 2.10 Let X_n be finite a.e.
(i) If $X_n \to_{\mu} X$, then there exists a subsequence $X_{n_k} \to_{a.e} X$.
(ii) If $\mu(\Omega) < \infty$ and $X_n \to_{a.e.} X$, then $X_n \to_{\mu} X$. †

Proof (i) For any k, there exists some n_k such that

$$P(|X_{n_k} - X| \geq 2^{-k}) < 2^{-k}.$$

Then

$$\mu(\cup_{m\geq k} \{|X_{n_m} - X| \geq \epsilon\}) \leq \mu(\cup_{m\geq k} \left\{|X_{n_m} - X| \geq 2^{-k}\right\}) \leq \sum_{m\geq k} 2^{-m} \to 0.$$

Thus from the previous proposition, $X_{n_k} \to_{a.e} X$.
(ii) is direct from the second part of Proposition 2.9. †

Example 2.8 Let $X_{2^n+k} = I(x \in [k/2^n, (k+1)/2^n)), 0 \leq k < 2^n$ be measurable functions in the Lebesgue measure space. Then it is easy to see $X_n \to_{\lambda} 0$ but does not converge to zero almost everywhere. While, there exists a subsequence converging to zero almost everywhere.

Example 2.9 In Example 2.7, $n^2 X_n \to_{a.e.} 0$ but $\lambda(|X_n| > \epsilon) \to \infty$. This example shows that $\mu(\Omega) < \infty$ in (ii) of Proposition 2.10 is necessary.

We now state the third important theorem.

Theorem 2.4 (Dominated Convergence Theorem) If $|X_n| \leq Y$ a.e. with Y integrable, and if $X_n \to_{\mu} X$ (or $X_n \to_{a.e.} X$), then $\int |X_n - X| d\mu \to 0$ and $\lim \int X_n d\mu = \int X d\mu$. †

Proof First, assume $X_n \to_{a.e} X$. Define $Z_n = 2Y - |X_n - X|$. Clearly, $Z_n \geq 0$ and $Z_n \to 2Y$. By the Fatou's lemma, we have

$$\int 2Y d\mu \leq \liminf_n \int (2Y - |X_n - X|) d\mu.$$

That is, $\limsup_n \int |X_n - X| d\mu \leq 0$ and the result holds. If $X_n \to_{\mu} X$ and the result does not hold for some subsequence of X_n, by Proposition 2.10, there exits a further sub-sequence converging to X almost surely. However, the result holds for this further subsequence. We obtain the contradiction. †

The existence of the dominating function Y is necessary, as seen in the counter example in Example 2.7. Finally, the following result describes the interchange between integral and limit or derivative.

Theorem 2.5 (Interchange of Integral and Limit or Derivatives) Suppose that $X(\omega, t)$ is measurable for each $t \in (a, b)$.
(i) If $X(\omega, t)$ is a.e. continuous in t at t_0 and $|X(\omega, t)| \leq Y(\omega), a.e.$ for $|t - t_0| < \delta$ with Y integrable, then

$$\lim_{t \to t_0} \int X(\omega, t) d\mu = \int X(\omega, t_0) d\mu.$$

(ii) Suppose $\frac{\partial}{\partial t} X(\omega, t)$ exists for a.e. ω, all $t \in (a, b)$ and $|\frac{\partial}{\partial t} X(\omega, t)| \leq Y(\omega), a.e.$ for all $t \in (a, b)$ with Y integrable. Then

$$\frac{\partial}{\partial t} \int X(\omega, t) d\mu = \int \frac{\partial}{\partial t} X(\omega, t) d\mu.$$

†

Proof (i) follows from the Dominated Convergence Theorem and the subsequence argument. (ii) can be seen from the following:

$$\frac{\partial}{\partial t} \int X(\omega, t) d\mu = \lim_{h \to 0} \int \frac{X(\omega, t + h) - X(\omega, t)}{h} d\mu.$$

Then from the conditions and (i), such a limit can be taken within the integration. †

2.4 Fubini Integration and Radon-Nikodym Derivative

2.4.1 Product of measures and Fubini-Tonelli theorem

Suppose that $(\Omega_1, \mathcal{A}_1, \mu_1)$ and $(\Omega_2, \mathcal{A}_2, \mu_2)$ are two measure spaces. Now we consider the product set $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$. Correspondingly, we define a class

$$\{A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}.$$

$A_1 \times A_2$ is called a *measurable rectangle set.* However, the above class is not a σ-field. We thus construct the σ-filed based on this class and denote

$$\mathcal{A}_1 \times \mathcal{A}_2 = \sigma(\{A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}).$$

To define a measure on this σ-field, denoted $\mu_1 \times \mu_2$, we can first define it on any rectangle set

$$(\mu_1 \times \mu_2)(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2).$$

Then $\mu_1 \times \mu_2$ is extended to all sets in the $\mathcal{A}_1 \times \mathcal{A}_2$ by the Caratheodory Extension theorem.

One simple example is the Lebesgue measure in a multi-dimensional real space R^k. We let $(R, \mathcal{B}, \lambda)$ be the Lebesgue measure in one-dimensional real space. Then we can use the above procedure to define $\lambda \times ... \times \lambda$ as a measure on $R^k = R \times ... \times R$. Clearly, for each cube in R^k, this measure gives the same value as the volume of the cube. In fact, this measure agrees with λ^k defined in Example 2.3.

With the product measure, we can start to discuss the integration with respect to this measure. Let $X(\omega_1, \omega_2)$ be the measurable function on the measurable space $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times$

$\mathcal{A}_2, \mu_1 \times \mu_2)$. The integration of X is denoted as $\int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2) d(\mu_1 \times \mu_2)$. In the case when the measurable space is real space, this integration is simply bivariate integration such like $\int_{R^2} f(x, y) dx dy$. As in the calculus, we are often concerned about whether we can integrate over x first then y or we can integrate y first then x. The following theorem gives the condition of changing the order of integration.

Theorem 2.6 (Fubini-Tonelli Theorem) Suppose that $X : \Omega_1 \times \Omega_2 \to R$ is $\mathcal{A}_1 \times \mathcal{A}_2$ measurable and $X \geq 0$. Then

$$\int_{\Omega_1} X(\omega_1, \omega_2) d\mu_1 \text{ is } \mathcal{A}_2 \text{ measurable,}$$

$$\int_{\Omega_2} X(\omega_1, \omega_2) d\mu_2 \text{ is } \mathcal{A}_1 \text{ measurable,}$$

and

$$\int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2) d(\mu_1 \times \mu_2) = \int_{\Omega_1} \left\{ \int_{\Omega_2} X(\omega_1, \omega_2) d\mu_2 \right\} d\mu_1 = \int_{\Omega_2} \left\{ \int_{\Omega_1} X(\omega_1, \omega_2) d\mu_1 \right\} d\mu_2.$$

†

As a corollary, suppose X is not necessarily non-negative but we can write $X = X^+ - X^-$. Then the above results hold for X^+ and X^-. Thus, if $\int_{\Omega_1 \times \Omega_2} |X(\omega_1, \omega_2)| d(\mu_1 \times \mu_2)$ is finite, then the above results hold.

Proof Suppose that we have shown the theorem holds for any indicator function $I_B(\omega_1, \omega_2)$, where $B \in \mathcal{A}_1 \times \mathcal{A}_2$. We construct a sequence of simple functions, denoted as $\tilde{X}_n$, increases to X. Clearly, $\int_{\Omega_1} \tilde{X}_n(\omega_1, \omega_2) d\mu_1$ is measurable and

$$\int_{\Omega_1 \times \Omega_2} \tilde{X}_n(\omega_1, \omega_2) d(\mu_1 \times \mu_2) = \int_{\Omega_2} \int_{\Omega_1} \left\{ \tilde{X}_n(\omega_1, \omega_2) d\mu_1 \right\} d\mu_2.$$

By the monotone convergence theorem, $\int_{\Omega_1} \tilde{X}_n(\omega_1, \omega_2) d\mu_1$ increases to $\int_{\Omega_1} X(\omega_1, \omega_2) d\mu_1$ almost everywhere. Further applying the monotone convergence theorem to both sides of the above equality, we obtain

$$\int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2) d(\mu_1 \times \mu_2) = \int_{\Omega_2} \int_{\Omega_1} \left\{ X(\omega_1, \omega_2) d\mu_1 \right\} d\mu_2.$$

Similarly,

$$\int_{\Omega_1 \times \Omega_2} X(\omega_1, \omega_2) d(\mu_1 \times \mu_2) = \int_{\Omega_1} \int_{\Omega_2} \left\{ X(\omega_1, \omega_2) d\mu_2 \right\} d\mu_1.$$

It remains to show $I_B(\omega_1, \omega_2)$ satisfies the theorem's results for $B \in \mathcal{A}_1 \times \mathcal{A}_2$.

To this end, we define what is called a monotone class: $\mathcal{M}$ is a monotone class if for any increasing sequence of sets $B_1 \subseteq B_2 \subseteq B_3 \ldots$ in the class, $\cup_i B_i$ belongs to $\mathcal{M}$. We then let $\mathcal{M}_0$ be the minimal monotone class in $\mathcal{A}_1 \times \mathcal{A}_2$ containing all the rectangles. The existence of such minimal class can be proved using the same construction as Proposition 2.3 and noting that $\mathcal{A}_1 \times \mathcal{A}_2$ itself is a monotone class. We show that $\mathcal{M}_0 = \mathcal{A}_1 \times \mathcal{A}_2$.

(a) $\mathcal{M}_0$ is a field: for $A, B \in \mathcal{M}_0$, it suffices to show that $A \cap B, A \cap B^c, A^c \cap B \in \mathcal{M}_0$. We consider

$$\mathcal{M}_A = \{B \in \mathcal{M}_0 : A \cap B, A \cap B^c, A^c \cap B \in \mathcal{M}_0\}.$$

It is straightforward to see that if A is a rectangle, then $B \in \mathcal{M}_A$ for any rectangle B and that $\mathcal{M}_A$ is a monotone class. Thus, $\mathcal{M}_A = \mathcal{M}_0$ for A being a rectangle. For general A, the previous result implies that all the rectangles are in $\mathcal{M}_A$. Clearly, $\mathcal{M}_A$ is a monotone class. Therefore, $\mathcal{M}_A = \mathcal{M}_0$ for any $A \in \mathcal{M}_0$. That is, for $A, B \in \mathcal{M}_0$, $A \cap B, A \cap B^c, A^c \cap B \in \mathcal{M}_0$.

(b) $\mathcal{M}_0$ is a σ-field. For any $B_1, B_2, ... \in \mathcal{M}_0$, we can write $\cup_i B_i$ as the union of increasing sets $B_1, B_1 \cup B_2, ...$ Since each set in the sequence is in $\mathcal{M}_0$ and $\mathcal{M}_0$ is a monotone class, $\cup_i B_i \in \mathcal{M}_0$. Thus, $\mathcal{M}_0$ is a σ-field so it must be equal to $\mathcal{A}_1 \times \mathcal{A}_2$.

Now we come back to show that for any $B \in \mathcal{A}_1 \times \mathcal{A}_2$, I_B satisfies the equality in Theorem 2.6. To do this, we define a class

$$\{B : B \in \mathcal{A}_1 \times \mathcal{A}_2 \text{ is measurable and } I_B \text{ satifies the equality in Theorem 2.6}\}.$$

Clearly, the class contains all the rectangles. Second, the class is a monotone class: suppose $B_1, B_2, ...$ is an increasing sequence of sets in the class, we apply the monotone convergence theorem to

$$\int_{\Omega_1 \times \Omega_2} I_{B_i} d(\mu_1 \times \mu_2) = \int_{\Omega_2} \left\{ \int_{\Omega_1} I_{B_i} d\mu_1 \right\} d\mu_2 = \int_{\Omega_1} \left\{ \int_{\Omega_2} I_{B_i} d\mu_2 \right\} d\mu_1$$

and note $I_{B_i} \to I_{\cup_i B_i}$. We conclude that $\cup_i B_i$ is also in the defined class. Therefore, from the previous result about the relationship between the monotone class and the σ-field, we obtain that the defined class should be the same as $\mathcal{A}_1 \times \mathcal{A}_2$. †

Example 2.10 Let $(\Omega, 2^\Omega, \mu^\#)$ be a counting measure space where $\Omega = \{1, 2, 3, ...\}$ and $(R, \mathcal{B}, \lambda)$ be the Lebesgue measure space. Define $f(x, y)$ be a bivariate function in the product of these two measure space as $f(x, y) = I(0 \leq x \leq y) \exp\{-y\}$. To evaluate the integral $f(x, y)$, we use the Fubini-Tonelli theorem and obtain

$$\int_{\Omega \times R} f(x, y) d\{\mu^\# \times \lambda\} = \int_\Omega \{ \int_R f(x, y) d\lambda(y) \} d\mu^\#(x) = \int_\Omega \exp\{-x\} d\mu^\#(x)$$

$$= \sum_{n=1}^\infty \exp\{-n\} = 1/(e-1).$$

2.4.2 Absolute continuity and Radon-Nikodym derivative

Let $(\Omega, \mathcal{A}, \mu)$ be a measurable space and let X be a non-negative measurable function on Ω. We define a set function ν as

$$\nu(A) = \int_A X d\mu = \int I_A X d\mu$$

for each $A \in \mathcal{A}$. It is easy to see that ν is also a measure on $(\Omega, \mathcal{A})$. X can be regarded as the derivative of the measure ν with respect μ (one can think about an example in real space).

However, one question is the opposite direction: if both μ and ν are the measures on $(\Omega, \mathcal{A})$, can we find a measurable function X such that the above equation holds? To answer this, we need to introduce the definition of absolute continuity.

Definition 2.6 If for any $A \in \mathcal{A}$, $\mu(A) = 0$ implies that $\nu(A) = 0$, then ν is said to be *absolutely continuous* with respect to μ, and we write $\nu \prec\prec \mu$. Sometimes it is also said that ν is *dominated* by μ. †

One equivalent condition to the above the condition is the following lemma.

Proposition 2.11 Suppose $\nu(\Omega) < \infty$. Then $\nu \prec\prec \mu$ if and only if for any $\epsilon > 0$, there exists a δ such that $\nu(A) < \epsilon$ whenever $\mu(A) < \delta$. †

Proof " $\Leftarrow''$ is clear. To prove " $\Rightarrow''$, we use the contradiction. Suppose there exists ϵ and a set A_n such that $\nu(A_n) > \epsilon$ and $\mu(A_n) < n^{-2}$. Since $\sum_n \mu(A_n) < \infty$, we have

$$\mu(\limsup_n A_n) \leq \sum_{m \geq n} \mu(A_n) \to 0.$$

Thus $\mu(\limsup_n A_n) = 0$. However, $\nu(\limsup_n A_n) = \lim_n \nu(\cup_{m \geq n} A_m) \geq \limsup_n \nu(A_n) \geq \epsilon$. It is a contradiction. †

The following Radon-Nikodym theorem says that if ν is dominated by μ, then a measurable function X satisfying the equation exists. Such X is called the *Radon-Nikodym derivative* of ν with respect μ, denoted by $d\nu/d\mu$.

Theorem 2.7 (Radon-Nikodym theorem) Let $(\Omega, \mathcal{A}, \mu)$ be a σ-finite measure space, and let ν be a measurable on $(\Omega, \mathcal{A})$ with $\nu \prec\prec \mu$. Then there exists a measurable function $X \geq 0$ such that $\nu(A) = \int_A X d\mu$ for all $A \in \mathcal{A}$. X is unique in the sense that if another measurable function Y also satisfies the equation, then $X = Y$, a.e. †

Before proving Theorem 2.7, we need the following Hahn decomposition theorem for any additive set function with real values, $\phi(A)$, which is defined on a measurable space $(\Omega, \mathcal{A})$ such that for countable disjoint sets $A_1, A_2, ...,$

$$\phi(\cup_n A_n) = \sum_n \phi(A_n).$$

The main difference from the usual measure definition is that $\phi(A)$ can be negative and must be finite.

Proposition 2.12 (Hahn Decomposition) For any additive set function ϕ, there exist disjoint sets A^+ and A^- such that $A^+ \cup A^- = \Omega$, $\phi(E) \geq 0$ for any $E \subset A^+$ and $\phi(E) \leq 0$ for any $E \subset A^-$. A^+ is called positive set and A^- is called negative set of ϕ. †

Proof Let $\alpha = \sup\{\phi(A) : A \in \mathcal{A}\}$. Suppose there exists a set A^+ such that $\phi(A^+) = \alpha < \infty$. Let $A^- = \Omega - A^+$. If $E \subset A^+$ and $\phi(E) < 0$, then $\phi(A^+ - E) \geq \alpha - \phi(E) > \alpha$, an impossibility. Thus, $\phi(E) \geq 0$. Similarly, for any $E \subset A^-$, $\phi(E) \leq 0$.

It remains to construct such A^+. Choose A_n such that $\phi(A_n) \to \alpha$. Let $A = \cup_n A_n$. For each n, we consider all possible intersection of $A_1, ..., A_n$, denoted by $\mathcal{B}_n = \{B_{ni} : 1 \leq i \leq 2^n\}$. Then the collection of $\mathcal{B}_n$ is a partition of A. Let C_n be the union of those B_{ni} in $\mathcal{B}_n$ such that $\phi(B_{ni}) > 0$. Then $\phi(A_n) \leq \phi(C_n)$. Moreover, for any $m < n$, $\phi(C_m \cup ... \cup C_n) \geq \phi(C_m \cup ... \cup C_{n-1})$. Let $A^+ = \cap_{m=1}^\infty \cup_{n \geq m} C_n$. Then $\alpha = \lim_m \phi(A_m) \leq \lim_m \phi(\cup_{n\geq m} C_n) = \phi(A^+)$. Then $\phi(A^+) = \alpha$. †

We now start to prove Theorem 2.7.

Proof We first show that this holds if $\mu(\Omega) < \infty$. Let Ξ be the class of non-negative functions g such that $\int_E g d\mu \leq \nu(E)$. Clearly, $0 \in \Xi$. If g and g' are in Ξ, then

$$\int_E \max(g, g') d\mu = \int_{E\cap\{g\geq g'\}} g d\mu + \int_{E\cap\{g<g'\}} g' d\mu \leq \int_{E\cap\{g\geq g'\}} d\nu + \int_{E\cap\{g<g'\}} d\nu = \nu(E).$$

Thus, $\max(g, g') \in \Xi$. Moreover, if g_n increases to g and $g_n \in \Xi$, then by the monotone convergence theorem, $g \in \Xi$.

Let $\alpha = \sup_{g\in\Xi} \int g d\mu$ then $\alpha \leq \nu(\Omega)$. Choose g_n in Ξ such that $\int g_n d\mu > \alpha - n^{-1}$. Define $f_n = \max(g_1, ..., g_n) \in \Xi$ and f_n increases to $f \in \Xi$. We have $\int f d\mu = \alpha$.

Define a measure $0 \leq \nu_s(E) = \nu(E) - \int_E f d\mu$. We will show that there exists set S_μ and S_ν such that $\mu(\Omega - S_\mu) = 0$, $\nu_s(\Omega - S_\nu) = 0$, and $S_\mu \cap S_\nu = \emptyset$. If this is true, then since $\nu \prec\prec \mu$, $\nu_s(\Omega - S_\mu) \leq \nu(\Omega - S_\mu) = 0$. Thus,

$$\nu_s(E) \leq \nu_s(E \cap (\Omega - S_\mu)) + \nu_s(E \cap (\Omega - S_\nu)) = 0.$$

This gives that $\nu(E) = \int_E f d\mu$. We prove the previous statement by contradiction. Let $A_n^+ \cup A_n^-$ be a Hahn decomposition for the the set function $\nu_s - n^{-1}\mu$ and let $M = \cup_n A_n^+$ so $M^c = \cap_n A_n^-$. Since $\nu_s(M^c) - n^{-1}\mu(M^c) \leq \nu_s(A_n^-) - n^{-1}\mu(A_n^-) \leq 0$, we have $\nu_s(M^c) \leq n^{-1}\mu(M^c) \to 0$. Then $\mu(M)$ must be positive. Therefore, there exists some $A = A_n^+$ such that $\mu(A) > 0$ and $\nu_s(E) \geq n^{-1}\mu(E)$ for any $E \subset A$. For such A, we have that for $\epsilon = 1/n$,

$$\begin{aligned}
\int_E (f + \epsilon I_A) d\mu &= \int_E f d\mu + \epsilon\mu(E \cap A) \\
&\leq \int_E f d\mu + \nu_s(E \cap A) \\
&\leq \int_{E\cap A} f d\mu + \nu_s(E \cap A) + \int_{E-A} f d\mu \\
&\leq \nu(E \cap A) + \int_{E-A} f d\mu \leq \nu(E \cap A) + \nu(E - A) = \nu(E).
\end{aligned}$$

In other words, $f + \epsilon I_A$ is in Ξ. However, $\int (f + \epsilon I_A) d\mu = \alpha + \epsilon\mu(A) > \alpha$. We obtain the contradiction.

We have proved the theorem for $\mu(\Omega) < \infty$. If μ is countably finite, there exists countable decomposition of Ω into $\{B_n\}$ such that $\mu(B_n) < \infty$. For the measures $\mu_n(A) = \mu(A \cap B_n)$ and $\nu_n(A) = \nu(A \cap B_n)$, $\nu_n \prec\prec \mu_n$ so we can find non-negative f_n such that

$$\nu(A \cap B_n) = \int_{A\cap B_n} f_n d\mu.$$

Then $\nu(A) = \sum_n \nu(A \cap B_n) = \int_A \sum_n f_n I_{B_n} d\mu$.

The function f satisfying the result must be unique almost everywhere. If two f_1 ad f_2 satisfy that $\int_A f_1 d\mu = \int_A f_2 d\mu$ then after choosing $A = \{f_1 - f_2 > 0\}$ and $A = \{f_1 - f_2 < 0\}$, we obtain $f_1 = f_2$ almost everywhere. †

Using the Radon-Nikodym derivative, we can transform the integration with respect to the measure μ to the integration with respect to the measure ν.

Proposition 2.13 Suppose ν and μ are σ-finite measure defined on a measure space $(\Omega, \mathcal{A})$ with $\nu \prec\prec \mu$, and suppose Z is a measurable function such that $\int Z d\nu$ is well defined. Then for any $A \in \mathcal{A}$,

$$\int_A Z d\nu = \int_A Z \frac{d\nu}{d\mu} d\mu.$$

†

Proof (i) If $Z = I_B$ where $B \in \mathcal{A}$, then

$$\int_A Z d\nu = \nu(A \cap B) = \int_{A \cap B} \frac{d\nu}{d\mu} d\mu = \int_A I_B \frac{d\nu}{d\mu} d\mu.$$

The result holds.
(ii) If $Z \geq 0$, we can find a sequence of simple function Z_n increasing to Z. Clearly, for Z_n,

$$\int_A Z_n d\nu = \int_A Z_n \frac{d\nu}{d\mu} d\mu.$$

Take limits on both sides and apply the monotone convergence theorem. We obtain the result.
(iii) For any Z, we write $Z = Z^+ - Z^-$. Then both Z^+ and Z^- are integrable. Thus,

$$\int Z d\nu = \int Z^+ d\nu - \int Z^- d\nu = \int Z^+ \frac{d\nu}{d\mu} d\mu - \int Z^- \frac{d\nu}{d\mu} d\mu = \int Z \frac{d\nu}{d\mu} d\mu.$$

†

2.4.3 X-induced measure

Let X be a measurable function defined on $(\Omega, \mathcal{A}, \mu)$. Then for any $B \in \mathcal{B}$, since $X^{-1}(B) \in \mathcal{A}$, we can define a set function on all the Borel sets as

$$\mu_X(B) = \mu(X^{-1}(B)).$$

Such μ_X is called a *measure induced by* X. Hence, we obtain a measure in the Borel σ-field $(R, \mathcal{B}, \mu_X)$.

Suppose that $(R, \mathcal{B}, \nu)$ is another measure space (often the counting measure or the Lebesgue measure) and μ_X is dominated by ν with the derivative f. Then f is called the *density of X* *with respect to the dominating measure ν.* Furthermore, we obtain that for any measurable function g from R to R,

$$\int_\Omega g(X(\omega))d\mu(\omega) = \int_R g(x)d\mu_X(x) = \int_R g(x)f(x)d\nu(x).$$

That is, the integration of $g(X)$ on the original measure space Ω can be transformed as the integration of $g(x)$ on R with respect to the induced-measure μ_X and can be further transformed as the integration of $g(x)f(x)$ with respect to the dominating measure ν.

When $(\Omega, \mathcal{A}, \mu) = (\Omega, \mathcal{A}, P)$ is a probability space, the above interpretation has a special meaning: X is now a random variable then the above equation becomes

$$E[g(X)] = \int_R g(x)f(x)d\nu(x).$$

We immediately recognize that $f(x)$ is the density function of X with respect to the dominating measure ν. Particularly, if ν is the counting measure, $f(x)$ is in fact the probability mass function; if ν is the Lebesgue measure, $f(x)$ is the probability density function in the usual sense. This fact has an important implication: any expectations regarding random variable X can be computed via its probability mass function or density function without referral to whatever probability measure space X is defined on. This is the reason why in most of statistical framework, we seldom mention the underlying measure space while only give either the probability mass function or the probability density function.

2.5 Probability Measure

2.5.1 Parallel definitions

Already discussed before, a probability measure space $(\Omega, \mathcal{A}, P)$ satisfies that $P(\Omega) = 1$ and random variable (or random vector in multi-dimensional real space) X is a measurable function on this space. The integration of X is equivalent to the expectation. The density or the mass function of X is the Radon-Nikydom derivative of the X-induced measure with respect to the Lebesgue measure or the counting measure in real space. By using the mass function or density function, statisticians unconsciously ignore the underlying probability measure space $(\Omega, \mathcal{A}, P)$. However, it is important for readers to keep in mind that whenever a density function or mass function is referred, we assume that above procedure has been worked out for some probability space.

Recall that $F(x) = P(X \leq x)$ is the cumulative distribution function of X. Clearly, $F(x)$ is a nondecreasing function with $F(-\infty) = 0$ and $F(\infty) = 1$. Moreover, $F(x)$ is right-continuous, meaning that $F(x_n) \to F(x)$, if x_n decreases to x. Interestingly, we can show that μ_F, the Lebesgue-Stieljes measure generated by F, is exactly the same measure as the one induced by X, i.e., P_X.

Since a probability measure space is a special case of general measure space, all the properties for the general measure space including the monotone convergence theorem, the Fatou's lemma, the dominating convergence theorem, and the Fubini-Tonelli theorem apply.

2.5.2 Conditional expectation and independence

Nevertheless, there are some features only specific to probability measure, which distinguish probability theory from general measure theory. Two of these important features are conditional probability and independence. We describe them in the following text.

In a probability measure space $(\Omega, \mathcal{A}, P)$, we know the conditional probability of an event A given another event B is defined as $P(A|B) = P(A \cap B)/P(B)$ and $P(A|B^c) = P(A \cap B^c)/P(B^c)$. This means: if B occurs, then the probability that A occurs is $P(A|B)$; if B does not occur, then the probability that A occurs if $P(A|B^c)$. Thus, such a conditional distribution can be thought as a measurable function assigned to the σ-field $\{\emptyset, B, B^c, \Omega\}$, which is equal

$$P(A|B)I_B(\omega) + P(A|B^c)I_{B^c}(\omega).$$

Such a simple example in fact characterizes the essential definition of conditional probability. Let $\aleph$ be the sub-σ-filed of $\mathcal{A}$. For any $A \in \mathcal{A}$, the *conditional probability* of A given $\aleph$ is a measurable function on $(\Omega, \aleph)$, denoted $P(A|\aleph)$, and satisfies that
(i) $P(A|\aleph)$ is measurable in $\aleph$ and integrable;
(ii) For any $G \in \aleph$,

$$\int_G P(A|\aleph)dP = P(A \cap G).$$

Theorem 2.8 (Existence and Uniqueness of Conditional Probability Function) The measurable function $P(A|\aleph)$ exists and is unique in the sense that any two functions satisfying (i) and (ii) are the same almost surely. †

Proof In the probability space $(\Omega, \aleph, P)$, we define a set function ν on $\aleph$ such that $\nu(G) = P(A \cap G)$ for any $G \in \aleph$. It can easily show ν is a measure and $P(G) = 0$ implies that $\nu(G) = 0$. Thus $\nu \prec\prec P$. By the Radon-Nikodym theorem, there exits a $\aleph$-measurable function X such that

$$\nu(G) = \int_G X dP.$$

Thus X satisfies the properties (i) and (ii). Suppose X and Y both are measurable in $\aleph$ and $\int_G X dP = \int_G Y dP$ for any $G \in \aleph$. That is, $\int_G (X - Y)dP = 0$. Particularly, we choose $G = \{X - Y \geq 0\}$ and $G = \{X - Y < 0\}$. We then obtain $\int |X - Y|dP = 0$. So $X = Y$, a.s. †

Some properties of the conditional probability $P(\cdot|\aleph)$ are the following.

Theorem 2.9 $P(\emptyset|\aleph) = 0, P(\Omega|\aleph) = 1$ a.e. and

$$0 \leq P(A|\aleph) \leq 1$$

for each $A \in \mathcal{A}$. if $A_1, A_2, ...$ is finite or countable sequence of disjoint sets in $\mathcal{A}$, then

$$P(\cup_n A_n|\aleph) = \sum_n P(A_n|\aleph).$$

†

The properties can be verified directly from the definition. Now we define the *conditional expectation* of a integrable random variable X given $\aleph$, denoted $E[X|\aleph]$, as
(i) $E[X|\aleph]$ is measurable in $\aleph$ and integrable;
(ii) For any $G \in \aleph$,

$$\int_G E[X|\aleph]dP = \int_G XdP,$$

equivalently; $E\left[E[X|\aleph]I_G\right] = E[XI_G], a.e.$

The existence and the uniqueness of $E[X|\aleph]$ can be shown similar to Theorem 2.8. The following properties are fundamental.

Theorem 2.10 Suppose X, Y, X_n are integrable.
(i) If $X = a$ a.s., then $E[X|\aleph] = a$.
(ii) For constants a and b, $E[aX + bY|\aleph] = aE[X|\aleph] + b[Y|\aleph]$.
(iii) If $X \leq Y$ a.s., then $E[X|\aleph] \leq E[Y|\aleph]$.
(iv) $|E[X|\aleph]| \leq E[|X||\aleph]$.
(v) If $\lim_n X_n = X$ a.s., $|X_n| \leq Y$ and Y is integrable, then $\lim_n E[X_n|\aleph] = E[X|\aleph]$.
(vi) If X is measurable in $\aleph$, then

$$E[XY|\aleph] = XE[Y|\aleph].$$

(vii) For two sub-σ fields $\aleph_1$ and $\aleph_2$ such that $\aleph_1 \subset \aleph_2$,

$$E\left[E[X|\aleph_2]|\aleph_1\right] = E[X|\aleph_1].$$

(viii) $P(A|\aleph) = E[I_A|\aleph]$. †

Proof (i)-(iv) be shown directly using the definition. To prove (v), we consider $Z_n = \sup_{m\geq n} |X_m - X|$. Then Z_n decreases to 0. From (iii), we have

$$|E[X_n|\aleph] - E[X|\aleph]| \leq E[Z_n|\aleph].$$

On the other hand, $E[Z_n|\aleph]$ decreases to a limit $Z \geq 0$. The result holds if we can show $Z = 0$ a.s. Note $E[Z_n|\aleph] \leq E[2Y|\aleph]$, by the dominated convergence theorem,

$$E[Z] = \int E[Z|\aleph]dP \leq \int E[Z_n|\aleph]dP \to 0.$$

Thus $Z = 0$ a.s.

To see (vi) holds, we first show it holds for a simple function $X = \sum_i x_i I_{B_i}$ where B_i are disjoint set in $\aleph$. For any $G \in \aleph$,

$$\int_G E[XY|\aleph]dP = \int_G XYdP = \sum_i x_i \int_{G\cap B_i} YdP = \sum_i x_i \int_{G\cap B_i} E[Y|\aleph]dP = \int_G XE[Y|\aleph]d.$$

Hence, $E[XY|\aleph] = XE[Y|\aleph]$. For any X, using the previous construction, we can find a sequence of simple functions X_n converging to X and $|X_n| \leq |X|$. Then we have

$$\int_G X_n YdP = \int_G X_n E[Y|\aleph]dP.$$

Note that $|X_nE[Y|\aleph]| = |E[X_nY|\aleph]| \leq E[|XY||\aleph]$. Taking limits on both sides and from the dominated convergence theorem, we obtain

$$\int_G XYdP = \int_G XE[Y|\aleph]dP.$$

Then $E[XY|\aleph] = XE[Y|\aleph]$.

For (vii), for any $G \in \aleph_1 \subset \aleph_2$, it is clear form that

$$\int_G E[X|\aleph_2]dP = \int_G XdP = \int_G E[X|\aleph_1]dP.$$

(viii) is clear from the definition of the conditional probability. †

How can we relate the above conditional probability and conditional expectation given a sub-σ field to the conditional distribution or density of X given Y? In R^2, suppose (X,Y) has joint density function $f(x,y)$ then it is known that the conditional density of X given $Y = y$ is equal to $f(x,y)/\int_x f(x,y)dx$ and the conditional expectation of X given $Y = y$ is equal to $\int_x xf(x,y)dx/\int_x f(x,y)dx$. To recover these formulae using the current definition, we define $\aleph = \sigma(Y)$, the σ-field generated by the class $\{\{Y \leq y\} : y \in R\}$. Then we can define the conditional probability $P(X \in B|\aleph)$ for any B in $(R, \mathcal{B})$. Since $P(X \in B|\aleph)$ is measurable in $\sigma(Y)$, $P(X \in B|\aleph) = g(B,Y)$ where $g(B,\cdot)$ is a measurable function. For any $\{Y \leq y\} \in \aleph$,

$$\int_{Y\leq y_0} P(X \in B|\aleph)dP = \int I(y \leq y_0)g(B,y)f_Y(y)dy = P(X \in B, Y \leq y_0)$$

$$= \int I(y \leq y_0)\int_B f(x,y)dxdy.$$

Differentiate with respect to y_0, we have $g(B,y)f_Y(y) = \int_B f(x,y)dx$. Thus,

$$P(X \in B|\aleph) = \int_B f(x|y)dx.$$

Thus, we note that the conditional density of $X|Y = y$ is in fact the density function of the conditional probability $P(X \in \cdot|\aleph)$ with respect to the Lebesgue measure.

On the other hand, $E[X|\aleph] = g(Y)$ for some measurable function $g(\cdot)$. Note that

$$\int I(Y \leq y_0)E[X|\aleph]dP = \int I(y \leq y_0)g(y)f_Y(y)dy = E[XI(Y \leq y_0)] = \int I(y \leq y_0)xf(x,y)dxdy.$$

We obtain $g(y) = \int xf(x,y)dx/\int f(x,y)dx$. Then $E[X|\aleph]$ is the same as the conditional expectation of X given $Y = y$.

Finally, we give the definition of independence: Two measurable sets or events A_1 and A_2 in $\mathcal{A}$ are *independent* if $P(A \cap B) = P(A)P(B)$. For two random variables X and Y, X and Y are said to independent if for any Borel sets B_1 and B_2, $P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2)$. In terms of conditional expectation, X is independent of Y implies that for any measurable function g, $E[g(X)|Y] = E[g(X)]$.

READING MATERIALS: You should read Lehmann and Casella, Sections 1.2 and 1.3. You may read Lehmann *Testing Statistical Hypotheses*, Chapter 2.

PROBLEMS

1. Let $\mathcal{O}$ be the class of all open sets in R. Show that the Borel σ-field $\mathcal{B}$ is also a σ-field generated by $\mathcal{O}$, i.e., $\mathcal{B} = \sigma(\mathcal{O})$.

2. Suppose $(\Omega, \mathcal{A}, \mu)$ is a measure space. For any set $C \in \mathcal{A}$, we define $\mathcal{A} \cap C$ as $\{A \cap C : A \in \mathcal{A}\}$. Show that $(\Omega \cap C, \mathcal{A} \cap C, \mu)$ is a measure space (it is called the measure space restricted to C).

3. Suppose $(\Omega, \mathcal{A}, \mu)$ is a measure space. We define a new class

$$\tilde{\mathcal{A}} = \{A \cup N : A \in \mathcal{A} \text{ and } N \text{ is contained in a set } B \in \mathcal{A} \text{ with } \mu(B) = 0\}.$$

Furthermore, we define a set function $\tilde{\mu}$ on $\tilde{\mathcal{A}}$: for any $A \cup N \in \tilde{\mathcal{A}}$, $\tilde{\mu}(A \cup N) = \mu(A)$. Show $(\Omega, \tilde{\mathcal{A}}, \tilde{\mu})$ is a measure space (it is called the completion of $(\Omega, \mathcal{A}, \mu)$).

4. Suppose $(R, \mathcal{B}, P)$ is a probability measure space. Let $F(x) = P((-\infty, x])$. Show

   (a) $F(x)$ is an increasing and right-continuous function with $F(-\infty) = 0$ and $F(\infty) = 1$. F is called a distribution function.

   (b) if denote μ_F as the Lebesgue-Stieljes measure generated from F, then $P(B) = \mu_F(B)$ for any $B \in \mathcal{B}$. *Hint*: use the uniqueness of measure extension in the Caratheodory extension theorem.

   Remark: In other words, any probability measure in the Borel σ-field can be considered as a Lebesgue-Stieljes measure generated from some distribution function. Obviously, a Lebesgue-Stieljes measure generated from some distribution function is a probability measure. This gives a one-to-one correspondence between probability measures and distribution functions.

5. Let $(R, \mathcal{B}, \mu_F)$ be a measure space, where $\mathcal{B}$ is the Borel σ-filed and μ_F is the Lebesgue-Stieljes measure generated from $F(x) = (1 - e^{-x})I(x \geq 0)$.

   (a) Show that for any interval $(a, b]$, $\mu_F((a, b]) = \int_{(a,b]} e^{-x} I(x \geq 0) d\mu(x)$, where μ is the Lebesgue measure in R.

   (b) Use the uniqueness of measure extension in the Carotheodory extension theorem to show $\mu_F(B) = \int_B e^{-x} I(x \geq 0) d\mu(x)$ for any $B \in \mathcal{B}$.

   (c) Show that for any measurable function X in $(R, \mathcal{B})$ with $X \geq 0$, $\int X(x) d\mu_F(x) = \int X(x) e^{-x} I(x \geq 0) d\mu(x)$. *Hint*: use a sequence of simple functions to approximate X.

   (d) Using the above result and the fact that for any Riemann integrable function, its Riemann integral is the same as its Lebesgue integral, calculate the integration $\int (1 + e^{-x})^{-1} d\mu_F(x)$.

6. If $X \geq 0$ is a measurable function on a measure space $(\Omega, \mathcal{A}, \mu)$ and $\int X d\mu = 0$, then $\mu(\{\omega : X(\omega) > 0\}) = 0$.

7. Suppose X is a measurable function and $\int |X| d\mu < \infty$. Show that for each $\epsilon > 0$, there exists a $\delta > 0$ such that $\int_A |X| d\mu < \epsilon$ whenever $\mu(A) < \delta$.

8. Let μ be the Borel measure in R and ν be the counting measure in the space $\Omega = \{1, 2, 3, ...\}$ such that $\nu(\{n\}) = 2^{-n}$ for $n = 1, 2, 3, ...$ Define a function $f(x, y) : R \times \Omega \mapsto R$ as $f(x, y) = I(y-1 \leq x < y)x$. Show $f(x, y)$ is a measurable function with respect to the product measure space $(R \times \Omega, \sigma(\mathcal{B} \times 2^{\Omega}), \mu \times \nu)$ and calculate $\int_{R \times \Omega} f(x, y) d(\mu \times \nu)(x, y)$.

9. F and G are two continuous generalized distribution functions. Use the Fubini-Tonelli theorem to show that for any $a \leq b$,
$$F(b)G(b) - F(a)G(a) = \int_{[a,b]} F dG + \int_{[a,b]} G dF \quad (\textit{integration by parts}).$$
Hint: consider the equality
$$\int_{[a,b]\times[a,b]} d(\mu_F \times \mu_G) = \int_{[a,b]\times[a,b]} I(x \geq y) d(\mu_F \times \mu_G) + \int_{[a,b]\times[a,b]} I(x < y) d(\mu_F \times \mu_G),$$
where μ_F and μ_G are the measures generated by F and G respectively.

10. Let μ be the Borel measure in R. We list all rational numbers in R as $r_1, r_2, ...$ Define ν as another measure such that for any $B \in \mathcal{B}$, $\nu(B) = \mu(B \cap [0, 1]) + \sum_{r_i \in B} 2^{-i}$. Show that neither $\nu \prec\prec \mu$ nor $\mu \prec\prec \nu$ is true; however, $\nu \prec\prec \mu + \nu$. Calculate the Radon-Nikodym derivative $d\nu/d(\mu + \nu)$.

11. X is a random variable in a probability measure space $(\Omega, \mathcal{A}, P)$. Let P_X be the probability measure induced by X. Show that for any measurable function $g : R \to R$ such that $g(X)$ is integrable,
$$\int_{\Omega} g(X(\omega)) dP(\omega) = \int_{R} g(x) dP_X(x).$$
Hint: first prove it for a simple function g.

12. $X_1, ..., X_n$ are i.i.d with Uniform(0,1). Let $X_{(n)}$ be $\max\{X_1, ..., X_n\}$. Calculate the conditional expectation $E[X_1|\sigma(X_{(n)})]$, or equivalently, $E[X_1|X_{(n)}]$.

13. X and Y are two random variables with density functions $f(x)$ and $g(y)$ in R. Define $A = \{x : f(x) > 0\}$ and $B = \{y : g(y) > 0\}$. Show P_X, the measure induced by X, is dominated by P_Y, the measured induced by Y, if and only if $\lambda(A \cap B^c) = 0$ (that is, A is almost contained in B). Here, λ is the Lebesgue measure in R. Use this result to show that the measure induced by $Uniform(0, 1)$ random variable is dominated by the measure induced by $N(0, 1)$ random variable but the opposite is not true.

14. Continue Question 9, Chapter 1. The distribution functions F_U and F_L are called the Frḗchet bounds. Show that F_L and F_U are singular with respect to Lebesgue measure λ^2 in $[0, 1]^2$; i.e., show that the corresponding probability measure P_L and P_U satisfy
$$P((X, Y) \in A) = 1, \quad \lambda^2(A) = 0$$

and

$$P((X,Y) \in A^c) = 0, \quad \lambda^2(A^c) = 1$$

for some set A (which will be different for P_L and P_U). This implies that F_L and F_U do not have densities with respect to Lebesgue measure on $[0,1]^2$.

15. Lehmann and Casella, page 63, problem 2.6

16. Lehmann and Casella, page 64, problem 2.11

17. Lehmann and Casella, page 64, problem 3.1

18. Lehmann and Casella, page 64, problem 3.3

19. Lehmann and Casella, page 64, problem 3.7

CHAPTER 3 LARGE SAMPLE THEORY

In many probabilistic and statistical problems, we are faced with a sequence of random variables (vectors), say $\{X_n\}$, and wish to understand the limit properties of X_n. As one example, let X_n be the number of heads appearing in n independent tossing coins. Interesting questions can be: what is the limit of the proportion of observing heads, X_n/n, when n is large? How accurate is X_n/n to estimate the probability of observing head in a flipping? Such theory studying the limit properties of a sequence of random variables (vectors) $\{X_n\}$ is called large sample theory. In this chapter, we always assume the existence of a probability measure space $(\Omega, \mathcal{A}, P)$ and suppose $X, X_n, n \geq 1$ are random variables (vectors) defined in this probability space.

3.1 Modes of Convergence in Real Space

3.1.1 Definition

Definition 3.1 X_n is said to *converge almost surely* to X, denoted by $X_n \rightarrow_{a.s.} X$, if there exists a set $A \subset \Omega$ such that $P(A^c) = 0$ and for each $\omega \in A$, $X_n(\omega) \rightarrow X(\omega)$. †

Remark 3.1. Note that

$$\{\omega : X_n(\omega) \rightarrow X(\omega)\}^c = \cup_{\epsilon>0} \cap_n \{\omega : \sup_{m \geq n} |X_m(\omega) - X(\omega)| > \epsilon\}.$$

Then the above definition is equivalent to

$$P(\sup_{m\geq n} |X_m - X| > \epsilon) \rightarrow 0 \ \text{ as } n \rightarrow \infty.$$

Such an equivalence is also implied in Proposition 2.9.

Definition 3.2 X_n is said to *converge in probability* to X, denoted by $X_n \rightarrow_p X$, if for every $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \rightarrow 0.$$

†

Definition 3.3 X_n is said to *converge in rth mean* to X, denote by $X_n \rightarrow_r X$, if

$$E[|X_n - X|^r] \rightarrow 0 \ \text{ as } n \rightarrow \infty \text{ for functions } X_n, X \in L_r(P),$$

where $X \in L_r(P)$ means $E[|X|^r] = \int |X|^r dP < \infty$. †

Definition 3.4 X_n is said to *converge in distribution* of X, denoted by $X_n \rightarrow_d X$ or $F_n \rightarrow_d F$ (or $\mathbf{L}(X_n) \rightarrow \mathbf{L}(X)$ with $\mathbf{L}$ referring to the "law" or "distribution"), if the distribution functions F_n and F of X_n and X satisfy

$$F_n(x) \rightarrow F(x) \quad \text{as } n \rightarrow \infty \text{ for each continuity point } x \text{ of } F.$$

†

Definition 3.5 A sequence of random variables $\{X_n\}$ is *uniformly integrable* if

$$\lim_{\lambda\rightarrow\infty} \limsup_{n\rightarrow\infty} E\left[|X_n| I(|X_n| \geq \lambda)\right] = 0.$$

†

3.1.2 Relationship among modes

The following theorem describes the relationship among all the convergence modes.

Theorem 3.1 (A) If $X_n \rightarrow_{a.s.} X$, then $X_n \rightarrow_p X$.
(B) If $X_n \rightarrow_p X$, then $X_{n_k} \rightarrow_{a.s.} X$ for some subsequence X_{n_k}.
(C) If $X_n \rightarrow_r X$, then $X_n \rightarrow_p X$.
(D) If $X_n \rightarrow_p X$ and $|X_n|^r$ is uniformly integrable, then $X_n \rightarrow_r X$.
(E) If $X_n \rightarrow_p X$ and $\limsup_n E|X_n|^r \leq E|X|^r$, then $X_n \rightarrow_r X$.
(F) If $X_n \rightarrow_r X$, then $X_n \rightarrow_{r'} X$ for any $0 < r' \leq r$.
(G) If $X_n \rightarrow_p X$, then $X_n \rightarrow_d X$.
(H) $X_n \rightarrow_p X$ if and only if for every subsequence $\{X_{n_k}\}$ there exists a further subsequence $\{X_{n_k,l}\}$ such that $X_{n_k,l} \rightarrow_{a.s.} X$.
(I) If $X_n \rightarrow_d c$ for a constant c, then $X_n \rightarrow_p c$. †

Remark 3.2 The results of Theorem 3.1 appear to be complicated; however, they can be well described in Figure 1 below.

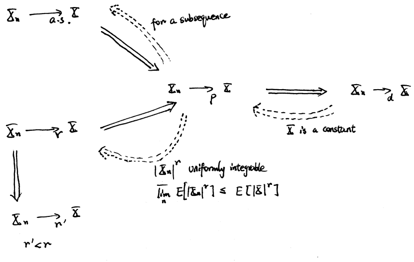


Figure 1: Relationship among Modes of Convergence

Proof (A) For any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \leq P(\sup_{m \geq n} |X_m - X| > \epsilon) \rightarrow 0.$$

(B) Since for any $\epsilon > 0$, $P(|X_n - X| > \epsilon) \rightarrow 0$, we choose $\epsilon = 2^{-m}$ then there exists a X_{n_m} such that

$$P(|X_{n_m} - X| > 2^{-m}) < 2^{-m}.$$

Particularly, we can choose n_m to be increasing. For the sequence $\{X_{n_m}\}$, we note that for any $\epsilon > 0$, when n_m is large,

$$P(\sup_{k \geq m} |X_{n_k} - X| > \epsilon) \leq \sum_{k \geq m} P(|X_{n_k} - X| > 2^{-k}) \leq \sum_{k \geq m} 2^{-k} \rightarrow 0.$$

Thus, $X_{n_m} \rightarrow_{a.s.} X$.
(C) We use the *Markov inequality*: for any positive and increasing function $g(\cdot)$ and random variable Y,

$$P(|Y| > \epsilon) \leq E[\frac{g(|Y|)}{g(\epsilon)}].$$

In particular, we choose $Y = |X_n - X|$ and $g(y) = |y|^r$. It gives that

$$P(|X_n - X| > \epsilon) \leq E[\frac{|X_n - X|^r}{\epsilon^r}] \rightarrow 0.$$

(D) It is sufficient to show that for any subsequence of $\{X_n\}$, there exists a further subsequence $\{X_{n_k}\}$ such that $E|X_{n_k} - X|^r \rightarrow 0$. For any subsequence of $\{X_n\}$, from (B), there exists a further subsequence $\{X_{n_k}\}$ such that $X_{n_k} \rightarrow_{a.s.} X$. We will show the result holds for $\{X_{n_k}\}$. For any ϵ, there exists λ such that

$$\limsup_{n_k} E[|X_{n_k}|^r I(|X_{n_k}|^r \geq \lambda)] < \epsilon.$$

Particularly, we choose λ (only depending on ϵ) such that $P(|X|^r = \lambda) = 0$. Then, it is clear that $|X_{n_k}|^r I(|X_{n_k}|^r \geq \lambda) \rightarrow_{a.s.} |X|^r I(|X|^r \geq \lambda)$. By the Fatou's Lemma,

$$E[|X|^r I(|X|^r \geq \lambda)] = \int \lim_n |X_{n_k}|^r I(|X_{n_k}|^r \geq \lambda) dP \leq \liminf_{n_k} E[|X_{n_k}|^r I(|X_{n_k}|^r \geq \lambda)] < \epsilon.$$

Therefore,

$$\begin{aligned}
& E[|X_{n_k} - X|^r] \\
\leq \; & E[|X_{n_k} - X|^r I(|X_{n_k}|^r < 2\lambda, |X|^r < 2\lambda)] + E[|X_{n_k} - X|^r I(|X_{n_k}|^r \geq 2\lambda \text{ or } |X|^r \geq 2\lambda)] \\
\leq \; & E[|X_{n_k} - X|^r I(|X_{n_k}|^r < 2\lambda, |X|^r < 2\lambda)] \\
& +2^r E[(|X_{n_k}|^r + |X|^r) I(|X_{n_k}|^r \geq 2\lambda \text{ or } |X|^r \geq 2\lambda)],
\end{aligned}$$

where the last inequality follows from the inequality $(x+y)^r \leq 2^r(\max(x,y))^r \leq 2^r(x^r + y^r), x \geq 0, y \geq 0$. Note that the first term converges to zero from the dominated convergence theorem.

Furthermore, when n_k is large, $I(|X_{n_k}| \geq 2\lambda) \leq I(|X| \geq \lambda)$ and $I(|X| \geq 2\lambda) \leq I(|X_{n_k}| \geq \lambda)$ almost surely. Then the second term is bounded by

$$2 * 2^r \left\{ E[|X_{n_k}|^r I(|X_{n_k}| \geq \lambda)] + E[|X|^r I(|X| \geq \lambda)] \right\},$$

which is smaller than $2^{r+1}\epsilon$. Thus,

$$\limsup_n E[|X_{n_k} - X|^r] \leq 2^{r+1}\epsilon.$$

Let ϵ tend to zero and the result holds.
(E) It is sufficient to show that for any subsequence of $\{X_n\}$, there exists a further subsequence $\{X_{n_k}\}$ such that $E[|X_{n_k} - X|^r] \to 0$. For any subsequence of $\{X_n\}$, from (B), there exists a further subsequence $\{X_{n_k}\}$ such that $X_{n_k} \to_{a.s.} X$. Define

$$Y_{n_k} = 2^r(|X_{n_k}|^r + |X|^r) - |X_{n_k} - X|^r \geq 0.$$

We apply the Fatou's Lemma to Y_n and obtain that

$$\int \liminf_{n_k} Y_{n_k} dP \leq \liminf_{n_k} \int Y_{n_k} dP.$$

It is equivalent to

$$2^{r+1} E[|X|^r] \leq \liminf_{n_k} \left\{ 2^r E[|X_{n_k}|^r] + 2^r E[|X|^r] - E[|X_{n_k} - X|^r] \right\}.$$

Thus,

$$\limsup_{n_k} E[|X_{n_k} - X|^r] \leq 2^r \left\{ \liminf_{n_k} E[|X_{n_k}|^r] - E[|X|^r] \right\} \leq 0.$$

The result holds.
(F) We need to use the *Hölder inequality* as follows

$$\int |f(x)g(x)| d\mu \leq \left\{ \int |f(x)|^p d\mu(x) \right\}^{1/p} \left\{ \int |g(x)|^q d\mu(x) \right\}^{1/q}, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

If we choose $\mu = P$, $f = |X_n - X|^{r'}$, $g \equiv 1$ and $p = r/r'$, $q = r/(r - r')$ in the Hölder inequality, we obtain

$$E[|X_n - X|^{r'}] \leq E[|X_n - X|^r]^{r'/r} \to 0.$$

(G) $X_n \to_p X$. If x is a continuity point of X, i.e., $P(X = x) = 0$, then for any $\epsilon > 0$,

$$\begin{aligned}
& P(|I(X_n \leq x) - I(X \leq x)| > \epsilon) \\
= \; & P(|I(X_n \leq x) - I(X \leq x)| > \epsilon, |X - x| > \delta) \\
& + P(|I(X_n \leq x) - I(X \leq x)| > \epsilon, |X - x| \leq \delta) \\
\leq \; & P(X_n \leq x, X > x + \delta) + P(X_n > x, X < x - \delta) + P(|X - x| \leq \delta) \\
\leq \; & P(|X_n - X| > \delta) + P(|X - x| \leq \delta).
\end{aligned}$$

The first term converges to zero as $n \to \infty$ since $X_n \to_p X$. The second term can be arbitrarily small if we choose δ is small, since $\lim_{\delta \to 0} P(|X - x| \leq \delta) = P(X = x) = 0$. Thus, we have shown that $I(X_n \leq x) \to_p I(X \leq x)$. From the dominated convergence theorem,

$$F_n(x) = E[I(X_n \leq x)] \to E[I(X \leq x)] = F_X(x).$$

Thus, $X_n \to_d X$.
(H) One direction follows from (B). To prove the other direction, we use the contradiction. Suppose there exists $\epsilon > 0$ such that $P(|X_n - X| > \epsilon)$ does not converge to zero. Then we can find a subsequence $\{X_{n'}\}$ such hat $P(|X_{n'} - X| > \epsilon) > \delta$ for some $\delta > 0$. However, by the condition, we can choose a further subsequence $X_{n''}$ such that $X_{n''} \to_{a.s.} X$ then $X_{n''} \to_p X$ from A. This is a contradiction.
(I) Let $X \equiv c$. It is clear from the following:

$$P(|X_n - c| > \epsilon) \leq 1 - F_n(c+\epsilon) + F_n(c-\epsilon) \to 1 - F_X(c+\epsilon) + F_X(c-\epsilon) = 0.$$

†

Remark 3.3 Denote $E[|X|^r]$ as μ_r. Then as proving (F) in Theorem 3.1., we obtain $\mu_r^{s-t}\mu_t^{r-s} \geq \mu_s^{r-t}$ where $r \geq s \geq t \geq 0$. Thus, $\log \mu_r$ is convex in r for $r \geq 0$. Furthermore, the proof of (F) says that $\mu_r^{1/r}$ is increasing in r.

Remark 3.4 For $r \geq 1$, we denote $E[|X|^r]^{1/r}$ as $\|X\|_r$ (or $\|X\|_{L_r(P)}$). Clearly, $\|X\|_r \geq 0$ and the equality holds if and only if $X = 0$ a.s. For any constant λ, $\|\lambda X\|_r = |\lambda|\|X\|_r$. Furthermore, we note that

$$E[|X+Y|^r] \leq E[(|X|+|Y|)|X+Y|^{r-1}] \leq E[|X|^r]^{1/r}E[|X+Y|^r]^{1-1/r} + E[|Y|^r]^{1/r}E[|X+Y|^r]^{1-1/r}.$$

Then we obtain a triangular inequality (called the *Minkowski's inequality*)

$$\|X+Y\|_r \leq \|X\|_r + \|Y\|_r.$$

Therefore, $\|\cdot\|_r$ in fact is a norm in the linear space $\{X : \|X\|_r < \infty\}$. Such a normed space is denoted as $L_r(P)$.

The following examples illustrate the results of Theorem 3.1.

Example 3.1 Suppose that X_n is degenerate at a point $1/n$; i.e., $P(X_n = 1/n) = 1$. Then X_n converges in distribution to zero. Indeed, X_n converges almost surely.

Example 3.2 $X_1, X_2, ...$ are i.i.d with standard normal distribution. Then $X_n \to_d X_1$ but X_n does not converge in probability to X_1.

Example 3.3 Let Z be a random variable with a uniform distribution in $[0,1]$. Let $X_n = I(m2^{-k} \leq Z < (m+1)2^{-k})$ when $n = 2^k + m$ where $0 \leq m < 2^k$. Then X_n converges in probability to zero but not almost surely. This example is already given in the second chapter.

Example 3.4 Let Z be $Uniform(0,1)$ and let $X_n = 2^n I(0 \leq Z < 1/n)$. Then $E[|X_n|^r]] \to \infty$ but X_n converges to zero almost surely.

The next theorem describes the necessary and sufficient conditions of convergence in moments from convergence in probability.

Theorem 3.2 (Vitali's theorem) Suppose that $X_n \in L_r(P)$, i.e., $\|X_n\|_r < \infty$, where $0 < r < \infty$ and $X_n \to_p X$. Then the following are equivalent:
(A) $\{|X_n|^r\}$ are uniformly integrable.

(B) $X_n \to_r X$.
(C) $E[|X_n|^r] \to E[|X|^r]$. †

Proof $(A) \Rightarrow (B)$ has been shown in proving (D) of Theorem 1.1. To prove $(B) \Rightarrow (C)$, first from the Fatou's lemma, we have

$$\liminf_n E[|X_n|^r] \geq E[|X|^r].$$

Second, we apply the Fatou's lemma to $2^r(|X_n - X|^r + |X|^r) - |X_n|^r \geq 0$ and obtain

$$E[2^r|X|^r - |X|^r] \leq 2^r \liminf_n E[|X_n - X|^r] + 2^r E[|X|^r] - \limsup_n E[|X_n|^r].$$

Thus,

$$\limsup_n E[|X_n|^r] \leq E[|X|^r] + 2^r \liminf_n E[|X_n - X|^r].$$

We conclude that $E[|X_n|^r] \to E[|X|^r]$.

To prove $(C) \Rightarrow (A)$, we note that for any λ such that $P(|X|^r = \lambda) = 0$, by the dominated convergence theorem,

$$\limsup_n E[|X_n|^r I(|X_n|^r \geq \lambda)] = \limsup_n \{E[|X_n|^r] - E[|X_n|^r I(|X_n|^r < \lambda)]\} = E[|X|^r I(|X|^r \geq \lambda)]$$

Thus,

$$\lim_{\lambda\to\infty} \limsup_n E[|X_n|^r I(|X_n|^r \geq \lambda)] = \lim_{\lambda\to\infty} \limsup_n E[|X|^r I(|X|^r \geq \lambda)] = 0.$$

†

From Theorem 3.2, we see that the uniform integrability plays an important role to ensure the convergence in moments. One sufficient condition to check the uniform integrability of $\{X_n\}$ is the *Liapunov condition*: if there exists a positive constant ϵ_0 such that $\limsup_n E[|X_n|^{r+\epsilon_0}] < \infty$, then $\{|X_n|^r\}$ satisfies the uniform integrability condition. This is because

$$E[|X_n|^r I(|X_n|^r \geq \lambda)] \leq \frac{E[|X_n|^{r+\epsilon_0}|]}{\lambda^{\epsilon_0}}.$$

3.1.3 Useful integral inequalities

We list some useful inequalities below, some of which have already been used. The first inequality is the Hölder inequality:

$$\int |f(x)g(x)|d\mu \leq \left\{\int |f(x)|^p d\mu(x)\right\}^{1/p} \left\{\int |g(x)|^q d\mu(x)\right\}^{1/q}, \quad \frac{1}{p} + \frac{1}{q} = 1.$$

We briefly describe how the *Hölder inequality* is derived. First, the following inequality holds (*Young's inequality*):

$$|ab| \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}, \quad a, b > 0,$$

where the equality holds if and only if $a = b$. This inequality is clear from its geometric meaning. In this inequality, we choose $a = f(x)/\int \{|f(x)|^p d\mu(x)\}^{1/p}$ and $b = g(x)/\int \{|g(x)|^q d\mu(x)\}^{1/q}$ and integrate over x on both side. It gives the Hölder inequality and the equality holds if and only if $f(x)$ is proportional to $g(x)$ almost surely. When $p = q = 2$, the inequality becomes

$$\int |f(x)g(x)|d\mu(x) \leq \left\{\int f(x)^2 d\mu(x)\right\}^{1/2} \left\{\int g(x)^2 d\mu(x)\right\}^{1/2},$$

which is the *Cauchy-Schwartz inequality.* One implication is that for non-trivial X and Y, $(E[|XY|])^2 \leq E[|X|^2]E[|Y|^2]$ and that the equality holds if and only if $|X| = c_0|Y|$ almost surely for some constant c_0.

A second important inequality is the Markov's inequality, which was used in proving (C) of Theorem 3.1:

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

where $g \geq 0$ is a increasing function in $[0, \infty)$. We can choose different g to obtain many similar inequalities. The proof of the Markov inequality is direct from the following:

$$P(|Y| > \epsilon) = E[I(|Y| > \epsilon)] \leq E[\frac{g(|Y|)}{g(\epsilon)} I(|Y| > \epsilon)] \leq E[\frac{g(|Y|)}{g(\epsilon)}].$$

If we choose $g(x) = x^2$ and X as $X - E[X]$ in the Markov inequality, we obtain

$$P(|X - E[X]| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2}.$$

This inequality is the *Chebychev's inequality* and gives an upper bound for controlling the tail probability of X using its variance.

In summary, we have introduced different modes of convergence for random variables and obtained the relationship among these modes. The same definitions and relationship can be generalized to random vectors. One additional remark is that since convergence almost surely or in probability are special definitions of convergence almost everywhere or in measure as given in the second chapter, all the theorems in Section 2.3.3 including the monotone convergence theorem, the Fatou's lemma and the dominated convergence theorem should apply. Convergence in distribution is the only one specific to probability measure. In fact, this model will be the main interest of the subsequent sections.

3.2 Convergence in Distribution

Among all the convergence modes of $\{X_n\}$, convergence in distribution is the weakest convergence. However, this convergence plays an important and sufficient role in statistical inference, especially when large sample behavior of random variables is of interest. We focus on such particular convergence in this section.

3.2.1 Portmanteau theorem

The following theorem gives all equivalent conditions to the convergence in distribution for a sequence of random variables $\{X_n\}$.

Theorem 3.3 (Portmanteau Theorem) The following conditions are equivalent.
(a) X_n converges in distribution to X.
(b) For any bounded continuous function $g(\cdot)$, $E[g(X_n)] \to E[g(X)]$.
(c) For any open set G in R, $\liminf_n P(X_n \in G) \geq P(X \in G)$.
(d) For any closed set F in R, $\limsup_n P(X_n \in F) \leq P(X \in F)$.
(e) For any Borel set O in R with $P(X \in \partial O) = 0$ where ∂O is the boundary of O, $P(X_n \in O) \to P(X \in O)$. †

Proof $(a) \Rightarrow (b)$. Without loss of generality, we assume $|g(x)| \leq 1$. We choose $[-M, M]$ such that $P(|X| = M) = 0$. Since g is continuous in $[-M, M]$, g is uniformly continuous in $[-M, M]$. Thus for any ϵ, we can partition $[-M, M]$ into finite intervals $I_1 \cup ... \cup I_m$ such that within each interval I_k, $\max_{I_k} g(x) - \min_{I_k} g(x) \leq \epsilon$ and X has no mass at all the endpoints of I_k (this is feasible since X has at most countable points with point masses). Therefore, if choose any point $x_k \in I_k, k = 1, ..., m$,

$$\begin{aligned} & |E[g(X_n)] - E[g(X)]| \\ \leq\ & E[|g(X_n)|I(|X_n| > M)] + E[|g(X)|I(|X| > M)] \\ & +|E[g(X_n)I(|X_n| \leq M)] - \sum_{k=1}^m g(x_k)P(X_n \in I_k)| \\ & +|\sum_{k=1}^m g(x_k)P(X_n \in I_k) - \sum_{k=1}^m g(x_k)P(X \in I_k)| \\ & +|E[g(X)I(|X| \leq M)] - \sum_{k=1}^m g(x_k)P(X \in I_k)| \\ \leq\ & P(|X_n| > M) + P(|X| > M) + 2\epsilon + \sum_{k=1}^m |P(X_n \in I_k) - P(X \in I_k)|. \end{aligned}$$

Thus, $\limsup_n |E[g(X_n)] - E[g(X)]| \leq 2P(|X| > M) + 2\epsilon$. Let $M \to \infty$ and $\epsilon \to 0$. We obtain (b).
$(b) \Rightarrow (c)$. For any open set G, we define a function

$$g(x) = 1 - \frac{\epsilon}{\epsilon + d(x, G^c)},$$

where $d(x, G^c)$ is the minimal distance between x and G^c, defined as $\inf_{y \in G^c} |x - y|$. Since for any $y \in G^c$,

$$d(x_1, G^c) - |x_2 - y| \leq |x_1 - y| - |x_2 - y| \leq |x_1 - x_2|,$$

we have $d(x_1, G^c) - d(x_2, G^c) \leq |x_1 - x_2|$. Then,

$$|g(x_1) - g(x_2)| \leq \epsilon^{-1}|d(x_1, G^c) - d(x_2, G^c)| \leq \epsilon^{-1}|x_1 - x_2|.$$

$g(x)$ is continuous and bounded. From (a), $E[g(X_n)] \to E[g(X)]$. Note $g(x) = 0$ if $x \notin G$ and $|g(x)| \leq 1$. Thus,

$$\liminf_n P(X_n \in G) \geq \liminf_n E[g(X_n)] \to E[g(X)].$$

Let $\epsilon \to 0$ and we obtain $E[g(X)]$ converges to $E[I(X \in G)] = P(X \in G)$.
$(c) \Rightarrow (d)$. This is clear by taking complement of F.
$(d) \Rightarrow (e)$. For any O with $P(X \in \partial O) = 0$, we have

$$\limsup_n P(X_n \in O) \leq \limsup_n P(X_n \in \bar{O}) \leq P(X \in \bar{O}) = P(X \in O),$$

and

$$\liminf_n P(X_n \in O) \geq \liminf_n P(X_n \in O^o) \geq P(X \in O^o) = P(X \in O).$$

Here, $\bar{O}$ and O^o are the closure and interior of O respectively.
$(e) \Rightarrow (a)$. It is clear by choosing $O = (-\infty, x]$ with $P(X \in \partial O) = P(X = x) = 0$. †

The conditions in Theorem 3.3 are necessary, as seen in the following examples.

Example 3.5 Let $g(x) = x$, a continuous but unbounded function. Let X_n be a random variable taking value n with probability $1/n$ and value 0 with probability $(1 - 1/n)$. Then $X_n \to_d 0$. However, $E[g(X)] = 1 \not\to 0$. This shows that the boundness of g in condition (b) is necessary.

Example 3.6 The continuity at boundary in (e) is also necessary: let X_n be degenerate at $1/n$ and consider $O = \{x : x > 0\}$. Then $P(X_n \in O) = 1$ but $X_n \to_d 0$.

3.2.2 Continuity theorem

Another way of verifying convergence in distribution of X_n is via the convergence of the characteristic functions of X_n, as given in the following theorem. This result is very useful in many applications.

Theorem 3.4 (Continuity Theorem) Let ϕ_n and ϕ denote the characteristic functions of X_n and X respectively. Then $X_n \to_d X$ is equivalent to $\phi_n(t) \to \phi(t)$ for each t. †

Proof To prove $\Rightarrow$ direction, from (b) in Theorem 3.1,

$$\phi_n(t) = E[e^{itX_n}] \to E[e^{itX}] = \phi(t).$$

We thus need to prove $\Leftarrow$ direction. This proof consists of the following steps.
Step 1. We show that for any ϵ, there exists a M such that $\sup_n P(|X_n| > M) < \epsilon$. This property is called *asymptotic tightness* of $\{X_n\}$. To see that, we note that

$$\begin{aligned} \frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi_n(t))dt &= E[\frac{1}{\delta}\int_{-\delta}^{\delta}(1-e^{itX_n})dt] \\ &= E[2(1-\frac{\sin\delta X_n}{\delta X_n})] \\ &\geq E[2(1-\frac{1}{|\delta X_n|})I(|X_n| > \frac{2}{\delta})] \\ &\geq P(|X_n| > \frac{2}{\delta}). \end{aligned}$$

However, the left-hand side of the inequality converges to

$$\frac{1}{\delta}\int_{-\delta}^{\delta}(1-\phi(t))dt.$$

Since $\phi(t)$ is continuous at $t=0$, this limit can be smaller than ϵ if we choose δ small enough. Let $M=\frac{2}{\delta}$. We obtain that when $n>N_0$, $P(|X_n|>M)<\epsilon$. Choose M larger then we can have $P(|X_k|>M)<\epsilon$, for $k=1,...,N_0$. Thus,

$$\sup_n P(|X_n|>M)<\epsilon.$$

Step 2. We show for any subsequence of $\{X_n\}$, there exists a further sub-sequence $\{X_{n_k}\}$ and the distribution function for X_{n_k}, denoted by F_{n_k}, converges to some distribution function. First, we need the Helly's Theorem.

Helly's Selection Theorem For every sequence $\{F_n\}$ of distribution functions, there exists a subsequence $\{F_{n_k}\}$ and a nondecreasing, right-continuous function F such that $F_{n_k}(x)\to F(x)$ at continuity points x of F. †

We defer the proof of the Helly's Selection Theorem to the end of the proof. Thus, from this theorem, for any subsequence of $\{X_n\}$, we can find a further subsequence $\{X_{n_k}\}$ such that $F_{n_k}(x)\to G(x)$ for some nondecreasing and right-continuous function G and the continuity points x of G. However, the Helly's Selection Theorem does not imply that G is a distribution function since $G(-\infty)$ and $G(\infty)$ may not be 0 or 1. But from the tightness of $\{X_{n_k}\}$, for any ϵ, we can choose M such that $F_{n_k}(-M)+(1-F_{n_k}(M))=P(|X_n|>M)<\epsilon$ and we can always choose M such that $-M$ and M are continuity points of G. Thus, $G(-M)+(1-G(M))<\epsilon$. Let $M\to\infty$ and since $0\le G(-M)\le G(M)\le 1$, we conclude that G must be a distribution function.

Step 3. We conclude that the subsequence $\{X_{n_k}\}$ in Step 2 converges in distribution to X. Since F_{n_k} weakly converges to $G(x)$ and $G(x)$ is a distribution function and $\phi_{n_k}(t)$ converges to $\phi(t)$, $\phi(t)$ must be the characteristic function corresponding to the distribution $G(x)$. From the uniqueness of the characteristic function in Theorem 1.1 (see the proof below), $G(x)$ is exactly the distribution of X. Therefore, $X_{n_k}\to_d X$. The theorem has been proved.

We need to prove the Helly's Selection Theorem: let $r_1,r_2,...$ be all the rational numbers. For r_1, we choose a subsequence of $\{F_n\}$, denoted by $F_{11},F_{12},...$ such that $F_{11}(r_1),F_{12}(r_1),...$ converges. Then for r_2, we choose a further subsequence from the above sequence, denote by $F_{21},F_{22},...$ such that $F_{21}(r_2),F_{22}(r_2),...$ converges. We continue this for all the rational numbers. We obtain a matrix of functions as follows:

$$\begin{pmatrix} F_{11} & F_{12} & \cdots \\ F_{21} & F_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}.$$

We finally select the diagonal functions, $F_{11},F_{22},....$ thus this subsequence converges for all the rational numbers. We denote their limits as $G(r_1),G(r_2),...$ Define $G(x)=\inf_{r_k>x}G(r_k)$. It is clear to see that G is nondecreasing. If x_k decreases to x, for any $\epsilon>0$, we can find r_s such that $r_s\ge x$ and $G(x)>G(r_s)-\epsilon$. Then when k is large, $G(x_k)-\epsilon\le G(r_s)-\epsilon<G(x)\le G(x_k)$.

That is, $\lim_k G(x_k) = G(x)$. Thus, G is right-continuous. If x is a continuity point of G, for any ϵ, we can find two sequence of rational number $\{r_k\}$ and $\{r_{k'}\}$ such that r_k decreases to x and $r_{k'}$ increases to x. Then after taking limits for the inequality $F_{ll}(r_{k'}) \leq F_{ll}(x) \leq F_{ll}(r_k)$, we have

$$G(r_{k'}) \leq \liminf_l F_{ll}(x) \leq \limsup_l F_{ll}(x) \leq G(r_k).$$

Let $k \to \infty$ then we obtain $\lim_l F_{ll}(x) = G(x)$.

It remains to prove Theorem 1.1, whose proof is deferred here: after substituting $\phi(t)$ in to the integration, we obtain

$$\frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\phi(t)dt = \frac{1}{2\pi}\int_{-T}^{T}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} dF(x)dt$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it} dtdF(x).$$

The interchange of the integrations follows from the Fubini's theorem. The last part is equal

$$\int_{-\infty}^{\infty} \left\{ \frac{sgn(x-a)}{\pi}\int_0^{T|x-a|} \frac{\sin t}{t}dt - \frac{sgn(x-b)}{\pi}\int_0^{T|x-b|} \frac{\sin t}{t}dt \right\} dF(x).$$

The integrand is bounded by $\frac{2}{\pi}\int_0^{\infty} \frac{\sin t}{t}dt$ and as $T \to \infty$, it converges to 0, if $x < a$ or $x > b$; $1/2$, if $x = a$ or $x = b$; 1, if $x \in (a, b)$. Therefore, by the dominated convergence theorem, the integral converges to

$$F(b-) - F(a) + \frac{1}{2}\{F(b) - F(b-)\} + \frac{1}{2}\{F(a) - F(a-)\}.$$

Since F is continuous at b and a, the limit is the same as $F(b) - F(a)$. Furthermore, suppose that F has a density function f. Then

$$F(x) - F(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1 - e^{-itx}}{it}\phi(t)dt.$$

Since $|\frac{\partial}{\partial x}\frac{1-e^{-itx}}{it}\phi(t)| \leq \phi(t)$, according to the interchange between derivative and integration, we obtain

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt.$$

†

The above theorem indicates that to prove the weak convergence of a sequence of random variables, it is sufficient to check the convergence of their characteristic functions. For example, if $X_1, ..., X_n$ are i.i.d Bernoulli(p), then the characteristic function of $\bar{X}_n = (X_1 + ... + X_n)/n$ is given by $(1 - p + pe^{it/n})^n$ converges to a function $\phi(t) = e^{itp}$, which is the characteristic function for a degenerate random variable $X \equiv p$. Thus $\bar{X}_n$ converges in distribution to p. Then from Theorem 3.1, $\bar{X}_n$ converges in probability to p.

Theorem 3.4 also has a multivariate version when X_n and X are k-dimensional random vectors: $X_n \to_d X$ if and only if $E[\exp\{it'X_n\}] \to E[\exp\{it'X\}]$, where t is any k-dimensional constant. Since the latter is equivalent to the weak convergence of $t'X_n$ to $t'X$, we conclude that the weak convergence of X_n to X is equivalent to the weak convergence of $t'X_n$ to $t'X$ for any t. That is, to study the weak convergence of random vectors, we can reduce to study the weak convergence of one-dimensional linear combination of the random vectors. This is the well-known *Cramér-Wold's device*:

Theorem 3.5 (The Cramér-Wold device) Random vector X_n in R^k satisfy $X_n \rightarrow_d X$ if and only $t'X_n \rightarrow_d t'X$ in R for all $t \in R^k$. †

3.2.3 Properties of convergence in distribution

Some additional results from convergence in distribution are the following theorems.

Theorem 3.6 (Continuous mapping theorem) Suppose $X_n \rightarrow_{a.s.} X$, or $X_n \rightarrow_p X$, or $X_n \rightarrow_d X$. Then for any continuous function $g(\cdot)$, $g(X_n)$ converges to $g(X)$ almost surely, or in probability, or in distribution. †

Proof If $X_n \rightarrow_{a.s.} X$, then clearly, $g(X_n) \rightarrow_{a.s} g(X)$. If $X_n \rightarrow_p X$, then for any subsequence, there exists a further subsequence $X_{n_k} \rightarrow_{a.s.} X$. Thus, $g(X_{n_k}) \rightarrow_{a.s.} g(X)$. Then $g(X_n) \rightarrow_p g(X)$ from (H) in Theorem 3.1. To prove that $g(X_n) \rightarrow_d g(X)$ when $X_n \rightarrow_d X$, we apply (b) of Theorem 3.3. †

Remark 3.5 Theorem 3.6 concludes that $g(X_n) \rightarrow_d g(X)$ if $X_n \rightarrow_d X$ and g is continuous. In fact, this result still holds if $P(X \in C(g)) = 1$ where $C(g)$ contains all the continuity points of g. That is, if g's discontinuity points take zero probability of X, the continuous mapping theorem holds.

Theorem 3.7 (Slutsky theorem) Suppose $X_n \rightarrow_d X$, $Y_n \rightarrow_p y$ and $Z_n \rightarrow_p z$ for some constant y and z. Then $Z_nX_n + T_n \rightarrow_d zX + y$. †

Proof We first show that $X_n + Y_n \rightarrow_d X + y$. For any $\epsilon > 0$,

$$P(X_n + Y_n \leq x) \leq P(X_n + Y_n \leq x, |Y_n - y| \leq \epsilon) + P(|Y_n - y| > \epsilon)$$

$$\leq P(X_n \leq x - y + \epsilon) + P(|Y_n - y| > \epsilon).$$

Thus,

$$\limsup_n F_{X_n+Y_n}(x) \leq \limsup_n F_{X_n}(x - y + \epsilon) \leq F_X(x - y + \epsilon).$$

On the other hand,

$$P(X_n + Y_n > x) = P(X_n + Y_n > x, |Y_n - y| \leq \epsilon) + P(|Y_n - y| > \epsilon)$$

$$\leq P(X_n > x - y - \epsilon) + P(|Y_n - y| > \epsilon).$$

Thus,

$$\limsup_n (1 - F_{X_n+Y_n}(x)) \leq \limsup_n P(X_n > x - y - \epsilon) \leq \limsup_n P(X_n \geq x - y - 2\epsilon)$$
$$\leq (1 - F_X(x - y - 2\epsilon)).$$

We obtain

$$F_X(x - y - 2\epsilon) \leq \liminf_n F_{X_n+Y_n}(x) \leq \limsup_n F_{X_n+Y_n}(x) \leq F_X(x + y + \epsilon).$$

Let $\epsilon \to 0$ then it holds

$$F_{X+y}(x-) \leq \liminf_n F_{X_n+Y_n}(x) \leq \limsup_n F_{X_n+Y_n}(x) \leq F_{X+y}(x).$$

Thus, $X_n + Y_n \to_d X + y$.

On the other hand, we have

$$P(|(Z_n - z)X_n| > \epsilon) \leq P(|Z_n - z| > \epsilon^2) + P(|Z_n - z| \leq \epsilon^2, |X_n| > \frac{1}{\epsilon}).$$

Thus,

$$\limsup_n P(|(Z_n - z)X_n| > \epsilon) \leq \limsup_n P(|Z_n - z| > \epsilon^2) + \limsup_n P(|X_n| \geq \frac{1}{2\epsilon}) \to P(|X| \geq \frac{1}{2\epsilon}).$$

Since ϵ is arbitrary, we conclude that $(Z_n - z)X_n \to_p 0$. Clearly $zX_n \to_d zX$. Hence, $Z_nX_n \to_d zX$ from the proof in the first half. Again, using the first half's proof, we obtain $Z_nX_n + Y_n \to_d zX + y$. †

Remark 3.6 In the proof of Theorem 3.7, if we replace $X_n + Y_n$ by $aX_n + bY_n$, we can show that $aX_n + bY_n \to_d aX + by$ by considering different cases of either a or b or both are non-zeros. Then from Theorem 3.5, $(X_n, Y_n) \to_d (X, y)$ in R^2. By the continuity theorem, we obtain $X_n + Y_n \to_d X + y$ and $X_nY_n \to_d Xy$. This immediately gives Theorem 3.7.

Both Theorems 3.6 and 3.7 are useful in deriving the convergence of some transformed random variables, as shown in the following examples.

Example 3.7 Suppose $X_n \to_d N(0, 1)$. Then by continuous mapping theorem, $X_n^2 \to_d \chi_1^2$.

Example 3.8 This example shows that g can be discontinuous in Theorem 3.6. Let $X_n \to_d X$ with $X \sim N(0, 1)$ and $g(x) = 1/x$. Although $g(x)$ is discontinuous at origin, we can still show that $1/X_n \to_d 1/X$, the reciprocal of the normal distribution. This is because $P(X = 0) = 0$. However, in Example 3.6 where $g(x) = I(x > 0)$, it shows that Theorem 3.6 may not be true if $P(X \in C(g)) < 1$.

Example 3.9 The condition $Y_n \to_p y$, where y is a constant, is necessary. For example, let $X_n = X \sim Uniform(0, 1)$. Let $Y_n = -X$ so $Y_n \to_d -\tilde{X}$, where $\tilde{X}$ is an independent random variable with the same distribution as X. However $X_n + Y_n = 0$ does not converge in distribution to the non-zero random variable $X - \tilde{X}$.

Example 3.10 Let $X_1, X_2, ...$ be a random sample from a normal distribution with mean μ and variance $\sigma^2 > 0$, then from the central limit theorem and the law of large number, which will be given later, we have

$$\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \sigma^2), \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \to_{a.s} \sigma^2.$$

Thus, from Theorem 3.7, it gives

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \to_d \frac{1}{\sigma}N(0, \sigma^2) \cong N(0, 1).$$

From the distribution theory, we know the left-hand side has a t-distribution with degrees of freedom $(n-1)$. Then this result says that in large sample, t_{n-1} can be approximated by a standard normal distribution.

3.2.4 Representation of convergence in distribution

As already seen before, working with convergence in distribution may not be easy, as compared with convergence almost surely. However, if we can represent convergence in distribution as convergence almost surely, many arguments can be simplified. The following famous theorem shows that such a representation does exist.

Theorem 3.8 (Skorohod's Representation Theorem) Let $\{X_n\}$ and X be random variables in a probability space $(\Omega, \mathcal{A}, P)$ and $X_n \to_d X$. Then there exists another probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ and a sequence of random variables $\tilde{X}_n$ and $\tilde{X}$ defined on this space such that $\tilde{X}_n$ and X_n have the same distributions, $\tilde{X}$ and X have the same distributions, and moreover, $\tilde{X}_n \to_{a.s.} \tilde{X}$. †

Before proving Theorem 3.8, we define the quantile function corresponding to a distribution function $F(x)$, denoted by $F^{-1}(p)$, for $p \in [0, 1]$,

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}.$$

Some properties regarding the quantile function are given in the following proposition.

Proposition 3.1 (a) F^{-1} is left-continuous.
(b) If X has continuous distribution function F, then $F(X) \sim Uniform(0, 1)$.
(c) Let $\xi \sim Uniform(0, 1)$ and let $X = F^{-1}(\xi)$. Then for all x, $\{X \leq x\} = \{\xi \leq F(x)\}$. Thus, X has distribution function F. †

Proof (a) Clearly, F^{-1} is nondecreasing. Suppose p_n increases to p then $F^{-1}(p_n)$ increases to some $y \leq F^{-1}(p)$. Then $F(y) \geq p_n$ so $F(y) \geq p$. Therefore $F^{-1}(p) \leq y$ by the definition of $F^{-1}(p)$. Thus $y = F^{-1}(p)$. F^{-1} is left-continuous.
(b) $\{X \leq x\} \subset \{F(X) \leq F(x)\}$. Thus, $F(x) \leq P(F(X) \leq F(x))$. On the other hand, $\{F(X) \leq F(x) - \epsilon\} \subset \{X \leq x\}$. Thus, $P(F(X) \leq F(x) - \epsilon) \leq F(x)$. Let $\epsilon \to 0$ and we obtain $P(F(X) \leq F(x)-) \leq F(x)$. Then if X is continuous, we have $P(F(X) \leq F(x)) = F(x)$ so

$F(X) \sim Uniform(0,1)$.
(c) $P(X \leq x) = P(F^{-1}(\xi) \leq x) = P(\xi \leq F(x)) = F(x)$. †

Proof Using the quantile function, we can construct the proof of Theorem 3.8. Let $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ be $([0,1], \mathcal{B} \cap [0,1], \lambda)$, where λ is the Borel measure. Define $\tilde{X}_n = F_n^{-1}(\xi)$, $\tilde{X} = F^{-1}(\xi)$, where ξ is uniform random variable on $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$. From (c) in the previous proposition, $\tilde{X}_n$ has a distribution F_n which is the same as X_n. It remains to show $\tilde{X}_n \to_{a.s.} \tilde{X}$.

For any $t \in (0,1)$ such that there is at most one value x such that $F(x) = t$ (it is easy to see t is the continuous point of F^{-1}), we have that for any $z < x$, $F(z) < t$. Thus, when n is large, $F_n(z) < t$ so $F_n^{-1}(t) \geq z$. We obtain $\liminf_n F_n^{-1}(t) \geq z$. Since z is any number less than x, we have $\liminf_n F_n^{-1}(t) \geq x = F^{-1}(t)$. On the other hand, from $F(x+\epsilon) > t$, we obtain when n is large enough, $F_n(x+\epsilon) > t$ so $F_n^{-1}(t) \leq x + \epsilon$. Thus, $\limsup_n F_n^{-1}(t) \leq x + \epsilon$. Since ϵ is arbitrary, we obtain $\limsup_n F_n^{-1}(t) \leq x$.

We conclude $F_n^{-1}(t) \to F^{-1}(t)$ for any t which is continuous point of F^{-1}. Thus $F_n^{-1}(t) \to F^{-1}(t)$ for almost every $t \in (0,1)$. That is, $\tilde{X}_n \to_{a.s.} \tilde{X}$. †

This theorem can be useful in a lot of arguments. For example, if $X_n \to_d X$ and one wishes to show some function of X_n, denote by $g(X_n)$, converges in distribution to $g(X)$, then by the representation theorem, we obtain $\tilde{X}_n$ and $\tilde{X}$ and $\tilde{X}_n \to_{a.s.} \tilde{X}$. Thus, if we can show $g(\tilde{X}_n) \to_{a.s.} g(\tilde{X})$, which is often easy to show, then of course, $g(\tilde{X}_n) \to_d g(\tilde{X})$. Since $g(\tilde{X}_n)$ has the same distribution as $g(X_n)$ and so are $g(\tilde{X})$ and $g(X)$, $g(X_n) \to_d g(X)$. Using this technique, readers should easily prove the continuous mapping theorem. Also see the diagram in Figure 2.

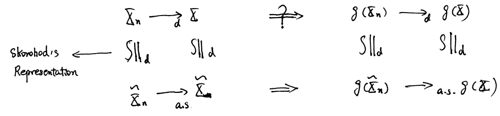

$X_n \to_d X$
?
$g(X_n) \to_d g(X)$
Skorohod's Representation
$\wr\|_d$
$\wr\|_d$
$\wr\|_d$
$\wr\|_d$
$\tilde{X}_n \xrightarrow[a.s]{} \tilde{X}$
$\Longrightarrow$
$g(\tilde{X}_n) \to_{a.s.} g(\tilde{X})$

Figure 2: Representation of Convergence in Distribution

Our final remark of this section is that all the results such as the continuous mapping theorem, the Slutsky theorem and the representation theorem can be in parallel given for the convergence of random vectors. The proofs for random vectors are based on the Cramé-Wold's device.

3.3 Summation of Independent Random Variables

The summation of independent random variables are commonly seen in statistical inference. Specially, many statistics can be expressed as the summation of i.i.d random variables. Thus, this section gives some classical large sample results for this type of statistics, which include the weak/strong law of large numbers, the central limit theorem, and the Delta method etc.

3.3.1 Preliminary lemma

Proposition 3.2 (Borel-Cantelli Lemma) For any events A_n,

$$\sum_{i=1}^{\infty} P(A_n) < \infty$$

implies $P(A_n, i.o.) = P(\{A_n\}$ occurs infinitely often$) = 0$; or equivalently, $P(\cap_{n=1}^{\infty} \cup_{m\geq n} A_m) = 0$. †

Proof

$$P(A_n, i.o) \leq P(\cup_{m\geq n} A_m) \leq \sum_{m\geq n} P(A_m) \to 0, \quad \text{as } n \to \infty.$$

†

As a result of the proposition, if for a sequence of random variables, $\{Z_n\}$, and for any $\epsilon > 0$, $\sum_n P(|Z_n| > \epsilon) < \infty$. Then with probability one, $|Z_n| > \epsilon$ only occurs finite times. That is, $Z_n \to_{a.s.} 0$.

Proposition 3.3 (Second Borel-Cantelli Lemma) For a sequence of independent events $A_1, A_2, ..., \sum_{n=1}^{\infty} P(A_n) = \infty$ implies $P(A_n, i.o.) = 1$. †

Proof Consider the complement of $\{A_n, i.o\}$. Note

$$P(\cup_{n=1}^{\infty} \cap_{m\geq n} A_m^c) = \lim_n P(\cap_{m\geq n} A_m^c) = \lim_n \prod_{m\geq n} (1 - P(A_m)) \leq \limsup_n \exp\{-\sum_{m\geq n} P(A_m)\} = 0.$$

†

Proposition 3.4 $X, X_1, ..., X_n$ are i.i.d with finite mean. Define $Y_n = X_n I(|X_n| \leq n)$. Then $\sum_{n=1}^{\infty} P(X_n \neq Y_n) < \infty$. †

Proof Since $E[|X|] < \infty$,

$$\sum_{n=1}^{\infty} P(X_n \neq Y_n) \leq \sum_{n=1}^{\infty} P(|X| \geq n) = \sum_{n=1}^{\infty} nP(n \leq |X| < (n+1)) \leq \sum_{n=1}^{\infty} E[|X|] < \infty.$$

From the Borel-Cantelli Lemma, $P(X_n \neq Y_n, i.o) = 0$. That is, for almost every $\omega \in \Omega$, when n is large enough, $X_n(\omega) = Y_n(\omega)$. †

3.3.2 Law of large numbers

We start to prove the weak and strong law of large numbers.

Theorem 3.9 (Weak Law of Large Number) If $X, X_1, ..., X_n$ are i.i.d with mean μ (so $E[|X|] < \infty$ and $\mu = E[X]$), then $\bar{X}_n \rightarrow_p \mu$. †

Proof Define $Y_n = X_n I(-n \leq X_n \leq n)$. Let $\bar{\mu}_n = \sum_{k=1}^n E[Y_k]/n$. Then by the Chebyshev's inequality,

$$P(|\bar{Y}_n - \bar{\mu}_n| \geq \epsilon) \leq \frac{Var(\bar{Y}_n)}{\epsilon^2} \leq \frac{\sum_{k=1}^n Var(X_k I(|X_k| \leq k))}{n^2\epsilon^2}.$$

Since

$$\begin{aligned} Var(X_k I(|X_k| \leq k)) &\leq E[X_k^2 I(|X_k| \leq k)] \\ &= E[X_k^2 I(|X_k| \leq k, |X_k| \geq \sqrt{k}\epsilon^2)] + E[X_k^2 I(|X_k| \leq k, |X| \leq \sqrt{k}\epsilon^2)] \\ &\leq kE[|X_k| I(|X_k| \geq \sqrt{k}\epsilon^2)] + k\epsilon^4, \end{aligned}$$

$$P(|\bar{Y}_n - \bar{\mu}_n| \geq \epsilon) \leq \frac{\sum_{k=1}^n E[|X| I(|X| \geq \sqrt{k}\epsilon^2)]}{n\epsilon^2} + \epsilon^2 \frac{n(n+1)}{2n^2}.$$

Thus, $\limsup_n P(|\bar{Y}_n - \bar{\mu}_n| \geq \epsilon) \leq \epsilon^2$. We conclude that $\bar{Y}_n - \bar{\mu}_n \rightarrow_p 0$. On the other hand, $\bar{\mu}_n \rightarrow \mu$. We obtain $\bar{Y}_n \rightarrow_p \mu$. This implies that for any subsequence, there is a further subsequence $\bar{Y}_{nk} \rightarrow_{a.s.} \mu$. Since X_n is eventually the same as Y_n for almost every ω from Proposition 3.4, we conclude $\bar{X}_{nk} \rightarrow_{a.s.} \mu$. This implies $X_n \rightarrow_p \mu$. †

Theorem 3.10 (Strong Law of Large Number) If $X_1, ..., X_n$ are i.i.d with mean μ then $\bar{X}_n \rightarrow_{a.s.} \mu$. †

Proof Without loss of generality, we assume $X_n \geq 0$ since if this is true, the result also holds for any X_n by $X_n = X_n^+ - X_n^-$.

Similar to Theorem 3.9, it is sufficient to show $\bar{Y}_n \rightarrow_{a.s.} \mu$, where $Y_n = X_n I(X_n \leq n)$. Note $E[Y_n] = E[X_1 I(X_1 \leq n)] \rightarrow \mu$ so

$$\sum_{k=1}^n E[Y_k]/n \rightarrow \mu.$$

Thus, if we denote $\tilde{S}_n = \sum_{k=1}^n (Y_k - E[Y_k])$ and we can show $\tilde{S}_n/n \rightarrow_{a.s.} 0$, then the result holds.

Note

$$Var(\tilde{S}_n) = \sum_{k=1}^n Var(Y_k) \leq \sum_{k=1}^n E[Y_k^2] \leq nE[X_1^2 I(X_1 \leq n)].$$

Then by the Chebyshev's inequality,

$$P(|\frac{\tilde{S}_n}{n}| > \epsilon) \leq \frac{1}{n^2\epsilon^2} Var(\tilde{S}_n) \leq \frac{E[X_1^2 I(X_1 \leq n)]}{n\epsilon^2}.$$

For any $\alpha > 1$, let $u_n = [\alpha^n]$. Then

$$\sum_{n=1}^{\infty} P(|\frac{\tilde{S}_{u_n}}{u_n}| > \epsilon) \leq \sum_{n=1}^{\infty} \frac{1}{u_n \epsilon^2} E[X_1^2 I(X_1 \leq u_n)] \leq \frac{1}{\epsilon^2} E[X_1^2 \sum_{u_n \geq X_1} \frac{1}{u_n}].$$

Since for any $x > 0$, $\sum_{u_n \geq x} \{\mu_n\}^{-1} < 2 \sum_{n \geq \log x / \log \alpha} \alpha^{-n} \leq K x^{-1}$ for some constant K, we have

$$\sum_{n=1}^{\infty} P(|\frac{\tilde{S}_{u_n}}{u_n}| > \epsilon) \leq \frac{K}{\epsilon^2} E[X_1] < \infty,$$

From the Borel-Cantelli Lemma in Proposition 3.2, $\tilde{S}_{u_n}/u_n \rightarrow_{a.s.} 0$.

For any k, we can find $u_n < k \leq u_{n+1}$. Thus, since $X_1, X_2, ... \geq 0$,

$$\frac{\tilde{S}_{u_n}}{u_n} \frac{u_n}{u_{n+1}} \leq \frac{\tilde{S}_k}{k} \leq \frac{\tilde{S}_{u_{n+1}}}{u_{n+1}} \frac{u_{n+1}}{u_n}.$$

After taking limits in the above, we have

$$\mu/\alpha \leq \liminf_k \frac{\tilde{S}_k}{k} \leq \limsup_k \frac{\tilde{S}_k}{k} \leq \mu\alpha.$$

Since α is arbitrary number larger than 1, let $\alpha \rightarrow 1$ and we obtain $\lim_k \tilde{S}_k/k = \mu$. The proof is completed. †

3.3.3 Central limit theorem

We now consider the central limit theorem. All the proofs can be based on the convergence of the corresponding characteristic function. The following lemma describes the approximation of a characteristic function.

Proposition 3.5 Suppose $E[|X|^m] < \infty$ for some integer $m \geq 0$. Then

$$|\phi_X(t) - \sum_{k=0}^{m} \frac{(it)^k}{k!} E[X^k]|/|t|^m \rightarrow 0, \quad \text{as } t \rightarrow 0.$$

†

Proof We note the following expansion for e^{itx},

$$e^{itx} = \sum_{k=1}^{m} \frac{(itx)^k}{k!} + \frac{(itx)^m}{m!} [e^{it\theta x} - 1],$$

where $\theta \in [0, 1]$. Thus,

$$|\phi_X(t) - \sum_{k=0}^{m} \frac{(it)^k}{k!} E[X^k]|/|t|^m \leq E[|X|^m |e^{it\theta X} - 1|]/m! \rightarrow 0,$$

as $t \to 0$. †

Theorem 3.11 (Central Limit Theorem) If $X_1, ..., X_n$ are i.i.d with mean μ and variance σ^2 then $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \sigma^2)$. †

Proof Denote $Y_n = \sqrt{n}(\bar{X}_n - \mu)$. We consider the characteristic function of Y_n.

$$\phi_{Y_n}(t) = \left\{\phi_{X_1-\mu}(t/\sqrt{n})\right\}^n.$$

Using Proposition 3.5, we have $\phi_{X_1-\mu}(t/\sqrt{n}) = 1 - \sigma^2 t^2/2n + o(1/n)$. Thus,

$$\phi_{Y_n}(t) \to \exp\{-\frac{\sigma^2 t^2}{2}\}.$$

The result holds. †

Theorem 3.12 (Multivariate Central Limit Theorem) If $X_1, ..., X_n$ are i.i.d random vectors in R^k with mean μ and covariance $\Sigma = E[(X-\mu)(X-\mu)']$, then $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \Sigma)$. †

Proof Similar to Theorem 3.11, but this time, we consider a multivariate characteristic function $E[\exp\{i\sqrt{n}t'(\bar{X}_n - \mu)\}]$. Note the result of Proposition 3.5 holds for this multivariate case. †

Theorem 3.13 (Liapunov Central Limit Theorem) Let $X_{n1}, ..., X_{nn}$ be independent random variables with $\mu_{ni} = E[X_{ni}]$ and $\sigma_{ni}^2 = Var(X_{ni})$. Let $\mu_n = \sum_{i=1}^n \mu_{ni}$, $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2$. If

$$\sum_{i=1}^n \frac{E[|X_{ni} - \mu_{ni}|^3]}{\sigma_n^3} \to 0,$$

then $\sum_{i=1}^n (X_{ni} - \mu_{ni})/\sigma_n \to_d N(0, 1)$. †

We skip the proof of Theorem 3.13 but try to give a proof for the following Theorem 3.14, for which Theorem 3.13 is a special case.

Theorem 3.14 (Lindeberg-Fell Central Limit Theorem) Let $X_{n1}, ..., X_{nn}$ be independent random variables with $\mu_{ni} = E[X_{ni}]$ and $\sigma_{ni}^2 = Var(X_{ni})$. Let $\sigma_n^2 = \sum_{i=1}^n \sigma_{ni}^2$. Then both $\sum_{i=1}^n (X_{ni} - \mu_{ni})/\sigma_n \to_d N(0, 1)$ and $\max\{\sigma_{ni}^2/\sigma_n^2 : 1 \le i \le n\} \to 0$ if and only if the Lindeberg condition

$$\frac{1}{\sigma_n^2}\sum_{i=1}^n E[|X_{ni} - \mu_{ni}|^2 I(|X_{ni} - \mu_{ni}| \ge \epsilon\sigma_n)] \to 0, \quad \text{for all } \epsilon > 0$$

holds. †

Proof " $\Leftarrow$": We first show that $\max\{\sigma_{nk}^2/\sigma_n^2 : 1 \le k \le n\} \to 0$.

$$\begin{aligned}\sigma_{nk}^2/\sigma_n^2 &\le E[|(X_{nk}-\mu_k)/\sigma_n|^2]\\ &\le \frac{1}{\sigma_n^2}\left\{E[I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)(X_{nk}-\mu_{nk})^2]+E[I(|X_{nk}-\mu_{nk}|<\epsilon\sigma_n)(X_{nk}-\mu_{nk})^2]\right\}\\ &\le \frac{1}{\sigma_n^2}E[I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)(X_{nk}-\mu_{nk})^2]+\epsilon^2.\end{aligned}$$

Thus,

$$\max_k\{\sigma_{nk}^2/\sigma_n^2\} \le \frac{1}{\sigma_n^2}\sum_{k=1}^n E[|X_{nk}-\mu_{nk}|^2 I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)]+\epsilon^2.$$

From the Lindeberg condition, we immediately obtain

$$\max_k\{\sigma_{nk}^2/\sigma_n^2\}\to 0.$$

To prove the central limit theorem, we let $\phi_{nk}(t)$ be the characteristic function of $(X_{nk}-\mu_{nk})/\sigma_n$. We note

$$\begin{aligned}&|\phi_{nk}(t)-(1-\frac{\sigma_{nk}^2}{\sigma_n^2}\frac{t^2}{2})|\\ \le&E\left[\left|e^{it(X_{nk}-\mu_{nk})/\sigma_n}-\sum_{j=0}^2\frac{(it)^j}{j!}\left(\frac{X_{nk}-\mu_{nk}}{\sigma_n}\right)^j\right|\right]\\ \le&E\left[I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)\left|e^{it(X_{nk}-\mu_{nk})/\sigma_n}-\sum_{j=0}^2\frac{(it)^j}{j!}\left(\frac{X_{nk}-\mu_{nk}}{\sigma_n}\right)^j\right|\right]\\ &+E\left[I(|X_{nk}-\mu_{nk}|<\epsilon\sigma_n)\left|e^{it(X_{nk}-\mu_{nk})/\sigma_n}-\sum_{j=0}^2\frac{(it)^j}{j!}\left(\frac{X_{nk}-\mu_{nk}}{\sigma_n}\right)^j\right|\right].\end{aligned}$$

From the expansion in proving Proposition 3.5, the inequality $|e^{itx}-(1+itx-t^2x^2/2)|\le t^2x^2$ so we apply it to the first half on the right-hand side. Additionally, from the Taylor expansion, $|e^{itx}-(1+itx-t^2x^2/2)|\le|t|^3|x|^3/6$ so we apply it to the second half of the right-hand side. Then, we obtain

$$\begin{aligned}&|\phi_{nk}(t)-(1-\frac{\sigma_{nk}^2}{\sigma_n^2}\frac{t^2}{2})|\\ \le&E\left[I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)t^2\left(\frac{X_{nk}-\mu_{nk}}{\sigma_n}\right)^2\right]\\ &+E\left[I(|X_{nk}-\mu_{nk}|<\epsilon\sigma_n)|t|^3\frac{|X_{nk}-\mu_{nk}|^3}{6\sigma_n^3}|\right]\\ \le&\frac{t^2}{\sigma_n^2}E[(X_{nk}-\mu_{nk})^2I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)]+\frac{\epsilon|t|^3}{6}\frac{\sigma_{nk}^2}{\sigma_n^2}.\end{aligned}$$

Therefore,

$$\sum_{k=1}^n|\phi_{nk}(t)-(1-\frac{t^2}{2}\frac{\sigma_{nk}^2}{\sigma_n^2})|\le\frac{t^2}{\sigma_n^2}\sum_{k=1}^n E[I(|X_{nk}-\mu_{nk}|\ge\epsilon\sigma_n)(X_{nk}-\mu_{nk})^2]+\frac{\epsilon|t|^3}{6}.$$

This summation goes to zero as $n \to \infty$ then $\epsilon \to 0$.

Since for any complex numbers $Z_1, ..., Z_m, W_1, ..., W_m$ with norm at most 1,

$$|Z_1 \cdots Z_m - W_1 \cdots W_m| \leq \sum_{k=1}^m |Z_k - W_k|,$$

we have

$$|\prod_{k=1}^n \phi_{nk}(t) - \prod_{k=1}^n (1 - \frac{t^2}{2}\frac{\sigma_{nk}^2}{\sigma_n^2}))| \leq \sum_{k=1}^n |\phi_{nk}(t) - (1 - \frac{t^2}{2}\frac{\sigma_{nk}^2}{\sigma_n^2})| \to 0.$$

On the other hand, from $|e^z - 1 - z| \leq |z|^2 e^{|z|}$,

$$|\prod_{k=1}^n e^{-t^2\sigma_{nk}^2/2\sigma_n^2} - \prod_{k=1}^n (1 - \frac{t^2}{2}\frac{\sigma_{nk}^2}{\sigma_n^2}))| \leq \sum_{k=1}^n |e^{-t^2\sigma_{nk}^2/2\sigma_n^2} - 1 + t^2\sigma_{nk}^2/2\sigma_n^2|$$

$$\leq \sum_{k=1}^n e^{t^2\sigma_{nk}^2/2\sigma_n^2} t^4\sigma_{nk}^4/4\sigma_n^4 \leq (\max_k\{\sigma_{nk}/\sigma_n\})^2 e^{t^2/2} t^4/4 \to 0.$$

We have

$$|\prod_{k=1}^n \phi_{nk}(t) - \prod_{k=1}^n e^{-t^2\sigma_{nk}^2/2\sigma_n^2}| \to 0.$$

The result thus follows by noticing

$$\prod_{k=1}^n e^{-t^2\sigma_{nk}^2/2\sigma_n^2} \to e^{-t^2/2}.$$

" $\Rightarrow$": First, we note that from $1 - \cos x \leq x^2/2$,

$$\frac{t^2}{2\sigma_n^2}\sum_{k=1}^n E[|X_{nk} - \mu_{nk}|^2 I(|X_{nk} - \mu_{nk}| > \epsilon\sigma_n)] \leq \frac{t^2}{2} - \sum_{k=1}^n \int_{|X_{nk}-\mu_{nk}|\leq\epsilon\sigma_n} \frac{t^2y^2}{2\sigma_n^2} dF_{nk}(y)$$

$$\leq \frac{t^2}{2} - \sum_{k=1}^n \int_{|X_{nk}-\mu_{nk}|\leq\epsilon\sigma_n} [1 - \cos(ty/\sigma_n)] dF_{nk}(y),$$

where F_{nk} is the distribution for $X_{nk} - \mu_{nk}$. On the other hand, since $\max\{\sigma_{nk}/\sigma_n\} \to 0$, $\max_k |\phi_{nk}(t) - 1| \to 0$ uniformly on any finite interval of t. Then

$$|\sum_{k=1}^n \log\phi_{nk}(t) - \sum_{k=1}^n (\phi_{nk}(t) - 1)| \leq \sum_{k=1}^n |\phi_{nk}(t) - 1|^2 \leq \max_k\{|\phi_{nk}(t) - 1|\} \sum_{k=1}^n |\phi_{nk}(t) - 1|$$

$$\leq \max_k\{|\phi_{nk}(t) - 1|\} \sum_{k=1}^n t^2\sigma_{nk}^2/\sigma_n^2.$$

Thus,

$$\sum_{k=1}^n \log\phi_{nk}(t) = \sum_{k=1}^n (\phi_{nk}(t) - 1) + o(1).$$

Since $\sum_{k=1}^n \log \phi_{nk}(t) \to -t^2/2$ uniformly in any finite interval of t, we obtain

$$\sum_{k=1}^n (1-\phi_{nk}(t)) = t^2/2 + o(1)$$

uniformly in finite interval of t. That is,

$$\sum_{k=1}^n \int (1-\cos(ty/\sigma_n))dF_{nk}(y) = t^2/2 + o(1).$$

Therefore, for any ϵ and for any $|t| \le M$, when n is large,

$$\frac{t^2}{2\sigma_n^2}\sum_{k=1}^n E[|X_{nk}-\mu_{nk}|^2 I(|X_{nk}-\mu_{nk}|>\epsilon\sigma_n)] \le \sum_{k=1}^n \int_{|X_{nk}-\mu_{nk}|>\epsilon\sigma_n} [1-\cos(ty/\sigma_n)]dF_{nk}(y) + \epsilon$$

$$\le 2\sum_{k=1}^n \int_{|X_{nk}-\mu_{nk}|>\epsilon\sigma_n} dF_{nk}(y) + \epsilon \le \frac{2}{\epsilon^2}\sum_{k=1}^n \frac{E[|X_{nk}-\mu_{nk}|^2]}{\sigma_n^2} + \epsilon \le 2/\epsilon^2 + \epsilon.$$

Let $t = M = 1/\epsilon^3$ and we obtain the Lindeberg condition. †

Remark 3.7 To see how Theorem 3.14 implies the result in Theorem 3.13, we note that

$$\frac{1}{\sigma_n^2}\sum_{i=1}^n E[|X_{nk}-\mu_{nk}|^2 I(|X_{nk}-\mu_{nk}|>\epsilon\sigma_n)] \le \frac{1}{\epsilon^3\sigma_n^3}\sum_{k=1}^n E[|X_{nk}-\mu_{nk}|^3].$$

We give some examples to show the application of the central limit theorems in statistics.

Example 3.11 This is one example from a simple linear regression. Suppose $X_j = \alpha + \beta z_j + \epsilon_j$ for $j = 1, 2, ...$ where z_j are known numbers not all equal and the ϵ_j are i.i.d with mean zero and variance σ^2. We know that the least square estimate for β is given by

$$\hat{\beta}_n = \sum_{j=1}^n X_j(z_j - \bar{z}_n)/\sum_{j=1}^n (z_j - \bar{z}_n)^2$$

$$= \beta + \sum_{j=1}^n \epsilon_j(z_j - \bar{z}_n)/\sum_{j=1}^n (z_j - \bar{z}_n)^2.$$

Assume

$$\max_{j\le n}(z_j - \bar{z}_n)^2/\sum_{j=1}^n (z_j - \bar{z}_n)^2 \to 0.$$

we can show that the Lindeberg condition is satisfied. Thus, we conclude that

$$\sqrt{n}\sqrt{\frac{\sum_{j=1}^n (z_j - \bar{z}_n)^2}{n}}(\hat{\beta}_n - \beta) \to_d N(0, \sigma^2).$$

Example 3.12 The example is taken from the randomization test for paired comparison. In a paired study comparing treatment vs control, $2n$ subjects are grouped into n pairs. For pair, it is decided at random that one subject receives treatment but not the other. Let (X_j, Y_j) denote the values of jth pairs with X_j being the result of the treatment. The usual paired t-test is based on the normality of $Z_j = X_j - Y_j$ which may be invalid in practice. The randomization test (sometimes called *permutation test*) avoids this normality assumption, solely based on the virtue of the randomization that the assignments of the treatment and the control are independent in the pair, i.e., conditional on $|Z_j| = z_j$, $Z_j = |Z_j|sgn(Z_j)$ is independent taking values $\pm|Z_j|$ with probability $1/2$, when treatment and control have no difference. Therefore, conditional on $z_1, z_2, ...$, the randomization t-test, based on the t-statistic $\sqrt{n-1}\bar{Z}_n/s_z$ where s_z^2 is $1/n\sum_{j=1}^n (Z_j - \bar{Z}_n)^2$, has a discrete distribution on 2^n equally likely values. We can simulate this distribution by the Monte Carlo method easily. Then if this statistic is large, there is strong evidence that treatment has large value. When n is large, such computation can be intimate, a better solution is to find an approximation. The Lindeberg-Feller central limit theorem can be applied if we assume

$$\max_{j\le n} z_j^2 / \sum_{j=1}^n z_j^2 \to 0.$$

It can be shown that this statistic has an asymptotic normal distribution $N(0,1)$. The details can be found in Ferguson, page 29.

Example 3.13 In Ferguson, page 30, an example of applying the central limit theorem is given for the signed-rank test for paired comparisons. Interested readers can find more details there.

3.3.4 Delta method

In many situation, the statistics are not simply the summation of independent random variables but a transformation of the latter. In this case, the Delta method can be used to obtain a similar result to the central limit theorem.

Theorem 3.15 (Delta method) For random vector X and X_n in R^k , if there exists two constant a_n and μ such that $a_n(X_n - \mu) \to_d X$ and $a_n \to \infty$, then for any function $g : R^k \mapsto R^l$ such that g has a derivative at μ, denoted by $\nabla g(\mu)$

$$a_n(g(X_n) - g(\mu)) \to_d \nabla g(\mu) X.$$

†

Proof By the Skorohod representation, we can construct $\tilde{X}_n$ and $\tilde{X}$ such that $\tilde{X}_n \sim_d X_n$ and $\tilde{X} \sim_d X$ ($\sim_d$ means the same distribution) and $a_n(\tilde{X}_n - \mu) \to_{a.s.} \tilde{X}$. Then $a_n(g(\tilde{X}_n) - g(\mu)) \to_{a.s.} \nabla g(\mu)\tilde{X}$. We obtain the result. †

As a corollary of Theorem 3.15, if $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \sigma^2)$, then for any differentiable function $g(\cdot)$, $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \to_d N(0, g'(\mu)^2\sigma^2)$.

Example 3.14 Let $X_1, X_2, ...$ be i.i.d with fourth moment. An estimate of the sample variance is $s_n^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^2$. We can use the Delta method in deriving the asymptotic distribution of s_n^2. Denote m_k as the kth moment of X_1 for $k \leq 4$. Note that $s_n^2 = (1/n)\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i/n)^2$ and

$$\sqrt{n}\left[\begin{pmatrix} \bar{X}_n \\ (1/n)\sum_{i=1}^n X_i^2 \end{pmatrix} - \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}\right] \to_d N\left(0, \begin{pmatrix} m_2 - m_1 & m_3 - m_1 m_2 \\ m_3 - m_1 m_2 & m_4 - m_2^2 \end{pmatrix}\right),$$

we can apply the Delta method with $g(x, y) = y - x^2$ to obtain

$$\sqrt{n}(s_n^2 - Var(X_1)) \to_d N(0, m_4 - (m_2 - m_1^2)^2).$$

Example 3.15 Let $(X_1, Y_1), (X_2, Y_2), ...$ be i.i.d bivariate samples with finite fourth moment. One estimate of the correlation among X and Y is

$$\hat{\rho}_n = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}},$$

where $s_{xy} = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$, $s_x^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $s_y^2 = (1/n)\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. To derive the large sample distribution of $\hat{\rho}_n$, we can first obtain the large sample distribution of (s_{xy}, s_x^2, s_y^2) using the Delta method as in Example 3.14 then further apply the Delta method with $g(x, y, z) = x/\sqrt{yz}$. We skip the details.

Example 3.16 The example is taken from the Pearson's Chi-square statistic. Suppose that one subject falls into K categories with probabilities $p_1, ..., p_K$, where $p_1 + ... + p_K = 1$. We actually observe $n_1, ..., n_k$ subjects in these categories from $n = n_1 + ... + n_K$ i.i.d subjects. The Pearson's statistic is defined as

$$\chi^2 = n\sum_{k=1}^K (\frac{n_k}{n} - p_k)^2/p_k,$$

which can be treated as $\sum$(observed count − expected count)^2/expected count. To obtain the asymptotic distribution of χ^2, we note that $\sqrt{n}(n_1/n - p_1, ..., n_K/n - p_K)$ has an asymptotic multivariate normal distribution. Then we can apply the Delta method to $g(x_1, ..., x_K) = \sum_{i=1}^K x_k^2$.

3.4 Summation of Non-independent Random Variables

In statistical inference, one will also encounter the summation of non-independent random variables. Theoretical results of the large sample theory for general non-independent random variables do not exist but for some summations with special structure, we have the similar results to the central limit theorem. These special cases include the U-statistics, the rank statistics, and the martingales.

3.4.1 U-statistics

We suppose $X_1, ..., X_n$ are i.i.d. random variables.

Definition 3.6 A *U-statistics* associated with $\tilde{h}(x_1, ..., x_r)$ is defined as

$$U_n = \frac{1}{r!\binom{n}{r}} \sum_{\beta} \tilde{h}(X_{\beta_1}, ..., X_{\beta_r}),$$

where the sum is taken over the set of all unordered subsets β of r different integers chosen from $\{1, ..., n\}$. †

One simple example is $\tilde{h}(x, y) = xy$. Then $U_n = (n(n-1))^{-1} \sum_{i \neq j} X_i X_j$. Many examples of U statistics arise from rank-based statistical inference. If let $X_{(1)}, ..., X_{(n)}$ be the ordered random variables of $X_1, ..., X_n$, one can see

$$U_n = E[\tilde{h}(X_1, ..., X_r)|X_{(1)}, ..., X_{(n)}].$$

Clearly, U_n is the summation of non-independent random variables.

If define $h(x_1, ..., x_r)$ as $(r!)^{-1} \sum_{(\tilde{x}_1, ..., \tilde{x}_r) \text{ is permutation of } (x_1, ..., x_r)} \tilde{h}(\tilde{x}_1, ..., \tilde{x}_r)$, then $h(x_1, ..., x_r)$ is permutation-symmetric and moreover,

$$U_n = \frac{1}{\binom{n}{r}} \sum_{\beta_1 < ... < \beta_r} h(\beta_1, ..., \beta_r).$$

In the last expression, h is called the *kernel* of the U-statistic U_n.

The following theorem says that the limit distribution of U is the same as the limit distribution of a sum of i.i.d random variables. Thus, the central limit theorem can be applied to U.

Theorem 3.16 Let $\mu = E[h(X_1, ..., X_r)]$. If $E[h(X_1, ..., X_r)^2] < \infty$, then

$$\sqrt{n}(U_n - \mu) - \sqrt{n} \sum_{i=1}^{n} E[U_n - \mu|X_i] \to_p 0.$$

Consequently, $\sqrt{n}(U_n - \mu)$ is asymptotically normal with mean zero and variance $r^2\sigma^2$, where, with $X_1, ..., X_r, \tilde{X}_1, ..., \tilde{X}_r$ i.i.d variables,

$$\sigma^2 = Cov(h(X_1, X_2, ..., X_r), h(X_1, \tilde{X}_2, ..., \tilde{X}_r)).$$

†

To prove Theorem 3.16, we need the following lemmas. Let $\mathcal{S}$ be a linear space of random variables with finite second moments that contain the constants; i.e., $1 \in \mathcal{S}$ and for any $X, Y \in \mathcal{S}$, $aX + bY \in \mathcal{S}_n$ where a and b are constants. For random variable T, a random variable S is called the *projection* of T on $\mathcal{S}$ if $E[(T - S)^2]$ minimizes $E[(T - \tilde{S})^2], \tilde{S} \in \mathcal{S}$.

Proposition 3.6 Let $\mathcal{S}$ be a linear space of random variables with finite second moments. Then S is the projection of T on $\mathcal{S}$ if and only if $S \in \mathcal{S}$ and for any $\tilde{S} \in \mathcal{S}$, $E[(T - S)\tilde{S}] = 0$.

Every two projections of T onto $\mathcal{S}$ are almost surely equal. If the linear space $\mathcal{S}$ contains the constant variable, then $E[T] = E[S]$ and $Cov(T - S, \tilde{S}) = 0$ for every $\tilde{S} \in \mathcal{S}$. †

Proof For any S and $\tilde{S}$ in $\mathcal{S}$,

$$E[(T - \tilde{S})^2] = E[(T - S)^2] + 2E[(T - S)\tilde{S}] + E[(S - \tilde{S})^2].$$

Thus, if S satisfies that $E[(T - S)\tilde{S}] = 0$, then $E[(T - \tilde{S})^2] \geq E[(T - S)^2]$. Thus, S is the projection of T on $\mathcal{S}$. On the other hand, if S is the projection, for any constant α, $E[(T - S - \alpha\tilde{S})^2]$ is minimized at $\alpha = 0$. Calculate the derivative at $\alpha = 0$ and we obtain $E[(T - S)\tilde{S}] = 0$.

If T has two projections S_1 and S_2, then from the above argument, we have $E[(S_1 - S_2)^2] = 0$. Thus, $S_1 = S_2, a.s.$ If the linear space $\mathcal{S}$ contains the constant variable, we choose $\tilde{S} = 1$. Then $0 = E[(T - S)\tilde{S}] = E[T] - E[S]$. Clearly, $Cov(T - S, \tilde{S}) = E[(T - S)\tilde{S}] = 0$. †

Proposition 3.7 Let $\mathcal{S}_n$ be linear space of random variables with finite second moments that contain the constants. Let T_n be random variables with projections S_n on to $\mathcal{S}_n$. If $Var(T_n)/Var(S_n) \to 1$ then

$$Z_n \equiv \frac{T_n - E[T_n]}{\sqrt{Var(T_n)}} - \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} \to_p 0.$$

†

Proof $E[Z_n] = 0$. Note that

$$Var(Z_n) = 2 - 2\frac{Cov(T_n, S_n)}{\sqrt{Var(T_n)Var(S_n)}}.$$

Since S_n is the projection of T_n, $Cov(T_n, S_n) = Cov(T_n - S_n, S_n) + Var(S_n) = Var(S_n)$. We have

$$Var(Z_n) = 2(1 - \sqrt{\frac{Var(S_n)}{Var(T_n)}}) \to 0.$$

By the Markov's inequality, we conclude that $Z_n \to_p 0$. †

The above lemma implies that if S_n is the summation of i.i.d random variables such that $(S_n - E[S_n])/\sqrt{Var(S_n)} \to_d N(0, \sigma^2)$, so is $(T_n - E[T_n])/\sqrt{Var(T_n)}$. The limit distribution of U-statistics is derived using this lemma.

We now start to prove Theorem 3.16.

Proof Let $\tilde{X}_1, ..., \tilde{X}_r$ be random variables with the same distribution as X_1 and they are independent of $X_1, ..., X_n$. Denote $\tilde{U}_n$ by $\sum_{i=1}^n E[U - \mu|X_i]$. We show that $\tilde{U}_n$ is the projection of U_n on the linear space $\mathcal{S}_n = \{g_1(X_1) + ... + g_n(X_n) : E[g_k(X_k)^2] < \infty, k = 1, ..., n\}$, which contains the constant variables. Clearly, $\tilde{U}_n \in \mathcal{S}_n$. For any $g_k(X_k) \in \mathcal{S}_n$,

$$E[(U_n - \tilde{U}_n)g_k(X_k)] = E[E[U_n - \tilde{U}_n|X_k]g_k(X_k)] = 0.$$

In fact, we can easily see that

$$\tilde{U}_n = \sum_{i=1}^n \frac{\binom{n-1}{r-1}}{\binom{n}{r}} E[h(\tilde{X}_1, ..., \tilde{X}_{r-1}, X_i) - \mu | X_i] = \frac{r}{n} \sum_{i=1}^n E[h(\tilde{X}_1, ..., \tilde{X}_{r-1}, X_i) - \mu | X_i].$$

Thus,

$$Var(\tilde{U}_n) = \frac{r^2}{n^2} \sum_{i=1}^n E[(E[h(\tilde{X}_1, ..., \tilde{X}_{r-1}, X_i) - \mu | X_i])^2]$$

$$= \frac{r^2}{n} Cov(E[h(\tilde{X}_1, ..., \tilde{X}_{r-1}, X_1) | X_1], E[h(\tilde{X}_1, ..., \tilde{X}_{r-1}, X_1) | X_1])$$

$$= \frac{r^2}{n} Cov(h(X_1, \tilde{X}_2, ..., \tilde{X}_r), h(X_1, X_2 ..., X_r)) = \frac{r^2 \sigma^2}{n},$$

where we use the equation

$$Cov(X, Y) = Cov(E[X|Z], E[Y|Z]) + E[Cov(X, Y|Z)].$$

Furthermore,

$$\begin{aligned} Var(U_n) &= \binom{n}{r}^{-2} \sum_{\beta} \sum_{\beta'} Cov(h(X_{\beta_1}, ..., X_{\beta_r}), h(X_{\beta'_1}, ..., X_{\beta'_r})) \\ &= \binom{n}{r}^{-2} \sum_{k=1}^r \sum_{\beta \text{ and } \beta' \text{ share } k \text{ components}} \\ &\qquad Cov(h(X_1, X_2, .., X_k, X_{k+1}, ..., X_r), h(X_1, X_2, ..., X_k, \tilde{X}_{k+1}, ..., \tilde{X}_r)). \end{aligned}$$

Since the number of β and β' sharing k components is equal to $\binom{n}{r}\binom{r}{k}\binom{n-r}{r-k}$, we obtain

$$Var(U_n) = \sum_{k=1}^r \frac{r!}{k!(r-k)!} \frac{(n-r)(n-r+1)\cdots(n-2r+k+1)}{n(n-1)\cdots(n-r+1)}$$

$$\times Cov(h(X_1, X_2, .., X_k, X_{k+1}, ..., X_r), h(X_1, X_2, ..., X_k, \tilde{X}_{k+1}, ..., \tilde{X}_r)).$$

The dominating term in U_n is the first term of order $1/n$ while the other terms are of order $1/n^2$. That is,

$$Var(U_n) = \frac{r^2}{n} Cov(h(X_1, X_2, ..., X_r), h(X_1, \tilde{X}_2, ..., \tilde{X}_r)) + O(\frac{1}{n^2}).$$

We conclude that $Var(U_n)/Var(\tilde{U}_n) \to 1$. From Proposition 3.7, it holds that

$$\frac{U_n - \mu}{\sqrt{Var(U_n)}} - \frac{\tilde{U}_n}{\sqrt{Var(\tilde{U}_n)}} \to_p 0.$$

Theorem 3.16 thus holds. †

Example 3.17 In a bivariate i.i.d sample $(X_1, Y_1), (X_2, Y_2), ...$, one statistic of measuring the agreement is called *Kendall's τ-statistic* given as

$$\hat{\tau} = \frac{4}{n(n-1)} \sum \sum_{i<j} I\left\{(Y_j - Y_i)(X_j - X_i) > 0\right\} - 1.$$

It can be seen that $\hat{\tau} + 1$ is a U-statistic of order 2 with the kernel

$$2I\left\{(y_2 - y_1)(x_2 - x_1) > 0\right\}.$$

Hence, by the above central limit theorem, $\sqrt{n}(\hat{\tau}_n + 1 - 2P((Y_2 - Y_1)(X_2 - X_1) > 0))$ has an asymptotic normal distribution with mean zero. The asymptotic variance can be computed as in Theorem 3.16.

3.4.2 Rank statistics

For a sequence of i.i.d random variables $X_1, ..., X_n$, we can order them from the smallest to the largest and denote by $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. The latter is called *order statistics* of the original sample. The *rank statistics*, denoted by $R_1, ..., R_n$ are the ranks of X_i among $X_1, ..., X_n$. Thus, if all the X's are different, $X_i = X_{(R_i)}$. When there are ties, R_i is defined as the average of all indices such that $X_i = X_{(j)}$ (sometimes called *midrank*). To avoid possible ties, we only consider the case that X's have continuous densities.

By name, a *rank statistic* is any function of the ranks. A linear rank statistic is a rank statistic of the special form $\sum_{i=1}^n a(i, R_i)$ for a given matrix $(a(i,j))_{n\times n}$. If $a(i,j) = c_i a_j$, then such statistic with form $\sum_{i=1}^n c_i a_{R_i}$ is called *simple linear rank statistic*, which will be our concern in this section. Here, c and a's are called the *coefficients* and *scores*.

Example 3.18 In two independent sample $X_1, ..., X_n$ and $Y_1, ..., Y_m$, a Wilcoxon statistic is defined as the summation of all the ranks of the second sample in the pooled data $X_1, ..., X_n$, $Y_1, ..., Y_m$, i.e.,

$$W_n = \sum_{i=n+1}^{n+m} R_i.$$

This is a simple linear rank statistic with c's are 0 and 1 for the first sample and the second sample respectively and the vector a is $(1, ..., n+m)$. There are other choices for rank statistics, for instance, the van der Waerden statistic $\sum_{i=n+1}^{n+m} \Phi^{-1}(R_i)$.

For order statistics and rank statistics, there are some useful properties.

Proposition 3.8 Let $X_1, ..., X_n$ be a random sample from continuous distribution function F with density f. Then

1. the vectors $(X_{(1)}, ..., X_{(n)})$ and $(R_1, ..., R_n)$ are independent;
2. the vector $(X_{(1)}, ..., X_{(n)})$ has density $n! \prod_{i=1}^n f(x_i)$ on the set $x_1 < ... < x_n$;
3. the variable $X_{(i)}$ has density $\binom{n-1}{i-1} F(x)^{i-1}(1-F(x))^{n-i} f(x)$; for F the uniform distribution on $[0,1]$, it has mean $i/(n+1)$ and variance $i(n-i+1)/[(n+1)^2(n+2)]$;

4. the vector $(R_1, ..., R_n)$ is uniformly distributed on the set of all $n!$ permutations of $1, 2, ..., n$;

5. for any statistic T and permutation $r = (r_1, ..., r_n)$ of $1, 2, ..., n$,

$$E[T(X_1, ..., X_n)|(R_1, .., R_n) = r] = E[T(X_{(r_1)}, .., X_{(r_n)})];$$

6. for any simple linear rank statistic $T = \sum_{i=1}^n c_i a_{R_i}$,

$$E[T] = n\bar{c}_n\bar{a}_n, \quad Var(T) = \frac{1}{n-1}\sum_{i=1}^n (c_i - \bar{c}_n)^2 \sum_{i=1}^n (a_i - \bar{a}_n)^2.$$

†

The proof of Proposition 3.8 is elementary so we skip. For simple linear rank statistic, a central limit theorem also exists:

Theorem 3.17 Let $T_n = \sum_{i=1}^n c_i a_{R_i}$ such that

$$\max_{i\le n} |a_i - \bar{a}_n| / \sqrt{\sum_{i=1}^n (a_i - \bar{a}_n)^2} \to 0, \quad \max_{i\le n} |c_i - \bar{c}_n| / \sqrt{\sum_{i=1}^n (c_i - \bar{c}_n)^2} \to 0.$$

Then $(T_n - E[T_n])/\sqrt{Var(T_n)} \to_d N(0,1)$ if and only if for every $\epsilon > 0$,

$$\sum_{(i,j)} I\left\{\sqrt{n}\frac{|a_i - \bar{a}_n||c_i - \bar{c}_n|}{\sqrt{\sum_{i=1}^n (a_i - \bar{a}_n)^2 \sum_{i=1}^n (c_i - \bar{c}_n)^2}} > \epsilon\right\} \frac{|a_i - \bar{a}_n|^2 |c_i - \bar{c}_n|^2}{\sum_{i=1}^n (a_i - \bar{a}_n)^2 \sum_{i=1}^n (c_i - \bar{c}_n)^2} \to 0.$$

We can immediately recognize that the last condition is similar to the Lindeberg condition. The proof can be found in Ferguson, Chapter 12.

Besides of rank statistics, there are other statistics based on ranks. For example, a simple linear *signed rank statistic* has the form

$$\sum_{i=1}^n a_{R_i^+} \mathrm{sign}(X_i),$$

where $R_1^+, ..., R_n^+$, called *absolute rank*, are the ranks of $|X_1|, ..., |X_n|$. In a bivariate sample $(X_1, Y_1), ..., (X_n, Y_n)$, one can define a statistic of the form

$$\sum_{i=1}^n a_{R_i} b_{S_i}$$

for two constant vector $(a_1, ..., a_n)$ and $(b_1, ..., b_n)$, where $(R_1, ..., R_n)$ and $(S_1, ..., S_n)$ are respective ranks of $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$. Such a statistic is useful for testing independence of X and Y. Another statistic is based on permutation test, as exemplified in Example 3.12. For all these statistics, some conditions ensure that the central limit theorem holds.

3.4.3 Martingales

In this section, we consider the central limit theorem for another type of the sum of non-independent random variables. These random variables are called martingale.

Definition 3.7 Let $\{Y_n\}$ be a sequence of random variables and $\mathcal{F}_n$ be sequence of σ-fields such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ....$ Suppose $E[|Y_n|] < \infty$. Then the sequence of pairs $\{(Y_n, \mathcal{F}_n)\}$ is called a *martingale* if

$$E[Y_n|\mathcal{F}_{n-1}] = Y_{n-1}, \quad a.s.$$

$\{(Y_n, \mathcal{F}_n)\}$ is a *submartingale* if

$$E[Y_n|\mathcal{F}_{n-1}] \geq Y_{n-1}, \quad a.s.$$

$\{(Y_n, \mathcal{F}_n)\}$ is a *supmartingale* if

$$E[Y_n|\mathcal{F}_{n-1}] \leq Y_{n-1}, \quad a.s.$$

†

The definition implies that $Y_1, ..., Y_n$ are measurable in $\mathcal{F}_n$. Sometimes, we say Y_n is adapted to $\mathcal{F}_n$. One simple example of a martingale is $Y_n = X_1 + ... + X_n$, where $X_1, X_2, ...$ are i.i.d with mean zero, and $\mathcal{F}_n$ is the σ-filed generated by $X_1, ..., X_n$. This is because

$$E[Y_n|\mathcal{F}_{n-1}] = E[X_1 + ... + X_n|X_1, ..., X_{n-1}] = Y_{n-1}.$$

For $Y_n = X_1^2 + ... + X_n^2$, one can verify that $\{(Y_n, \mathcal{F}_n)\}$ is a submartingale. In fact, from one submartingale, one can construct many submartingales as shown in the following lemma.

Proposition 3.9 Let $\{(Y_n, \mathcal{F}_n)\}$ be a martingale. For any measurable and convex function ϕ, $\{(\phi(Y_n), \mathcal{F}_n)\}$ is a submartingale. †

Proof Clearly, $\phi(Y_n)$ is adapted to $\mathcal{F}_n$. It is sufficient to show

$$E[\phi(Y_n)|\mathcal{F}_{n-1}] \geq \phi(Y_{n-1}).$$

This follows from the well-known *Jensen's inequality*: for any convex function ϕ,

$$E[\phi(Y_n)|\mathcal{F}_{n-1}] \geq \phi(E[Y_n|\mathcal{F}_{n-1}]) = \phi(Y_{n-1}).$$

†

Particularly, the Jensen's inequality is given in the following lemma.

Proposition 3.10 For any random variable X and any convex measurable function ϕ,

$$E[\phi(X)] \geq \phi(E[X]).$$

†

Proof We first claim that for any x_0, there exists a constant k_0 such that for any x,

$$\phi(x) \geq \phi(x_0) + k_0(x - x_0).$$

The line $\phi(x_0) + k_0(x - x_0)$ is called the supporting line for $\phi(x)$ at x_0. By the convexity, we have that for any $x' < y' < x_0 < y < x$,

$$\frac{\phi(x_0) - \phi(x')}{x_0 - x'} \leq \frac{\phi(y) - \phi(x_0)}{y - x_0} \leq \frac{\phi(x) - \phi(x_0)}{x - x_0}.$$

Thus, $\frac{\phi(x)-\phi(x_0)}{x-x_0}$ is bounded and decreasing as x decreases to x_0. Let the limit be k_0^+ then

$$\frac{\phi(x) - \phi(x_0)}{x - x_0} \geq k_0^+.$$

I.e.,

$$\phi(x) \geq k_0^+(x - x_0) + \phi(x_0).$$

Similarly,

$$\frac{\phi(x') - \phi(x_0)}{x' - x_0} \leq \frac{\phi(y') - \phi(x_0)}{y' - x_0} \leq \frac{\phi(x) - \phi(x_0)}{x - x_0}.$$

Then $\frac{\phi(x')-\phi(x_0)}{x'-x_0}$ is increasing and bounded as x' increases to x_0. Let the limit be k_0^- then

$$\phi(x') \geq k_0^-(x' - x_0) + \phi(x_0).$$

Clearly, $k_0^+ \geq k_0^-$. Combining those two inequalities, we obtain

$$\phi(x) \geq \phi(x_0) + k_0(x - x_0)$$

for $k_0 = (k_0^+ + k_0^-)/2$. We choose $x_0 = E[X]$ then

$$\phi(X) \geq \phi(E[X]) + k_0(X - E[X]).$$

The Jensen's inequality holds by taking the expectation on both sides. †

If $\{(Y_n, \mathcal{F}_n)\}$ is a submartingale, we can write

$$\begin{aligned} Y_n &= \sum_{j=1}^{n} (Y_j - E[Y_j|\mathcal{F}_{j-1}]) \\ &\quad + \sum_{j=1}^{n} E[Y_j - Y_{j-1}|\mathcal{F}_{j-1}] \\ &= M_n + A_n, \end{aligned}$$

where $\mathcal{F}_0$ is the null σ-field and $Y_0 = EY_1$. Note that $\{M_n, \mathcal{F}_n)\}$ is a martingale and that A_n is measurable in $\mathcal{F}_{n-1}$. Thus any submartingale can be written as the summation of a martingale and a random variable predictable in $\mathcal{F}_{n-1}$. We now state the limit theorems for the martingales.

Theorem 3.18 (Martingale Convergence Theorem) Let $\{(X_n, \mathcal{F}_n)\}$ be submartingale. If $K = \sup_n E[|X_n|] < \infty$, then $X_n \to_{a.s.} X$ where X is a random variable satisfying $E[|X|] \leq K$. †

The proof needs the maximal inequality for a submartingale and the up-crossing inequality.

Proof We first prove the following maximal inequality: for $\alpha > 0$,

$$P(\max_{i\leq n} X_i \geq \alpha) \leq \frac{1}{\alpha} E[|X_n|].$$

To see that, we note that

$$\begin{aligned} & P(\max_{i\leq n} X_i \geq \alpha) \\ = & \sum_{i=1}^n P(X_1 < \alpha, ..., X_{i-1} < \alpha, X_i \geq \alpha) \\ \leq & \sum_{i=1}^n E[I(X_1 < \alpha, ..., X_{i-1} < \alpha, X_i \geq \alpha) \frac{X_i}{\alpha}] \\ = & \frac{1}{\alpha} \sum_{i=1}^n E[I(X_1 < \alpha, ..., X_{i-1} < \alpha, X_i \geq \alpha) X_i]. \end{aligned}$$

Since $E[X_n|X_1, ..., X_{n-1}] \geq X_{n-1}$, $E[X_n|X_1, ..., X_{n-2}] \geq E[X_{n-1}|X_1, ..., X_{n-2}]$ and so on. We obtain $E[X_n|X_1, ..., X_i] \geq E[X_{i+1}|X_1, ..., X_i] \geq X_i$ for $i = 1, ..., n-1$. Thus,

$$P(\max_{i\leq n} X_i \geq \alpha) \leq \frac{1}{\alpha} \sum_{i=1}^n E[I(X_1 < \alpha, ..., X_{i-1} < \alpha, X_i \geq \alpha) E[X_n|X_1, ..., X_i]]$$

$$\leq \frac{1}{\alpha} E[X_n \sum_{i=1}^n I(X_1 < \alpha, ..., X_{i-1} < \alpha, X_i \geq \alpha)] \leq \frac{1}{\alpha} E[X_n] \leq \frac{1}{\alpha} E[|X_n|].$$

For any interval $[\alpha, \beta]$ ($\alpha < \beta$), we define a sequence of numbers $\tau_1, \tau_2, ...$ as follows:
τ_1 is the smallest j such that $1 \leq j \leq n$ and $X_j \leq \alpha$ and is n if there is not such j;
τ_{2k} is the smallest j such that $\tau_{2k-1} < j \leq n$ and $X_j \geq \beta$, and is n if there is not such j;
τ_{2k+1} is the smallest j such $\tau_{2k} < j \leq n$ and $X_j \leq \alpha$, and is n if there is not such j.
A random variable U, called upcrossings of $[\alpha, \beta]$ by $X_1, ..., X_n$, is the largest i such that $X_{\tau_{2i-1}} \leq \alpha < \beta \leq X_{\tau_{2i}}$. We then show that

$$E[U] \leq \frac{E[|X_n|] + |\alpha|}{\beta - \alpha}.$$

Let $Y_k = \max\{0, X_k - \alpha\}$ and $\theta = \beta - \alpha$. It is easy to see $Y_1, ..., Y_n$ is a submartingale. The τ_k are unchanged if the definitions $X_j \leq \alpha$ is replaced by $Y_j = 0$ and $X_j \geq \beta$ by $Y_j \geq \theta$, and so U is also the number of upcrossings of $[0, \theta]$ by $Y_1, .., Y_n$. We also obtain

$$E[Y_{\tau_{2k+1}} - Y_{\tau_{2k}}] = \sum_{1\leq k_1<k_2\leq n} E[(Y_{k_2} - Y_{k_1}) I(\tau_{2k+1} = k_2, \tau_{2k} = k_1)]$$

$$= \sum_{k_1=1}^{n-1} \sum_{k'=2}^{n} E[I(\tau_{2k} = k_1, k_1 < k' \leq \tau_{2k+1})(Y_{k'} - Y_{k'-1})]$$

$$= \sum_{k_1=1}^{n-1} \sum_{k'=2}^{n} E[I(\tau_{2k} = k_1, k_1 < k')(1 - I(\tau_{2k+1} < k'))(Y_{k'} - Y_{k'-1})].$$

By the definition, if $\{\tau_{2k-1} = i\}$ is measurable in $\mathcal{F}_i$ for $i = 1, ..., n$, where $\mathcal{F}_i$ is the σ-field generated by $Y_1, ..., Y_i$, then

$$\{\tau_{2k} = j\} = \cup_{i=1}^{j-1} \{\tau_{2k-1} = i, Y_{i+1} < \theta, ..., Y_{j-1} \leq \theta, Y_j \geq \theta\}$$

belongs to the σ-field $\mathcal{F}_j$ and $\{\tau_{2k} = n\} = \{\tau_{2k} \leq n-1\}^c$ lies in $\mathcal{F}_n$. Similarly, if $\{\tau_{2k} = i\} \in \mathcal{F}_i$ for any $i = 1, ..., n$, so is $\{\tau_{2k+1} = i\} \in \mathcal{F}_i$ for any $i = 1, ..., n$. Thus, by the deduction, we obtain that for any $i = 1, ..., n$, $\{\tau_k = i\}$ is in $\mathcal{F}_i$. Then,

$$E[I(\tau_{2k} = k_1, k_1 < k')(1 - I(\tau_{2k+1} < k'))(Y_{k'} - Y_{k'-1})]$$

$$= E[I(\tau_{2k} = k_1, k_1 < k')(1 - I(\tau_{2k+1} < k'))(E[Y_{k'}|\mathcal{F}_{k'-1}] - Y_{k'-1})] \geq 0.$$

We conclude that $E[Y_{\tau_{2k+1}} - Y_{\tau_{2k}}] \geq 0$.

Since τ_k is strictly increasing and $\tau_n = n$,

$$Y_n = Y_{\tau_n} \geq Y_{\tau_n} - Y_{\tau_1} = \sum_{k=2}^{n} (Y_{\tau_k} - Y_{\tau_{k-1}}) = \sum_{2 \leq k \leq n, k \text{ even}} (Y_{\tau_k} - Y_{\tau_{k-1}}) + \sum_{2 \leq k \leq n, k \text{ odd}} (Y_{\tau_k} - Y_{\tau_{k-1}}).$$

When k is even, $Y_{\tau_k} - Y_{\tau_k - 1} \geq \theta$ and the total number of such k is U. The expectation of the second half is non-negative. We obtain

$$E[Y_n] \geq \theta E[U].$$

Thus,

$$E[U] \leq \frac{\theta}{E}[Y_n] \leq \frac{E[|X| + |\alpha|}{\beta - \alpha}.$$

With the maximal inequality, we can start to prove the martingale convergence theorem. Let U_n be the number of upcrossings of $[\alpha, \beta]$ by $X_1, ..., X_n$. Then

$$E[U_n] \leq \frac{K + |\alpha|}{\beta - \alpha}.$$

Let $X^* = \limsup_n X_n$ and $X_* = \liminf_n X_n$. If $X_* < \alpha < \beta < X^*$, then U_n must go to infinity. Since U_n is bounded with probability 1, $P(X_* < \alpha < \beta < X^*) = 0$. Now

$$\{X_* < X^*\} = \cup_{\alpha < \beta, \alpha, \beta \text{ are rational numbers}} \{X_* < \alpha < \beta < X^*\}.$$

We obtain $P(X_* = X^*) = 1$. That is, X_n converges to their common values X. By the Fatou's lemma, $E[|X|] \leq \liminf_n E[|X_n|] \leq K$. X is integrable and finite with probability 1. † .

As a corollary of the martingale convergence theorem, we obtain

Corollary 3.1 If $\mathcal{F}_n$ is increasing σ-field and denote $\mathcal{F}_\infty$ as the σ-field generated by $\cup_{n=1}^\infty \mathcal{F}_n$, then for any random variable Z with $E[|Z|] < \infty$, it holds

$$E[Z|\mathcal{F}_n] \rightarrow_{a.s.} E[Z|\mathcal{F}_\infty].$$

†

Proof Denote $Y_n = E[Z_n|\mathcal{F}_n]$. Clearly, Y_n is a martingale adapted to $\mathcal{F}_n$. Moreover, $E[|Y_n|] \leq E[|Z|]$. By the martingale convergence theorem, Y_n converges to some random variable Y almost surely. Clearly, Y is measurable in $\mathcal{F}_\infty$. We then show Y_n is uniformly integrable. Since $Y_n \leq E[|Z_n||\mathcal{F}_n]$, we may assume Z is non-negative. For any $\epsilon > 0$, there exists a δ such that $E[ZI_A] < \epsilon$ whenever $P(A) < \delta$ (since the measure $E[ZI_A]$ is absolutely continuous with respect to the measure P). Note that for a large α, consider the set $A = \{P(E[Z|\mathcal{F}_n] \geq \alpha)\}$. Since

$$P(A) = E[I(E[Z|\mathcal{F}_n] \geq \alpha)] \leq \frac{1}{\alpha}E[Z],$$

we can choose α large enough (independent of n) such that $P(A) < \delta$. Thus, $E[ZI(E[Z|\mathcal{F}_n] \geq \alpha)] < \epsilon$ for any n. We conclude $E[Z|\mathcal{F}_n]$ is uniformly integrable. With the uniform integrability, we have that for any $A \in \mathcal{F}_k$, $\lim_n \int_A Y_n dP = \int_A Y dP$. Note that $\int_A Y_n dP = \int_A Z dP$ for $n > k$. Thus, $\int_A Y dP = \int_A Z dP = \int_A E[Z|\mathcal{F}_\infty] dP$. This is true for any $A \in \cup_{n=1}^\infty \mathcal{F}_\infty$ so it is also true for any $A \in \mathcal{F}_\infty$. Since Y is measurable in $\mathcal{F}_\infty$, $Y = E[Z|\mathcal{F}_\infty], a.s.$ †

Finally, a similar theorem to the Lindeberg-Feller central limit theorem also exists for the martingales.

Theorem 3.19 (Martingale Central Limit Theorem) Let $(Y_{n1}, \mathcal{F}_{n1}), (Y_{n2}, \mathcal{F}_{n2}), ...$ be a martingale. Define $X_{nk} = Y_{nk} - Y_{n,k-1}$ with $Y_{n0} = 0$ thus $Y_{nk} = X_{n1} + ... + X_{nk}$. Suppose that

$$\sum_k E[X_{nk}^2|\mathcal{F}_{n,k-1}] \rightarrow_p \sigma^2$$

where σ is a positive constant and that

$$\sum_k E[X_{nk}^2 I(|X_{nk}| \geq \epsilon)|\mathcal{F}_{n,k-1}] \rightarrow_p 0$$

for each $\epsilon > 0$. Then

$$\sum_k X_{nk} \rightarrow_d N(0, \sigma^2).$$

†

The proof is based on the approximation of the characteristic function and we skip the details here.

3.5 Some Notation

In a probability space $(\Omega, \mathcal{A}, P)$, let $\{X_n\}$ be random variables (random vectors). We introduce the following notation: $X_n = o_p(1)$ denotes that X_n converges in probability to zero, $X_n = O_p(1)$ denotes that X_n is bounded in probability; i.e.,

$$\lim_{M\to\infty} \limsup_n P(|X_n| \geq M) = 0.$$

It is easy to see $X_n = O_p(1)$ is equivalent to saying X_n is uniformly tight. Furthermore, for a sequence of random variable $\{r_n\}$, $X_n = o_p(r_n)$ means that $|X_n|/r_n \to_p 0$ and $X_n = O_p(r_n)$ means that $|X_n|/r_n$ is bounded in probability.

There are many rules of calculus with o and O symbols. For instance, some commonly used formulae are (R_n is a deterministic sequence)

$$o_p(1) + o_p(1) = o_p(1), \quad O_p(1) + O_p(1) = O_p(1), \quad O_p(1)o_p(1) = o_p(1),$$

$$(1 + o_p(1))^{-1} = 1 + o_p(1), \quad o_p(R_n) = R_n o_p(1), \quad O_p(R_n) = R_n O_p(1),$$

$$o_p(O_p(1)) = o_p(1).$$

Furthermore, if a real function $R(\cdot)$ satisfies that $R(h) = o(|h|^p)$ as $h \to 0$, then $R(X_n) = o_p(|X_n|^p)$; if $R(h) = O(|h|^p)$ as $h \to 0$, then $R(X_n) = O_p(|X_n|^p)$. Readers should be able to prove these results without difficulty.

READING MATERIALS: You should read Lehmann and Casella, Section 1.8, Ferguson, Part 1, Part 2, Part 3 12-15

PROBLEMS

1. (a) If $X_1, X_2, ...$ are i.i.d $N(0,1)$, then $X_{(n)}/\sqrt{2\log n} \to_p 1$ where $X_{(n)}$ is the maximum of $X_1, ..., X_n$. *Hint*: use the following inequality: for any $\delta > 0$,

$$\frac{\delta}{\sqrt{2\pi}} e^{-(1+\delta)y^2/2} y \leq \int_y^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \leq \frac{e^{-y^2(1-\delta)/2}}{\sqrt{\delta}}.$$

   (b) If $X_1, X_2, ...$ are i.i.d $Uniform(0,1)$, derive the limit distribution of $n(1 - X_{(n)})$.

2. Suppose that $U \sim Uniform(0,1), \alpha > 0$, and

$$X_n = (n^\alpha / \log(n+1)) I_{[0, n^{-\alpha}]}(U).$$

   (a) Show that $X_n \to_{a.s.} 0$ and $E[X_n] \to 0$.

   (b) Can you find a random variable Y with $|X_n| \leq Y$ for all n with $E[Y] < \infty$?

(c) For what values of α does the uniform integrability condition

$$\limsup_{n\to\infty} E[|X_n| I_{|X_n|\geq M}] \to 0, \quad \text{as } M \to \infty$$

hold?

3. (a) Show by example that distribution functions having densities can converge in distribution even if the densities do not converge. *Hint*: Consider $f_n(x) = 1 + \cos 2\pi nx$ in $[0, 1]$.

   (b) Show by example that distributions with densities can converge in distribution to a limit that has no density.

   (c) Show by example that discrete distributions can converge in distribution to a limit that has a density.

4. *Stirling's formula.* Let $S_n = X_1 + ... + X_n$, where the $X_1, ..., X_n$ are independent and each has the Poisson distribution with parameters 1. Calculate or prove successively:

   (a) Calculate the expectation of $\{(S_n - n)/\sqrt{n}\}^-$, the negative part of $(S_n - n)/\sqrt{n}$.

   (b) Show $\{(S_n - n)/\sqrt{n}\}^- \to_d Z^-$, where Z has a standard normal distribution.

   (c) Show

$$E\left[\left\{\frac{S_n - n}{\sqrt{n}}\right\}^-\right] \to E[Z^-].$$

   (d) Use the above results to derive the Stirling's formula:

$$n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n}.$$

5. This problem gives an alternative way of proving the Slutsky theorem. Let $X_n \to_d X$ and $Y_n \to_p y$ for some constant y. Assume X_n and Y_n are both measurable functions on the same probability measure space $(\Omega, \mathcal{A}, P)$. Then $(X_n, Y_n)'$ can be considered as a bivariate random variable into R^2.

   (a) Show $(X_n, Y_n)' \to_d (X, y)'$. *Hint*: show the characteristic function of $(X_n, Y_n)'$ converges using the dominated convergence theorem.

   (b) Use the continuous mapping theorem to prove the Slutsky theorem. *Hint*: first show $Z_n X_n \to_d zX$ using the function $g(x, z) = xz$; then show $Z_n X_n + Y_n \to_d zX + y$ using the function $\tilde{g}(x, y) = x + y$.

6. Suppose that $\{X_n\}$ is a sequence of random variables in a probability measure space. Show that, if $E[g(X_n)] \to E[g(X)]$ for all continuous g with bounded support (that is, $g(x)$ is zero when x is outside a bounded interval), then $X_n \to_d X$. *Hint*: verify (c) of the Portmanteau Theorem. Follow the proof for (c) by considering $g(x) = 1 - \epsilon/[\epsilon + d(x, G^c \cup (-M, M)^c)]$ for any M.

7. Suppose that $X_1, ..., X_n$ are i.i.d with distribution function $G(x)$. Let $M_n = \max\{X_1, .., X_n\}$.

   (a) If $G(x) = (1 - \exp\{-\alpha x\})I(x > 0)$, what is the limit distribution of $M_n - \alpha^{-1} \log n$?

(b) If

$$G(x) = \begin{cases} 0 & \text{if } x \leq 1, \\ 1 - x^{-\alpha} & \text{if } x \geq 1, \end{cases}$$

where $\alpha > 0$, what is the limit distribution of $n^{-1/\alpha} M_n$?

(c) If

$$G(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - (1-x)^{\alpha} & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x \geq 1, \end{cases}$$

where $\alpha > 0$, what is the limit distribution of $n^{1/\alpha}(M_n - 1)$?

8. (a) Suppose that $X_1, X_2, ...$ are i.i.d in R^2 with distribution giving probability θ_1 to $(1,0)$, probability θ_2 to $(0,1)$, θ_3 to $(0,0)$ and θ_4 to $(-1,-1)$ where $\theta_j \geq 0$ for $j = 1,2,3,4$ and $\theta_1 + ... + \theta_4 = 1$. Find the limiting distribution of $\sqrt{n}(\bar{X}_n - E[X_1])$ and describe the resulting approximation to the distribution of $\bar{X}_n$.

(b) Suppose that $X_1, ..., X_n$ is a sample from the Poisson distribution with parameter $\lambda > 0$: $P(X_1 = k) = \exp\{-\lambda\}\lambda^k/k!$, $k = 0, 1, ...$ Let $Z_n = [\sum_{i=1}^n I(X_i = 1)]/n$. What is the joint asymptotic distribution of $\sqrt{n}((\bar{X}_n, Z_n)' - (\lambda, \lambda e^{-\lambda}))$? Let $p_1(\lambda) = P(X_1 = 1)$. What is the asymptotic distribution of $\hat{p}_1 = p_1(\bar{X}_n)$? What is the joint asymptotic distribution of $(Z_n, \hat{p}_1)$ (after centering and rescaling)?

(c) If X_n possesses a t-distribution with n degrees of freedom, then $X_n \to_d N(0,1)$ as $n \to \infty$. Show this.

9. Suppose that X_n converges in distribution to X. Let $\phi_n(t)$ and $\phi(t)$ be the characteristic functions of X_n and X respectively. We know that $\phi_n(t) \to \phi(t)$ for each t. The following procedure shows that if $\sup_n E[|X_n|] < C_0$ for some constant C_0, the convergence pointwise of the characteristic functions can be strengthened to the convergence uniformly in any bounded interval,

$$\sup_{|t|<M} |\phi_n(t) - \phi(t)| \to 0$$

for any constant M. Verify each of the following steps.

(a) Show that $E[|X_n|] = \int_0^\infty P(|X_n| \geq t)dt$ and $E[|X|] = \int_0^\infty P(|X| \geq t)dt$. *Hint*: write $P(|X_n| \geq t) = E[I(|X_n| \geq t)]$ then apply the Fubini-Tonelli theorem.

(b) Show that $P(|X_n| \geq t) \to P(|X| \geq t)$ almost everywhere (with respect to the Lebsgue measure). Then apply the Fatou's lemma to show that $E[|X|] \leq C_0$.

(c) Show that both $\phi_n(t)$ and $\phi(t)$ satisfy: for any t_1, t_2,

$$|\phi_n(t_1) - \phi_n(t_2)| \leq C_0|t_1 - t_2|,$$

$$|\phi(t_1) - \phi(t_2)| \leq C_0|t_1 - t_2|.$$

That is, ϕ_n and ϕ are uniformly continuous.

(d) Show that $\sup_{t \in [-M,M]} |\phi_n(t) - \phi(t)| \to 0$. *Hint*: first partition $[-M, M]$ into equally spaced $-M = t_0 < t_1 < ... < t_m = M$; then for t in one of these intervals, say $[t_k, t_{k+1}]$, use the inequality

$$|\phi_n(t) - \phi(t)| \leq |\phi_n(t) - \phi_n(t_k)| + |\phi_n(t_k) - \phi(t_k)| + |\phi(t_k) - \phi(t)|.$$

10. Suppose that $X_1, ..., X_n$ are i.i.d from the uniform distribution in $[0,1]$. Derive the asymptotic distribution of *Gini's mean difference*, which is defined as $\binom{n}{2}^{-1} \sum\sum_{i<j} |X_i - X_j|$.

11. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d from a bivariate distribution with bounded fourth moments. Derive the limit distribution of $U = \binom{n}{2}^{-1} \sum\sum_{i<j} (Y_j - Y_i)(X_j - X_i)$. Write the expression in terms of the moments of (X_1, Y_1).

12. Let $Y_1, Y_2, ...$ be independent random variables with mean 0 and variance σ^2. Let $X_n = (\sum_{k=1}^n Y_k)^2 - n\sigma^2$ and show that $\{X_n\}$ is a martingale.

13. Suppose that $X_1, ..., X_n$ are independent $N(0,1)$ random variables, and let $Y_i = X_i^2$ for $i = 1, ..., n$. Thus $\sum_{i=1}^n Y_i^2 \sim \chi_n^2$.

    (a) Show that $\sqrt{n}(\bar{Y}_n - 1) \rightarrow_d N(0, \sigma^2)$ and find σ^2.

    (b) Show that for each $r > 0$, $\sqrt{n}(\bar{Y}_n^r - 1) \rightarrow_d N(0, V(r)^2)$ and find $V(r)^2$ as a function of r.

    (c) Show that
    $$\frac{\sqrt{n}\{\bar{Y}_n^{1/3} - (1 - 2/(9n))\}}{\sqrt{2/9}} \rightarrow_d N(0,1).$$
    Does this agree with your result in (b).

    (d) Make normal probability plots to compare the approximations in (a) and (c) (the transformation in (c) is called the "Wilson-Hilferty" transformation of a χ^2-random variable.

14. Suppose that $X_1, X_2, ...$ are i.i.d positive random variables, and define $\bar{X}_n = \sum_{i=1}^n X_i/n$, $H_n = 1/\{n^{-1}\sum_{i=1}^n (1/X_i)\}$, and $G_n = \{\prod_{i=1}^n X_i\}^{1/n}$ to be the arithmetic, harmonic and geometric means respectively. We know that $\bar{X}_n \rightarrow_{a.s.} E[X_1] = \mu$ if and only if $E[|X_i|]$ is finite.

    (a) Use the strong law of large numbers together with appropriate additional hypotheses to show that $H_n \rightarrow_{a.s.} 1/\{E[1/X_1]\} \equiv h$ and $G_n \rightarrow_{a.s.} \exp\{E[\log X_1]\} \equiv g$.

    (b) Find the joint limiting distribution of $\sqrt{n}(\bar{X}_n - \mu, H_n - h, G_n - g)$. You will need to impose or assume additional moment conditions to be able to prove this. Specify these additional assumptions carefully.

    (c) Suppose that $X_i \sim Gamma(r, \lambda)$ with $r > 0$. Find what values of r are the hypotheses you impose in (c) satisfied? Compute the covariance of the limiting distribution in (c) as explicitly as you can in this case.

    (d) Show that $\sqrt{n}(G_n/\bar{X}_n - g/\mu) \rightarrow_d N(0, V^2)$. Compute V explicitly when $X_i \sim Gamma(r, \lambda)$ with r satisfying the conditions you found in (d).

15. Suppose that $(N_{11}, N_{12}, N_{21}, N_{22})$ has multinomial distribution with (n, p) where $p = (p_{11}, p_{12}, p_{21}, p_{22})$ and $\sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1$. Thus, N's can be treated as counts in a 2× table. The log-odds ratio is defined by
    $$\psi = \log \frac{p_{12}p_{21}}{p_{11}p_{22}}.$$

(a) Suggest an estimator of ψ, say $\hat{\psi}_n$.

(b) Show that the estimator you proposed in (a) is asymptotically normal and compute the asymptotic variance of your estimator. *Hint:* The vectors of N's is the sum of n independent Multinomial(1, p) random vectors $\{Y_i, i = 1, ..., n\}$.

16. Suppose that $X_i \sim Bernoulli(p_i)$, $i = 1, .., n$ are independent. Show that if

$$\sum_{i=1}^{n} p_i(1-p_i) \to \infty,$$

then

$$\frac{\sqrt{n}(\bar{X}_n - \bar{p}_n)}{\sqrt{n^{-1}\sum_{i=1}^{n} p_i(1-p_i)}} \to_d N(0,1).$$

Give one example $\{p_i\}$ for which the above convergence in distribution holds and another example for which it fails.

17. Suppose that $X_1, ..., X_n$ are independent with common mean μ but with variances $\sigma_1^2, ..., \sigma_n^2$ respectively.

(a) Show that $\bar{X}_n \to_p \mu$ if $\sum_{i=1}^{n} \sigma_i^2 = o(n^2)$.

(b) Now suppose that $X_i = \mu + \sigma_i \epsilon_i$ where $\epsilon_1, ..., \epsilon_n$ are i.i.d with distribution function F with $E[\epsilon_1] = 0$ and $var(\epsilon_1) = 1$. Show that if

$$\max_{i \le n} \sigma_i^2 / \sum_{i=1}^{n} \sigma_i^2 \to 0$$

then with $\bar{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n} \sigma_i^2$,

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

Hence show that if furthermore $\bar{\sigma}^2 \to \sigma_0^2$, then $\sqrt{n}(\bar{X}_n - \mu) \to_d N(0, \sigma_0^2)$.

(c) If $\sigma_i^2 = Ai^r$ for some constant A, show that $\max_{i \le n} \sigma_i^2 / \sum_{i=1}^{n} \sigma_i^2 \to 0$ but $\bar{\sigma}_n^2$ has not limit. In this case, $n^{(1-r)/2}(\bar{X}_n - \mu) = O_p(1)$.

18. Suppose that $X_1, ..., X_n$ are independent with common mean μ but with variances $\sigma_1^2, ..., \sigma_n^2$ respectively, the same as the previous question. Consider the estimator of μ: $T_n = \sum_{i=1}^{n} \omega_{ni} X_i$, where $\omega = (\omega_{n1}, ..., \omega_{nn}))$ is a vector of weights with $\sum_{i=1}^{n} \omega_{ni} = 1$.

(a) Show that all the estimators T_n have the mean μ and the choice of weights minimizing $var(T_n)$ is

$$\omega_{ni}^{opt} = \frac{1/\sigma_i^2}{\sum_{j=1}^{n}(1/\sigma_j^2)}, \quad i = 1, ..., n.$$

(b) Compute $var(T_n)$ when $\omega = \omega^{opt}$ and show $T_n \to_p \mu$ if $\sum_{i=1}^{n}(1/\sigma_i^2) \to \infty$.

(c) Suppose $X_i = \mu + \sigma_i \epsilon_i$ where $\epsilon_1, ..., \epsilon_n$ are i.i.d with distribution function F with $E[\epsilon_1] = 0$ and $var(\epsilon_1) = 1$. Show that

$$\sqrt{\sum_{i=1}^{n}(1/\sigma_i^2)}(T_n - \mu) \rightarrow_d N(0,1)$$

if $\max_{i \leq n}(1/\sigma_i^2)/\sum_{j=1}^{n}(1/\sigma_j^2) \rightarrow 0$, where ω chosen as ω^{opt}.

(d) Compute $var(T_n)/var(\bar{X}_n)$ when $\omega = \omega^{opt}$ in the case $\sigma_i^2 = Ar^i$ for $r = 0.25, 0.5, 0.75$ and $n = 5, 10, 20, 50, 100, \infty$.

19. Ferguson, page 6 and page 7, problems 1-7

20. Ferguson, page 11 and page 12, problems 1-8

21. Ferguson, page 18, problems 1-5

22. Ferguson, page 23, page 24 and page 25, problems 1-8

23. Ferguson, page 34 and page 35, problems 1-10

24. Ferguson, page 42 and page 43, problems 1-6

25. Ferguson, page 49 and page 50, problems 1-6

26. Ferguson, page 54 and page 55, problems 1-4

27. Ferguson, page 60, problems 1-4

28. Ferguson, page 65 and page 66, problems 1-3

29. Read Ferguson, pages 87-92 and do problems 3-6

30. Ferguson, page 100, problems 1-2

31. Lehmann and Casella, page 75, problems 8.2, 8.3

32. Lehmann and Casella, page 76, problems 8.8, 8.10, 8.11, 8.12, 8.14, 8.15, 8.16, 8.17 8.18

33. Lehmann and Casella, page 77, problems 8.19, 8.20, 8.21, 8.22, 8.23, 8.24, 8.25, 8.26

CHAPTER 4 POINT ESTIMATION AND EFFICIENCY

The objective of science is to make general conclusions based on observed empirical data or phenomenon. The differences among different scientific areas are scientific tools implemented and scientific approaches to derive the decisions. However, they follow a similar procedure as follows:
(A) a class of mathematical models is proposed to model scientific phenomena or processes;
(B) an estimated model is derived using the empirical data;
(C) the obtained model is validated using more and new observations; if wrong, go back to (A).
Usually, in (A), the class of mathematical models is proposed based on either past experience or some physical laws. (B) is the step where all different scientific tools can play by using mathematical methods to determine the model. (C) is the step of model validation. Undoubtedly eac step is important.

In statistical science, (A) corresponds to proposing a class of distribution functions, denoted by $\mathcal{P}$, to describe the probabilistic mechanisms of data generation. (B) consists of all kinds of statistical methods to decide which distribution in the class of (A) fits the data best. (C) is how one can validate or test the goodness of the distribution obtained in (B). Our goal of this course is mainly on (B), which is called statistical inference step.

One good estimation approach should be able to estimate model parameters with reasonable accuracy. Such accuracy is characterized by either unbiasedness in finite sample performance or consistency in large sample performance. Furthermore, by accounting for randomness in data generation, we also want the estimation to be somewhat robust to intrinsic random mechanism. This robustness is characterized by the variance of the estimates. Thus, an ideally best estimator should have no bias and have the smallest variance in any finite sample. Unfortunately, although such estimators may exist for some models, most of models do not. One compromise is to seek an estimator which has no bias and has the smallest variance in large sample, i.e., an estimate which is asymptotically unbiased and efficient. Fortunately, such an estimator exists for most of models.

In this chapter, we review some commonly-used estimation approaches, with particular attention to the estimation providing the unbiased and smallest variance estimators if they exist. The smallest variance for finite sample is characterized by the Cramér-Rao bound (efficiency bound in finite sample). Such a bound also turns out to be the efficiency bound in large sample, where we show that the asymptotic variance of any regular estimators in regular models can not be smaller than this bound.

4.1 Introductory Examples

A *model* $\mathcal{P}$ is a collection of probability distributions for the data we observe. Parameters of interest are simply some functionals on $\mathcal{P}$, denoted by $\nu(P)$ for $P \in \mathcal{P}$.

Example 4.1 Suppose X is a non-negative random variable.
Case A. Suppose that $X \sim \text{Exponential}(\theta), \theta > 0$; thus $p_\theta(x) = \theta e^{-\theta x} I(x \geq 0)$. $\mathcal{P}$ consists of distribution function which are indexed by a finite-dimensional parameter θ. $\mathcal{P}$ is a parametric model. $\nu(p_\theta) = \theta$ is parameter of interest.
Case B. Suppose $\mathcal{P}$ consists of the distribution functions with density $p_{\lambda,G} = \int_0^\infty \lambda \exp\{-\lambda x\} dG(\lambda)$, where $\lambda \in R$ and G is any distribution function. Then $\mathcal{P}$ consists of the distribution functions

which are indexed by both real parameter λ and functional parameter G. $\mathcal{P}$ is a semiparametric model. $\nu(p_{\lambda,G}) = \lambda$ or G or both can be parameters of interest.
Case C. $\mathcal{P}$ consists of all distribution functions in $[0, \infty)$. $\mathcal{P}$ is a nonparametric model. $\nu(P) = \int x dP(x)$, the mean of the distribution function, can be parameter of interest.

Example 4.2 Suppose that $X = (Y, Z)$ is a random vector on $R^+ \times R^d$.
Case A. Suppose $X \sim P_\theta$ with $Y|Z = z \sim$ exponential$(\lambda e^{\theta' z})$ for $y \geq 0$. This is a parametric model with parameter space $\Theta = R^+ \times R^d$.
Case B. Suppose $X \sim P_{\theta,\lambda}$ with $Y|Z = z \sim \lambda(y)e^{\theta' z} \exp\{-\Lambda(y)e^{\theta' z}\}$ where $\Lambda(y) = \int_0^y \lambda(y)dy$. This is a semiparametric model, the Cox proportional hazards model for survival analysis, with parameter space $(\theta, \lambda) \in R \times \{\lambda(y) : \lambda(y) \geq 0, \int_0^\infty \lambda(y)dy = \infty\}$.
Case C. Suppose $X \sim P$ on $R^+ \times R^d$ where P is completely arbitrary. This is a nonparametric model.

Example 4.3 Suppose $X = (Y, Z)$ is a random vector in $R \times R^d$.
Case A. Suppose that $X = (Y, Z) \sim P_\theta$ with $Y = \theta' Z + \epsilon$ where $\theta \in R^d$ and $\epsilon \sim N(0, \sigma^2)$. This is a parametric model with parameter space $(\theta, \sigma) \in R^d \times R^+$.
Case B. Suppose $X = (Y, Z) \sim P_\theta$ with $Y = \theta' Z + \epsilon$ where $\theta \in R^d$ and $\epsilon \sim G$ with density g is independent of Z. This is a semiparametric model with parameters (θ, g).
Case C. Suppose $X = (Y, Z) \sim P$ where P is an arbitrary probability distribution on $R \times R^d$. This is a nonparametric model.

For a given data, there are many reasonable models which can be used to describe data. A good model is usually preferred if it is compatible with underlying mechanism of data generation, has as few model assumption as possible, can be presented in simple ways, and inference is feasible. In other words, a good model should make sense, be flexible and parsimonious, and be easy for inference.

4.2 Methods of Point Estimation: A Review

There have been a number of estimation methods proposed for many statistical models. However, some methods may work well from some statistical models but may not work well for others. In the following sections, we list a few of these methods, along with examples.

4.2.1 Least square estimation

The least square estimation is the most classical estimation method. This method estimates the parameters by minimizing the summed square distance between the observed quantities and the expected quantities.

Example 4.4 Suppose n i.i.d observations (Y_i, Z_i), $i = 1, ..., n$, are generated from the distribution in Example 4.3. To estimate θ, one method is to minimize the least square function

$$\sum_{i=1}^n (Y_i - \theta' Z_i)^2.$$

This gives the least square estimate for θ as

$$\hat{\theta} = (\sum_{i=1}^{n} Z_i Z_i')^{-1} (\sum_{i=1}^{n} Z_i Y_i).$$

It can show that $E[\hat{\theta}] = \theta$. Note that this estimation does not use any distribution function in ϵ so applies to all three cases.

4.2.2 Uniformly minimal variance and unbiased estimation

Sometimes, one seeks an estimate which is unbiased for parameters of interest. Furthermore, one wants such an estimate to have the least variation. If such an estimator exists, we call it the *uniformly minimal variance and unbiased estimator* (UMVUE) (an estimator T is unbiased for the parameter θ if $E[T] = \theta$). It should be noted that such an estimator may not exist.

The UMVUE often exists for distributions in the exponential family, whose probability density functions are of form

$$p_\theta(x) = h(x)c(\theta)\exp\{\eta_1(\theta)T_1(x) + ...\eta_s(\theta)T_s(x)\},$$

where $\theta \in R^d$ and $T(x) = (T_1(x), ..., T_s(x))$ is the s-dimensional statistics. The following lemma describes how one can find a UMVUE for θ from an unbiased estimator.

Definition 4.1 $T(X)$ is called a *sufficient statistic* for $X \sim p_\theta$ with respect to θ if the conditional distribution of X given $T(X)$ is independent of θ. $T(X)$ is a *complete statistic* with respect to θ if for any measurable function g, $E_\theta[g(T(X))] = 0$ for any θ implies $g = 0$, where E_θ denotes the expectation under the density function with parameter θ. †

It is easy to check that $T(X)$ is sufficient if and only if $p_\theta(x)$ can be factorized into $g_\theta(T(x))h(x)$. Thus, in the exponential family, $T(X) = (T_1(X), ..., T_s(X))$ is sufficient. Additionally, if the exponential family is of full-rank (i.e., $\{(\eta_1(\theta), ..., \eta_s(\theta)) : \theta \in \Theta\}$ contains a cube in s-dimensional space), $T(X)$ is also a complete statistic. The proof can be referred to Theorem 6.22 in Lehmann and Casella (1998).

Proposition 4.1 Suppose $\hat{\theta}(X)$ is an unbiased estimator for θ; i.e., $E[\hat{\theta}(X)] = \theta$. If $T(X)$ is a sufficient statistics of X, then $E[\hat{\theta}(X)|T(X)]$ is unbiased and moreover,

$$Var(E[\hat{\theta}(X)|T(X)]) \leq Var(\hat{\theta}(X)),$$

with the equality if and only if with probability 1, $\hat{\theta}(X) = E[\hat{\theta}(X)|T(X)]$. †

Proof $E[\hat{\theta}(X)|T]$ is clearly unbiased and moreover, by the Jensen's inequality,

$$Var(E[\hat{\theta}(X)|T]) = E[(E[\hat{\theta}(X)|T])^2] - E[\hat{\theta}(X)]^2 \leq E[\hat{\theta}(X)^2] - \theta^2 = Var(\hat{\theta}(X)).$$

The equality holds if and only if $E[\hat{\theta}(X)|T] = \hat{\theta}(X)$ with probability 1. †

Proposition 4.2 If $T(X)$ is complete sufficient and $\hat{\theta}(X)$ is unbiased, then $E[\hat{\theta}(X)|T(X)]$ is the unique UMVUE for θ. †

Proof For any unbiased estimator for θ, denoted by $\tilde{T}(X)$, we obtain from Proposition 4.1 that $E[\tilde{T}(X)|T(X)]$ is unbiased and

$$Var(E[\tilde{T}(X)|T(X)]) \leq Var(\tilde{T}(X)).$$

Since $E[E[\tilde{T}(X)|T(X)] - E[\hat{\theta}(X)|T(X)]] = 0$ and $E[\tilde{T}(X)|T(X)]$ and $E[\hat{\theta}(X)|T(X)]$ are independent of θ, the completeness of $T(X)$ gives that

$$E[\tilde{T}(X)|T(X)] = E[\hat{\theta}(X)|T(X)].$$

That is, $Var(E[\hat{\theta}(X)|T(X)]) \leq Var(\tilde{T}(X))$. Thus, $E[\hat{\theta}(X)|T(X)]$ is the UMVUE. The above arguments also show that such a UMVUE is unique. †

Proposition 4.2 suggests two ways to derive the UMVUE in the presence of a complete sufficient statistic $T(X)$: one way is to find an unbiased estimator of θ then calculate the conditional expectation of this unbiased estimator given $T(X)$; another way is to directly find a function $g(T(X))$ such that $E[g(T(X))] = \theta$. The following example describes these two methods.

Example 4.5 Suppose $X_1, ..., X_n$ are i.i.d according to the uniform distribution $U(0, \theta)$ and we wish to obtain a UMVUE of $\theta/2$. From the joint density of $X_1, ..., X_n$ given by

$$\frac{1}{\theta^n} I(X_{(n)} < \theta) I(X_{(1)} > 0),$$

one can easily show $X_{(n)}$ is complete and sufficient for θ. Note $E[X_1] = \theta/2$. Thus, a UMVUE for $\theta/2$ is given by

$$E[X_1|X_{(n)}] = \frac{n+1}{n} \frac{X_{(n)}}{2}.$$

The other way is to directly find a function $g(X_{(n)}) = \theta/2$ by noting

$$E[g(X_{(n)})] = \frac{1}{\theta^n} \int_0^\theta g(x) n x^{n-1} dx = \theta/2.$$

Thus, we have

$$\int_0^\theta g(x) x^{n-1} dx = \frac{\theta^{n+1}}{2n}.$$

We differentiate both sides with respect to θ and obtain $g(x) = \frac{n+1}{n} \frac{x}{2}$. Hence, we again obtain the UMVUE for $\theta/2$ is equal to $(n+1)X_{(n)}/2n$.

Many more examples of the UMVUE can be found in Chapter 2 of Lehmann and Casella (1998).

4.2.3 Robust estimation

In some regression problems, one may be concerned about outliers. For example, in a simple linear regression, an extreme outlier may affect the fitted line greatly. One estimation approach called robust estimation approach is to propose an estimator which is little influenced by extreme observations. Often, for n i.i.d observations $X_1, ..., X_n$, the robust estimation approach is to minimize an objective function with the form $\sum_{i=1}^n \phi(X_i; \theta)$.

Example 4.6 In linear regression, a model for (Y, X) is given by

$$Y = \theta' X + \epsilon,$$

where ϵ has mean zero. One robust estimator is to minimize

$$\sum_{i=1}^n |Y_i - \theta' X_i|$$

and the obtained estimator is called the least absolute deviation estimator. A more general objective function is to minimize

$$\sum_{i=1}^n \phi(Y_i - \theta' X_i),$$

where $\phi(x) = |x|^k, |x| \leq C$ and $\phi(x) = C^k$ when $|x| > C$.

4.2.4 Estimating functions

In recent statistical inference, more and more estimators are based on estimating functions. The use of estimating functions has been extensively seen in semiparametric model. An estimating function for θ is a measurable function $f(X; \theta)$ with $E[f(X; \theta)] = 0$ or approximating zero. Then an estimator for θ using n i.i.d observations can be constructed by solving the estimating equation

$$\sum_{i=1}^n f(X_i; \theta) = 0.$$

The estimating function is useful, especially when there are other parameters in the model but only θ is parameters of interest.

Example 4.7 We still consider the linear regression example. We can see that for any function $W(X)$, $E[XW(X)(Y - \theta' X)] = 0$. Thus an estimating equation for θ can be constructed as

$$\sum_{i=1}^n X_i W(X_i)(Y_i - \theta' X_i) = 0.$$

Example 4.8 Still in the regression example but we now assume the median of ϵ is zero. It is easy to see that $E[XW(X)sgn(Y - \theta' X)] = 0$. Then an estimating equation for θ can be constructed as

$$\sum_{i=1}^n X_i W(X_i) sgn(Y_i - \theta' X_i) = 0.$$

4.2.5 Maximum likelihood estimation

The most commonly used method, at least in parametric models, is the maximum likelihood estimation method: If n i.i.d observations $X_1, ..., X_n$ are generated from a distribution function with densities $p_\theta(x)$, then it is reasonable to believe that the best value for θ should be the one maximizing the observed likelihood function, which is defined as

$$L_n(\theta) = \prod_{i=1}^n p_\theta(X_i).$$

The obtained estimator $\hat{\theta}$ is called the maximum likelihood estimator for θ. Many nice properties are possessed by the maximum likelihood estimators and we will particularly investigate this issue in next chapter. Recent development has also seen the implementation of the maximum likelihood estimation in semiparametric models and nonparametric models.

Example 4.9 Suppose $X_1, ..., X_n$ are i.i.d. observations from $\exp(\theta)$. Then the likelihood function for θ is equal to

$$L_n(\theta) = \theta^n \exp\{-\theta(X_1 + ... + X_n)\}.$$

The maximum likelihood estimator for θ is given by $\hat{\theta} = \bar{X}_n$.

Example 4.10 The setting is Case B of Example 1.2. Suppose $(Y_1, Z_1), ..., (Y_n, Z_n)$ are i.i.d with the density function $\lambda(y) e^{\theta' z} \exp\{-\Lambda(y) e^{\theta' z}\} g(z)$, where $g(z)$ is the known density function of $Z = z$. Then the likelihood function for the parameters (θ, λ) is given by

$$L_n(\theta, \lambda) = \prod_{i=1}^n \left\{ \lambda(Y_i) e^{\theta' Z_i} \exp\{-\Lambda(Y_i) e^{\theta' Z_i}\} g(Z_i) \right\}.$$

It turns out that the maximum likelihood estimators for (θ, λ) do not exist. One way is to let Λ be a step function with jumps at $Y_1, ..., Y_n$ and let $\lambda(Y_i)$ be the jump size, denoted as p_i. Then the likelihood function becomes

$$L_n(\theta, p_1, ..., p_n) = \prod_{i=1}^n \left\{ p_i e^{\theta' Z_i} \exp\{-\sum_{Y_j \leq Y_i} p_j e^{\theta' Z_i}\} g(Z_i) \right\}.$$

The maximum likelihood estimators for $(\theta, p_1, ..., p_n)$ are given as: $\hat{\theta}$ solves the equation

$$\sum_{i=1}^n Z_i \left[1 - \frac{\sum_{Y_j \geq Y_i} Z_j e^{\theta' Z_j}}{\sum_{Y_j \geq Y_i} e^{\theta' Z_j}} \right] = 0$$

and

$$p_i = \frac{1}{\sum_{Y_j \geq Y_i} e^{\theta' Z_j}}.$$

4.2.6 Bayesian estimation

In this estimation approach, the parameter θ in the model distributions $\{p_\theta(x)\}$ is treated as a random variable with some prior distribution $\pi(\theta)$. The estimator for θ is defined as a value depending on the data and minimizing the expected loss function or the maximal loss function, where the loss function is denoted as $l(\theta, \hat{\theta}(X))$. The usual loss function includes the quadratic loss $(\theta - \hat{\theta}(X))^2$, the absolute loss $|\theta - \hat{\theta}(X)|$ etc. It often turns out that $\hat{\theta}(X)$ can be determined from the posterior distribution of $P(\theta|X) = P(X|\theta)P(\theta)/P(X)$.

Example 4.11 Suppose $X \sim N(\theta, 1)$, where θ has an improper prior distribution of being uniform in $(-\infty, \infty)$. It is clear that the estimator $\hat{\theta}(X)$, minimizing the quadratic loss $E[(\theta - \hat{\theta}(X))^2]$, is the posterior mean $E[\theta|X] = X$.

4.2.7 Concluding remarks

We have reviewed a few methods which are seen in many statistical problems. However we have not exhausted all estimation approaches. Other estimation methods include the conditional likelihood estimation, the profile likelihood estimation, the partial likelihood estimation, the empirical Bayesian estimation, the minimax estimation, the rank estimation, L-estimation and etc.

With a number of estimators, one natural question is to decide which estimator is the best choice. The first criteria is that the estimator must be unbiased or at least consistent with the true parameter. Such a property is called the first order efficiency. In order to make a precise estimation, we may also want the estimator to have as small variance as possible. The issue then becomes the second order efficiency, which we will discuss in the next section.

4.3 Cram$\acute{e}$r-Rao Bounds for Parametric Models

4.3.1 Information bound in one-dimensional model

First, we assume the model is one-dimensional parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R$. We assume:
A. $X \sim P_\theta$ on $(\Omega, \mathcal{A})$ with $\theta \in \Theta$.
B. $p_\theta = dP_\theta/d\mu$ exists where μ is a σ-finite dominating measure.
C. $T(X) \equiv T$ estimates $q(\theta)$ has $E_\theta[|T(X)|] < \infty$; set $b(\theta) = E_\theta[T] - q(\theta)$.
D. $q'(\theta) \equiv \dot{q}(\theta)$ exists.

Theorem 4.1 (Information bound or Cram$\acute{e}$r-Rao Inequality) Suppose:
(C1) Θ is an open subset of the real line.
(C2) There exists a set B with $\mu(B) = 0$ such that for $x \in B^c$, $\partial p_\theta(x)/\partial\theta$ exists for all θ. Moreover, $A = \{x : p_\theta(x) = 0\}$ does not depend on θ.
(C3) $I(\theta) = E_\theta[\dot{l}_\theta(X)^2] > 0$ where $\dot{l}_\theta(x) = \partial \log p_\theta(x)/\partial\theta$. Here, $I(\theta)$ is the called the *Fisher information* for θ and $\dot{l}_\theta$ is called the *score function* for θ.
(C4) $\int p_\theta(x)d\mu(x)$ and $\int T(x)p_\theta(x)d\mu(x)$ can both be differentiated with respect to θ under the integral sign.

(C5) $\int p_\theta(x)d\mu(x)$ can be differentiated twice under the integral sign.
If (C1)-(C4) hold, then

$$Var_\theta(T(X)) \geq \frac{\{\dot{q}(\theta)+\dot{b}(\theta)\}^2}{I(\theta)},$$

and the lower bound is equal to $\dot{q}(\theta)^2/I(\theta)$ if T is unbiased. Equality holds for all θ if and only if for some function $A(\theta)$, we have

$$\dot{l}_\theta(x) = A(\theta)\{T(x) - E_\theta[T(X)]\}, \quad a.e.\mu.$$

If, in addition, (C5) holds, then

$$I(\theta) = -E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right\} = -E_\theta[\ddot{l}_\theta(X)].$$

†

Proof Note

$$q(\theta)+b(\theta) = \int T(x)p_\theta(x)d\mu(x) = \int_{A^c\cap B^c} T(x)p_\theta(x)d\mu(x).$$

Thus from (C2) can (C4),

$$\dot{q}(\theta)+\dot{b}(\theta) = \int_{A^c\cap B^c} T(x)\dot{l}_\theta(x)p_\theta(x)d\mu(x) = E_\theta[T(X)\dot{l}_\theta(X)].$$

On the other hand, since $\int_{A^c\cap B^c} p_\theta(x)d\mu(x) = 1$,

$$0 = \int_{A^c\cap B^c} \dot{l}_\theta(x)p_\theta(x)d\mu(x) = E_\theta[\dot{l}_\theta(X)].$$

Then

$$\dot{q}(\theta)+\dot{b}(\theta) = Cov(T(X), \dot{l}_\theta(X)).$$

By the Cauchy-Schwartz inequality, we obtain

$$|\dot{q}(\theta)+\dot{b}(\theta)| \leq Var(T(X))Var(\dot{l}_\theta(X)).$$

The equality holds if and only if

$$\dot{l}_\theta(X) = A(\theta)\left\{T(X) - E_\theta[T(X)]\right\}, a.s.$$

Finally, if (C5) holds, we further differentiate

$$0 = \int \dot{l}_\theta(x)p_\theta(x)d\mu(x)$$

and obtain

$$0 = \int \ddot{l}_\theta(x)p_\theta(x)d\mu(x) + \int \dot{l}_\theta(x)^2 p_\theta(x)d\mu(x).$$

Thus, we obtain the equality $I(\theta) = -E_\theta[\ddot{l}_\theta(X)]$. †

Theorem 4.1 implies that the variance of any unbiased estimator has a lower bound $\dot{q}(\theta)^2/I(\theta)$, which is intrinsic to the parametric model. Especially, if $q(\theta) = \theta$, then the lower bound for the variance of unbiased estimator for θ is the inverse of the information. The following examples calculate this bound for some parametric models.

Example 4.12 Suppose $X_1, ..., X_n$ are i.i.d $Poisson(\theta)$. The density function for $(X_1, ..., X_n)$ is given by

$$p_\theta(X_1, ..., X_n) = -n\theta + n\bar{X}_n \log\theta - \sum_{i=1}^n \log(X_i!).$$

Thus,

$$l_\theta(X_1, ..., X_n) = \frac{n}{\theta}(\bar{X}_n - \theta).$$

It is direct to check all the regularity conditions of Theorem 3.1 are satisfied. Then $I_n(\theta) = n^2/\theta^2 Var(\bar{X}_n) = n/\theta$. The Carmér-Rao bound for θ is equal to θ/n. On the other hand, we note $\bar{X}_n$ is an unbiased estimator of θ. Moreover, since $\bar{X}_n$ is the complete statistic for θ. $\bar{X}_n$ is indeed the UMVUE of θ. Note $Var(\bar{X}_n) = \theta/n$. We conclude that $\bar{X}_n$ attains the lower bound. However, although $T_n = \bar{X}_n^2 - n^{-1}\bar{X}_n$ is unbiased for θ^2 and it is UMVUE of θ^2, we find $Var(T_n) = 4\theta^3/n + 2\theta^2/n^2 >$ the Cramér-Rao lower bound for θ^2. In other words, some UMVUE attains the lower bound but some do not.

Example 4.13 Suppose $X_1, ..., X_n$ are i.i.d with density $p_\theta(x) = g(x - \theta)$ where g is known density. This family is the one-dimensional location model. Assume g' exists and the regularity conditions in Theorem 4.1 are satisfied. Then

$$I_n(\theta) = nE_\theta[\frac{g'(X-\theta)}{g(X-\theta)}^2] = n\int \frac{g'(x)^2}{g(x)}dx.$$

Note the information does not depend on θ.

Example 4.14 Suppose $X_1, ..., X_n$ are i.i.d with density $p_\theta(x) = g(x/\theta)/\theta$ where g is a known density function. This model is one-dimensional scale model with the common shape g. It is direct to calculate

$$I_n(\theta) = \frac{n}{\theta^2}\int (1 + y\frac{g'(y)}{g(y)})^2 g(y)dy.$$

4.3.2 Information bound in multi-dimensional model

We can extend Theorem 4.1 to the case in which the model is k-dimensional parametric family: $\mathcal{P} = \{P_\theta : \theta \in \Theta \subset R^k\}$. Similar to Assumptions A-C, we assume P_θ has density function p_θ with respect to some σ-finite dominating measure μ; $T(X)$ is an estimator for $q(\theta)$ with $E_\theta[|T(X)|] < \infty$ and $b(\theta) = E_\theta[T(X)] - q(\theta$ is the bias of $T(X)$; $\dot{q}(\theta) = \nabla q(\theta)$ exists.

Theorem 4.2 (Information inequality) Suppose that
(M1) Θ an open subset in R^k.
(M2) There exists a set B with $\mu(B) = 0$ such that for $x \in B^c$, $\partial p_\theta(x)/\partial\theta_i$ exists for all θ and

$i = 1, ..., k$. The set $A = \{x : p_\theta(x) = 0\}$ does no depend on θ.
(M3) The $k \times k$ matrix $I(\theta) = (I_{ij}(\theta)) = E_\theta[\dot{l}_\theta(X)\dot{l}_\theta(X)'] > 0$ is a positive definite where

$$\dot{l}_{\theta_i}(x) = \frac{\partial}{\partial\theta_i}\log p_\theta(x).$$

Here, $I(\theta)$ is called the Fisher information matrix for θ and $\dot{l}_\theta$ is called the score for θ.
(M4) $\int p_\theta(x)d\mu(x)$ and $\int T(x)p_\theta(x)d\mu(x)$ can both be differentiated with respect to θ under the integral sign.
(M5) $\int p_\theta(x)d\mu(x)$ can be differentiated twice with respect to θ under the integral sign.
If (M1)-(M4) holds, than

$$Var_\theta(T(X)) \geq (\dot{q}(\theta) + \dot{b}(\theta))'I^{-1}(\theta)(\dot{q}(\theta) + \dot{b}(\theta))$$

and this lower bound is equal $\dot{q}(\theta)'I(\theta)^{-1}\dot{q}(\theta)$ if $T(X)$ is unbiased. If, in addition, (M5) holds, then

$$I(\theta) = -E_\theta[\ddot{l}_{\theta\theta}(X)] = -\Big(E_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right\}\Big).$$

†

Proof Under (M1)-(M4), we have

$$\dot{q}(\theta) + \dot{b}(\theta) = \int T(x)\dot{l}_\theta(x)p_\theta(x)d\mu(x) = E_\theta[T(x)\dot{l}_\theta(X)].$$

On the other hand, from $\int p_\theta(x)d\mu(x) = 1$, $0 = E_\theta[\dot{l}_\theta(X)]$. Thus,

$$\begin{aligned}
& |\left\{\dot{q}(\theta) + \dot{b}(\theta)\right\}' I(\theta)^{-1}\left\{\dot{q}(\theta) + \dot{b}(\theta)\right\}| \\
= \;& |E_\theta[T(X)(\dot{q}(\theta) + \dot{b}(\theta))'I(\theta)^{-1}\dot{l}_\theta(X)]| \\
= \;& |Cov_\theta(T(X), (\dot{q}(\theta) + \dot{b}(\theta))'I(\theta)^{-1}\dot{l}_\theta(X))| \\
\leq \;& \sqrt{Var_\theta(T(X))(\dot{q}(\theta) + \dot{b}(\theta))'I(\theta)^{-1}(\dot{q}(\theta) + \dot{b}(\theta))}.
\end{aligned}$$

We obtain the information inequality. In addition, if (M5) holds, we further differentiate $\int \dot{l}_\theta(x)p_\theta(x)d\mu(x) = 0$ and obtain the then

$$I(\theta) = -E_\theta[\ddot{l}_{\theta\theta}(X)] = -\Big(E_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right\}\Big).$$

†

Example 4.15 The Weibull family $\mathcal{P}$ is the parametric model with densities

$$p_\theta(x) = \frac{\beta}{\alpha}(\frac{x}{\alpha})^{\beta-1}\exp\left\{-(\frac{x}{\alpha})^\beta\right\}I(x \geq 0)$$

with respect to the Lebesgue measure where $\theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)$. We can easily calculate that

$$\dot{l}_\alpha(x) = \frac{\beta}{\alpha}\left\{(\frac{x}{\alpha})^\beta - 1\right\},$$

$$\dot{l}_\beta(x) = \frac{1}{\beta} - \frac{1}{\beta}\log\left\{(\frac{x}{\alpha})^\beta\right\}\left\{(\frac{x}{\alpha})^\beta - 1\right\}.$$

Thus, the Fisher information matrix is

$$I(\theta) = \begin{pmatrix} \beta^2/\alpha^2 & -(1-\gamma)/\alpha \\ -(1-\gamma)/\alpha & \{\pi^2/6 + (1-\gamma)^2\}/\beta^2 \end{pmatrix},$$

where γ is the Euler's constant ($\gamma \approx 0.5777...$). The computation of $I(\theta)$ is simplified by noting that $Y \equiv (X/\alpha)^\beta \sim \text{Exponential}(x)$.

4.3.3 Efficient influence function and efficient score function

From the above proof, we also note that the lower bound is attained for an unbiased estimator $T(X)$ if and only if $T(X) = \dot{q}(\theta)'I^{-1}(\theta)\dot{l}_\theta(X)$, the latter is called the *efficient influence function* for estimating $q(\theta)$ and its variance, which is equal to $\dot{q}(\theta)'I(\theta)^{-1}\dot{q}(\theta)$, is called the *information bound* for $q(\theta)$. If we regard $q(\theta)$ as a function on all the distributions of $\mathcal{P}$ and denote $\nu(P_\theta) = q(\theta)$, then in some literature, the efficient influence function and the information bound for $q(\theta)$ can be represented as $\tilde{l}(X, P_\theta|\nu, \mathcal{P})$ and $I^{-1}(P_\theta|\nu, \mathcal{P})$, both implying that the efficient influence function and the information matrix are meant for a fixed model $\mathcal{P}$, for a parameter of interest $\nu(P_\theta) = q(\theta)$, and at a fixed distribution P_θ.

Proposition 4.3 The information bound $I^{-1}(P|\nu, \mathcal{P})$ and the efficient influence function $\tilde{l}(\cdot, P|\nu, \mathcal{P})$ are invariant under smooth changes of parameterization. †

Proof Suppose $\gamma \mapsto \theta(\gamma)$ is a one-to-one continuously differentiable mapping of an open subset Γ of R^k onto Θ with nonsingular differential $\dot{\theta}$. The model of distribution can be represented as $\{P_{\theta(\gamma)} : \gamma \in \Gamma\}$. Thus, the score for γ is $\dot{\theta}(\gamma)\dot{l}_\theta(X)$ so the information matrix for γ is equal to

$$\dot{\theta}(\gamma)'I(\theta(\gamma))\dot{\theta}(\gamma),$$

which is the same as the information matrix for $\theta = \theta(\gamma)$. The efficient influence function for γ is equal to

$$(\dot{\theta}(\gamma)\dot{q}(\theta(\gamma)))'I(\gamma)^{-1}\dot{l}_\gamma = \dot{q}(\theta(\gamma))'I(\theta(\gamma))^{-1}\dot{l}_\theta$$

and it is the same as the efficient influence function for θ. †

The proposition implies that the information bound and the efficient influence function for some ν in a family of distribution are independent of the parameterization used in the model. However, with some natural and simple parameterization, the calculation of the information bound and the efficient influence function can be directly done along the definition. Especially, we look into a specific parameterization where $\theta' = (\nu', \eta')$ and $\nu \in \mathcal{N} \subset R^m$, $\eta \in \mathcal{H} \subset R^{k-m}$. ν can be regarded as a map mapping P_θ to one of component of θ, ν, and it is the parameter of interest while η is a nuisance parameter. We want to assess the cost of not knowing η by comparing the information bounds and the efficient influence functions for ν in the model $\mathcal{P}$ (η is unknown parameter) and $\mathcal{P}_\eta$ (η is known and fixed).

In the model $\mathcal{P}$, we can decompose

$$\dot{l}_\theta = \begin{pmatrix} \dot{l}_1 \\ \dot{l}_2 \end{pmatrix}, \quad \tilde{l}_\theta = \begin{pmatrix} \tilde{l}_1 \\ \tilde{l}_2 \end{pmatrix},$$

where $\dot{l}_1$ is the score for ν and $\dot{l}_2$ is the score for η, $\tilde{l}_1$ is the efficient influence function for ν and $\tilde{l}_2$ is the efficient influence function for η. Correspondingly, we can decompose the information matrix $I(\theta)$ into

$$I(\theta) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where $I_{11} = E_\theta[\dot{l}_1\dot{l}_1']$, $I_{12} = E_\theta[\dot{l}_1\dot{l}_2']$, $I_{21} = E_\theta[\dot{l}_2\dot{l}_1']$, and $I_{22} = E_\theta[\dot{l}_2\dot{l}_2']$. Thus,

$$I^{-1}(\theta) = \begin{pmatrix} I_{11\cdot 2}^{-1} & -I_{11\cdot 2}^{-1}I_{12}I_{22}^{-1} \\ -I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1} & I_{22\cdot 1}^{-1} \end{pmatrix} \equiv \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix},$$

where

$$I_{11\cdot 2} = I_{11} - I_{12}I_{22}^{-1}I_{21}, \quad I_{22\cdot 1} = I_{22} - I_{21}I_{11}^{-1}I_{12}.$$

Since the information bound for estimating ν is equal to

$$I^{-1}(P_\theta|\nu, \mathcal{P}) = \dot{q}(\theta)'I^{-1}(\theta)\dot{q}(\theta),$$

where $q(\theta) = \nu$, and

$$\dot{q}(\theta) = \begin{pmatrix} I_{m\times m} & 0_{m\times(k-m)} \end{pmatrix},$$

we obtain the information bound for ν is given by

$$I^{-1}(P_\theta|\nu, \mathcal{P}) = I_{11\cdot 2}^{-1} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}.$$

The efficient influence function for ν is given by

$$\tilde{l}_1 = \dot{q}(\theta)'I^{-1}(\theta)\dot{l}_\theta = I_{11\cdot 2}^{-1}\dot{l}_1^*,$$

where $\dot{l}_1^* = \dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2$. It is easy to check

$$I_{11\cdot 2} = E[\dot{l}_1^*(\dot{l}_1^*)'].$$

Thus, l_1^* is called the *efficient score function* for ν in $\mathcal{P}$.

Now we consider the model $\mathcal{P}_\eta$ with η known and fixed. It is clear the information bound for ν is just I_{11}^{-1} and the efficient influence function for ν is equal to $I_{11}^{-1}\dot{l}_1$.

Since $I_{11} > I_{11\cdot 2} = I_{11} - I_{12}I_{22}^{-1}I_{21}$, we conclude that knowing η increases the Fisher information for ν and decreases the information bound for ν. Moreover, knowledge of η does not increase information about ν if and only if $I_{12} = 0$. In this case, $\tilde{l}_1 = I_{11}^{-1}\dot{l}_1$ and $l_1^* = l_1$.

Example 4.16 Suppose

$$\mathcal{P} = \left\{P_\theta : p_\theta = \phi((x-\nu)/\eta)/\eta, \nu \in R, \eta > 0\right\}.$$

Note that
$$\dot{l}_\nu(x) = \frac{x-\nu}{\eta^2}, \quad \dot{l}_\eta(x) = \frac{1}{\eta}\left\{\frac{(x-\nu)^2}{\eta^2} - 1\right\}.$$
Then the information matrix $I(\theta)$ is given by by
$$I(\theta) = \begin{pmatrix} \eta^{-2} & 0 \\ 0 & 2\eta^{-2} \end{pmatrix}.$$

Then we can estimate the ν equally well whether we know the variance or not.

Example 4.17 If we reparameterize the above model as
$$P_\theta = N(\nu, \eta^2 - \nu^2), \quad \eta^2 > \nu^2.$$

The easy calculation shows that $I_{12}(\theta) = \nu\eta/(\eta^2-\nu^2)^2$. Thus lack of knowledge of η in this parameterization does change the information bound for estimation of ν.

We provide a nice geometric way of calculating the efficient score function and the efficient influence function for ν. For any θ, the linear space $L_2(P_\theta) = \{g(X) : E_\theta[g(X)^2] < \infty\}$ is a Hilbert space with the inner product defined as
$$< g_1, g_2 >= E[g_1(X)g_2(X)].$$

On this Hilbert space, we can define the concept of the projection. For any closed linear space $\mathcal{S} \subset L_2(P_\theta)$ and any $g \in L_2(P_\theta)$, the projection of g on $\mathcal{S}$ is $\tilde{g} \in \mathcal{S}$ such that $g - \tilde{g}$ is orthogonal to any g^* in $\mathcal{S}$ in the sense that
$$E[(g(X) - \tilde{g}(X))g^*(X)] = 0, \quad \forall g^* \in \mathcal{S}.$$

The *orthocomplement* of $\mathcal{S}$ is a linear space with all the $g \in L_2(P)$ such that g is orthogonal to any $g^* \in \mathcal{S}$. The above concepts agree with the usual definition in the Euclidean space. The following theorem describes the calculation of the efficient score function and the efficient influence function.

Theorem 4.3 A. The efficient score function $\dot{l}_1^*(\cdot, P_\theta|\nu, \mathcal{P})$ is the projection of the score function $\dot{l}_1$ on the orthocomplement of $[\dot{l}_2]$ in $L_2(P_\theta)$, where $[\dot{l}_2]$ is the linear span of the components of $\dot{l}_2$.
B. The efficient influence function $\tilde{l}(\cdot, P_\theta|\nu, \mathcal{P}_\eta)$ is the projection of the efficient influence function $\tilde{l}_1$ on $[\dot{l}_1]$ in $L_2(P_\theta)$. †

Proof A. Suppose the projection of $\dot{l}_1$ on $[\dot{l}_2]$ is equal to $\Sigma\dot{l}_2$ for some matrix Σ. Since $E[(\dot{l}_1 - \Sigma\dot{l}_2)\dot{l}_2'] = 0$, we obtain $\Sigma = I_{12}I_{22}^{-1}$ then the projection on the orthocomplement of $[\dot{l}_2]$ is equal to $\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2$, which is the same as $\dot{l}_1^*$.
B. After the algebra, we note
$$\tilde{l}_1 = I_{11\cdot 2}^{-1}(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2) = (I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1})(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2) = I_{11}^{-1}\dot{l}_1 - I_{11}^{-1}I_{12}\tilde{l}_2.$$

Since from A, $\tilde{l}_2$ is orthogonal to $\dot{l}_1$, the projection of $\tilde{l}_1$ on $[\dot{l}_1]$ is equal $I_{11}^{-1}\dot{l}_1$, which is the efficient influence function $\tilde{l}(\cdot, P_\theta|\nu, \mathcal{P}_\eta)$. †

The following table describes the relationship among all these terminologies.

Term	**Notation**	$\mathcal{P}$ (η unknown)	$\mathcal{P}_\eta$ (η known)
efficient score	$\dot{l}_1^*(, P\|\nu, \cdot)$	$\dot{l}_1^* = \dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2$	$\dot{l}_1$
information	$I(P\|\nu, \cdot)$	$E[\dot{l}_1^*(\dot{l}_1^*)'] = I_{11} - I_{12}I_{22}^{-1}I_22$	I_{11}
efficient influence information	$\tilde{l}_1(\cdot, P\|\nu, \cdot)$	$\tilde{l}_1 = I^{11}\dot{l}_1 + I^{12}\dot{l}_2 = I_{11\cdot 2}^{-1}\dot{l}_1^*$ $= I_{11}^{-1}\dot{l}_1 - I_{11}^{-1}I_{12}\tilde{l}_2$	$I_{11}^{-1}\dot{l}_1$
information bound	$I^{-1}(P\|\nu, \cdot)$	$I^{11} = I_{11\cdot 2}^{-1} = I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1}$	I_{11}^{-1}

4.4 Asymptotic Efficiency Bound

4.4.1 Regularity conditions and asymptotic efficiency theorems

The Cramér-Rao bound can be considered as the lower bound for any unbiased estimator in finite sample. One may ask whether such a bound still holds in large sample. To be specific, we suppose $X_1, ..., X_n$ are i.i.d P_θ ($\theta \in R$) and an estimator T_n for θ satisfies that

$$\sqrt{n}(T_n - \theta) \to_d N(0, V(\theta)^2).$$

Then the question is whether $V(\theta)^2 \geq 1/I(\theta)$. Unfortunately, this may not be true as the following example due to Hodges gives one counterexample.

Example 4.18 Let $X_1, ..., X_n$ be i.i.d $N(\theta, 1)$ so that $I(\theta) = 1$. Let $|a| < 1$ and define

$$T_n = \begin{cases} \bar{X}_n & \text{if} |\bar{X}_n| > n^{-1/4} \\ a\bar{X}_n & \text{if} |\bar{X}_n| \leq n^{-1/4}. \end{cases}$$

Then

$$\begin{aligned} \sqrt{n}(T_n - \theta) &= \sqrt{n}(\bar{X}_n - \theta)I(|\bar{X}_n| > n^{-1/4}) + \sqrt{n}(a\bar{X}_n - \theta)I(|\bar{X}_n| \leq n^{-1/4}) \\ &=_d ZI(|Z + \sqrt{n}\theta| > n^{1/4}) + \left\{aZ + \sqrt{n}(a-1)\theta\right\} I(|Z + \sqrt{n}\theta| \leq n^{1/4}) \\ &\to_{a.s.} ZI(\theta \neq 0) + aZI(\theta = 0). \end{aligned}$$

Thus, the asymptotic variance of $\sqrt{n}T_n$ is equal 1 for $\theta \neq 0$ and a^2 for $\theta = 0$. The latter is smaller than the Cramér-Rao bound. In other words, T_n is a superefficient estimator.

To avoid the Hodge's superefficient estimator, we need impose some conditions to T_n in addition to the weak convergence of $\sqrt{n}(T_n - \theta)$. One such condition is called locally regular condition in the following sense.

Definition 4.2 $\{T_n\}$ is a *locally regular estimator* of θ at $\theta = \theta_0$ if, for every sequence $\{\theta_n\} \subset \Theta$ with $\sqrt{n}(\theta_n - \theta) \to t \in R^k$, under P_{θ_n},

$$\text{(local regularity)} \quad \sqrt{n}(T_n - \theta_n) \to_d Z, \quad \text{as} \quad n \to \infty$$

where the distribution of Z depend on θ_0 but not on t. Thus the limit distribution of $\sqrt{n}(T_n - \theta_n)$ does not depend on the direction of approach t of θ_n to θ_0. $\{T_n\}$ is a locally Gaussian regular if Z has normal distribution. †

In the above definition, $\sqrt{n}(T_n - \theta_n) \to_d Z$ under P_{θ_n} is equivalent to saying that for any bounded and continuous function g, $E_{\theta_n}[g(\sqrt{n}(T_n - \theta_n))] \to E[g(Z)]$. One can consider a locally regular estimator as the one whose limit distribution is locally stable: if data are generated under a model not far from a given model, the limit distribution of centralized estimator remains the same.

Furthermore, the locally regular condition, combining with the following two additional conditions, gives the results that the Cramér-Rao bound is also the asymptotic lower bound:

(C1) (*Hellinger differentiability*) A model $\mathcal{P} = \{P_\theta : \theta \in R^k\}$ is a parametric model dominated by a σ-finite measure μ. It is called a Hellinger-differentiable parametric model if

$$\|\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}h'\dot{l}_\theta\sqrt{p_\theta}\|_{L_2(\mu)} = o(|h|),$$

where $p_\theta = dP_\theta/d\mu$.

(C2) (*Local Asymptotic Normality* (LAN)) In a model $\mathcal{P} = \{P_\theta : \theta \in R^k\}$ dominated by a σ-finite measure μ, suppose $p_\theta = dP_\theta/d\mu$. Let $l(x;\theta) = \log p(x,\theta)$ and let

$$l_n(\theta) = \sum_{i=1}^n l(X_i;\theta)$$

be the log-likelihood function of $X_1, ..., X_n$. The local asymptotic normality condition at θ_0 is

$$l_n(\theta_0 + n^{-1/2}t) - l_n(\theta_0) \to_d N(-\frac{1}{2}t'I(\theta_0)t, t'I(\theta_0)t)$$

under P_{θ_0}.

Both conditions (C1) and (C2) are the smooth conditions imposed on the parametric models. In other words, we do not allow a model whose parameterization is irregular. An irregular model is seldom encountered in practical use.

The following theorem gives the main results.

Theorem 4.4 (Hájek's convolution theorem) Under conditions (C1)-(C2) with $I(\theta_0)$ non-singular. For any locally regular estimator of θ, $\{T_n\}$, the limit distribution of $\sqrt{n}(T_n - \theta_0)$ under P_{θ_0} satisfies

$$Z =^d Z_0 + \Delta_0,$$

where $Z_0 \sim N(0, I^{-1}(\theta_0))$ is independent of Δ_0. †

As a corollary, if $V(\theta_0)^2$ is the asymptotic variance of $\sqrt{n}(T_n - \theta_0)$, then $V(\theta_0)^2 \geq I^{-1}(\theta_0)$. Thus, the Cramér-Rao bound is a lower bound for the asymptotic variances of any locally regular estimators. Furthermore, we obtain the following corollary from Theorem 4.4.

Corollary 4.1 Suppose that $\{T_n\}$ is a locally regular estimator of θ at θ_0 and that $U : R^k \to R^+$ is bowl-shaped loss function; i.e., $U(x) = U(-x)$ and $\{x : U(x) \leq c\}$ is convex for any $c \geq 0$. Then

$$\liminf_n E_{\theta_0}[U(\sqrt{n}(T_n - \theta_0))] \geq E[U(Z_0)],$$

where $Z_0 \sim N(0, I(\theta_0)^{-1})$. †

Corollary 4.2 (H*á*jek-Le Cam asymptotic minmax theorem) Suppose that (C2) holds, that T_n is any estimator of θ, and U is bowl-shaped. Than

$$\lim_{\delta\to 0}\liminf_n \sup_{\theta:\sqrt{n}|\theta-\theta_0|\le\delta} E_\theta[U(\sqrt{n}(T_n-\theta))] \ge E[U(Z_0)],$$

where $Z_0 \sim N(0, I(\theta_0)^{-1})$. †

In summary, the two corollaries conclude that the asymptotic loss of any regular estimators is at least the loss given by the distribution Z_0. Thus, from this point of view, Z_0 is also the distribution of most efficiency. The proofs of the two corollaries are beyond this book so are skipped.

4.4.2 Le Cam's lemmas

Before proving Theorem 4.4, we introduce the contiguity definition and the Le Cam's lemmas. Consider a sequence of measure spaces $(\Omega_n, \mathcal{A}_n, \mu_n)$ and on each measure space, we have two probability measure P_n and Q_n with $P_n \prec\prec \mu_n$ and $Q_n \prec\prec \mu_n$. Let $p_n = dP_n/d\mu_n$ and $q_n = dQ_n/d\mu_n$ be the corresponding densities of P_n and Q_n. We define the likelihood ratios

$$L_n = \begin{cases} q_n/p_n & \text{if } p_n > 0 \\ 1 & \text{if } q_n = p_n = 0 \\ n & \text{if } q_n > 0 = p_n. \end{cases}$$

Definition 4.3 (*Contiguity*) The sequence $\{Q_n\}$ is contiguous to $\{P_n\}$ if for every sequence $B_n \in \mathcal{A}_n$ for which $P_n(B_n) \to 0$ it follows that $Q_n(B_n) \to 0$. †

Thus contiguity of $\{Q_n\}$ to $\{P_n\}$ means that Q_n is "asymptotically absolutely continuous" with respect to P_n. We denote $\{Q_n\} \lhd \{P_n\}$. Two sequences are contiguous to each other if $\{Q_n\} \lhd \{P_n\}$ and $\{P_n\} \lhd \{Q_n\}$ and we write $\{P_n\} \lhd\rhd \{Q_n\}$.

Definition 4.4 (*Asymptotic orthogonality*) The sequence $\{Q_n\}$ is asymptotically orthogonal to $\{P_n\}$ if there exists a sequence $B_n \in \mathcal{A}_n$ such that $Q_n(B_n) \to 1$ and $P_n(B_n) \to 0$. †

Proposition 4.4 (Le Cam's first lemma) Suppose under P_n, $L_n \to_d L$ with $E[L] = 1$. Then $\{Q_n\} \lhd \{P_n\}$. On the contrary, if $\{Q_n\} \lhd \{P_n\}$ and under P_n, $L_n \to_d L$, then $E[L] = 1$. †

Proof We fist prove the first half of the lemma. Let $B_n \in \mathcal{A}_n$ with $P_n(B_n) \to 0$. Then $I_{\Omega_n - B_n}$ converges to 1 in probability under P_n. Since L_n is asymptotically tight, $(L_n, I_{\Omega_n-B_n})$ is asymptotically tight under P_n. Thus, by the Helly's lemma, for every subsequence of $\{n\}$, there exists a further subsequence such that $(L_n, I_{\Omega_n-B_n}) \to_d (L, 1)$. By the Protmanteau Lemma, since $(v, t) \mapsto vt$ is continuous and nonnegative,

$$\liminf_n Q_n(\Omega_n - B_n) \ge \liminf_n \int I_{\Omega_n-B_n}\frac{dQ_n}{dP_n}dP_n \ge E[L] = 1.$$

We obtain $Q_n(B_n) \to 0$. Thus $\{Q_n\} \lhd \{P_n\}$.

We then prove the second half of the lemma. The probability measure $R_n = (P_n + Q_n)/2$ dominate both P_n and Q_n. Note that $\{dP_n/dQ_n\}$, $\{L_n\}$ and $W_n = dP_n/dR_n$ are tight with respect to $\{Q_n\}$, $\{P_n\}$ and $\{R_n\}$. By the Prohov's theorem, for any subsequence, there exists a further subsequence such that

$$\frac{dP_n}{dQ_n} \to_d U, \quad \text{under } Q_n,$$

$$L_n = \frac{dQ_n}{dP_n} \to_d L, \quad \text{under } P_n,$$

$$W_n = \frac{dP_n}{dR_n} \to_d W, \quad \text{under } R_n$$

for certain random variables U, V, and W. Since $E_{R_n}[W_n] = 1$ and $0 \le W_n \le 2$, we obtain $E[W] = 1$. For a given bounded, continuous function f, define $g(\omega) = f(\omega/(2-\omega))(2-\omega)$ for $0 \le \omega < 2$ and $g(2) = 0$. Then g is continuous. Thus,

$$E_{Q_n}[f(\frac{dP_n}{dQ_n})] = E_{R_n}[f(\frac{dP_n}{dQ_n})\frac{dQ_n}{dR_n}] = E_{R_n}[g(W_n)] \to E[f(\frac{W}{2-W})(2-W)].$$

Since $E_{Q_n}[f(dP_n/dQ_n)] \to E[f(U)]$, we have

$$E[f(U)] = E[f(\frac{W}{2-W})(2-W)].$$

Choose f_m in the above expression such that $f_m \le 1$ and f_m decreases to $I_{\{0\}}$. From the dominated convergence theorem, we have

$$P(U=0) = E[I_{\{0\}}(\frac{W}{2-W})(2-W)] = 2P(W=0).$$

However, since

$$P_n(\{\frac{dP_n}{dQ_n} \le \epsilon_n\} \cap \{q_n > 0\}) \le \int_{dP_n/dQ_n \le \epsilon_n} \frac{dP_n}{dQ_n} dQ_n \le \epsilon_n \to 0$$

and $\{Q_n\} \lhd \{P_n\}$,

$$P(U=0) = \lim_n P(U \le \epsilon_n) \le \liminf_n Q_n(\frac{dP_n}{dQ_n} \le \epsilon_n) = \liminf_n Q_n(\{\frac{dP_n}{dQ_n} \le \epsilon_n\} \cap \{q_n > 0\}) = 0.$$

That is, $P(W=0) = 0$. Similar to the above deduction, we obtain that

$$E[f(L)] = E[f(\frac{2-W}{W})W].$$

Choose f_m in the expression such that $f_m(x)$ increase to x. By the monotone convergence theorem, we have

$$E[L] = E[(2-W)I(W>0)] = 2P(W>0) - 1 = 1.$$

†

As a corollary, we have

Corollary 4.3 If $\log L_n \to_d N(-\sigma^2/2, \sigma^2)$ under P_n, then $\{Q_n\} \triangleleft \{P_n\}$. †

Proof Under P_n, $L_n \to_d \exp\{-\sigma^2/2 + \sigma Z\}$ where the limit has mean 1. The result thus follows from Proposition 4.4. †

Proposition 4.5 (Le Cam's third lemma) Let P_n and Q_n be sequence of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$, and let $X_n : \Omega_n \to R^k$ be a sequence of random vectors. Suppose that $Q_n \triangleleft P_n$ and under P_n,

$$(X_n, L_n) \to_d (X, L).$$

Then $G(B) = E[I_B(X)L]$ defines a probability measure, and under Q_n, $X_n \to_d G$. †

Proof Because $V \geq 0$, for countable disjoint sets $B_1, B_2, ...$, by the monotone convergence theorem,

$$G(\cup B_i) = E[\lim_n (I_{B_1} + ... + I_{B_n})L] = \lim_n \sum_{i=1}^n E[I_{B_i} L] = \sum_{i=1}^\infty G(B_i).$$

From Proposition 4.4, $E[L] = 1$. Then $G(\Omega) = 1$. G is a probability measure. Moreover, for any measurable simple function f, it is easy to see

$$\int f dG = E[f(X)L].$$

Thus, this equality holds for any measurable function f. In particular, for continuous and nonnegative function f, $(x, v) \mapsto f(x)v$ is continuous and nonnegative. Thus,

$$\liminf E_{Q_n}[f(X_n)] \geq \liminf \int f(X_n) \frac{dQ_n}{dP_n} dP_n \geq E[f(X)L].$$

Thus, under Q_n, $X_n \to_d G$. †

Remark 4.1 In fact, the name Le Cam's third lemma is often reserved for the following result. If under P_n,

$$(X_n, \log L_n) \to_d N_{k+1}\Big(\begin{pmatrix} \mu \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau & \sigma^2 \end{pmatrix} \Big),$$

then under Q_n, $X_n \to_d N_k(\mu + \tau, \Sigma)$. This result follows from Proposition 4.5 by noticing that the characteristic function of the limit distribution G is equal to $E[e^{itX} e^Y]$, where (X, Y) has the joint distribution

$$N_{k+1}\Big(\begin{pmatrix} \mu \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau & \sigma^2 \end{pmatrix} \Big).$$

Such a characteristic function is equal $\exp\{it'(\mu + \tau) - t'\Sigma t/2\}$, which is the characteristic function for $N_k(\mu + \tau, \Sigma)$.

4.4.3 Proof of the convolution theorem

Equipped with the Le Cam's two lemmas, we start to prove the convolution result in Theorem 4.4.

Proof of Theorem 4.4 We divide the proof into the following steps.
Step I. We first prove that the Hellinger differentiability condition (C1) implies that $P_{\theta_0}[\dot{l}_{\theta_0}] = 0$, the Fisher information $I(\theta_0) = E_{\theta_0}[\dot{l}_{\theta_0}\dot{l}'_{\theta_0}]$ exists, and moreover, for every convergent sequence $h_n \to h$, as $n \to \infty$,

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0 + h_n/\sqrt{n}}}{p_{\theta_0}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h'\dot{l}_{\theta_0}(X_i) - \frac{1}{2}h' I_{\theta_0} h + r_n,$$

where $r_n \to_p 0$. To see that , we abbreviate p_n, p, g as $p_{\theta_0+h/\sqrt{n}}$, p_{θ_0}, $h'\dot{l}_{\theta_0}$. Since $\sqrt{n}(\sqrt{p_n} - \sqrt{p})$ converges in $L_2(\mu)$ to $g\sqrt{p}/2$, $\sqrt{p_n}$ converges to $\sqrt{p}$ in $L_2(\mu)$. Then

$$E[g] = \int \frac{1}{2} g\sqrt{p} 2\sqrt{p} d\mu = \lim_{n\to\infty} \int \sqrt{n}(\sqrt{p_n} - \sqrt{p})(\sqrt{p_n} + \sqrt{p}) d\mu = 0.$$

Thus, $E_{\theta_0}[\dot{l}_{\theta_0}] = 0$. Let $W_{ni} = 2(\sqrt{p_n(X_i)/p(X_i)} - 1)$. We have

$$Var(\sum_{i=1}^{n} W_{ni} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i)) \leq E[(\sqrt{n} W_{ni} - g(X_i))^2] \to 0,$$

$$E[\sum_{i=1}^{n} W_{ni}] = 2n(\int \sqrt{p_n}\sqrt{p} d\mu - 1) = -n \int [\sqrt{p_n} - \sqrt{p}]^2 d\mu \to -\frac{1}{4} E[g^2].$$

Here, $E[g^2] = h' I(\theta_0) h$. By the Chebyshev's inequality, we obtain

$$\sum_{i=1}^{n} W_{ni} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \frac{1}{4} E[g^2] + a_n,$$

where $a_n \to_p 0$.

Next, by the Taylor expansion,

$$\log \prod_{i=1}^{n} \frac{p_n}{p}(X_i) = 2\sum_{i=1}^{n} \log(1 + \frac{1}{2} W_{ni}) = \sum_{i=1}^{n} W_{ni} - \frac{1}{4}\sum_{i=1}^{n} W_{ni}^2 + \frac{1}{2} \sum_{i=1}^{n} W_{ni}^2 R(W_{ni}),$$

where $R(x) \to 0$ as $x \to 0$. Since $E[(\sqrt{n} W_{ni} - g(X_i))^2] \to 0$, $nW_{ni}^2 = g(X_i)^2 + A_{ni}$ where $E[|A_{ni}|] \to 0$. Then $\sum_{i=1}^{n} W_{ni}^2 \to_p E[g^2]$. Moreover,

$$nP(|W_{ni}| > \epsilon\sqrt{2}) \leq nP(g(X_i)^2 > n\epsilon^2) + nP(|A_{ni}| > n\epsilon^2) \leq \epsilon^{-2} E[g^2 I(g^2 > n\epsilon^2)] + \epsilon^{-2} E[|A_{ni}|] \to 0.$$

The left-hand side is the upper bound for $P(\max_{1\leq i\leq n} |W_{ni}| > \epsilon)$. Thus, $\max_{1\leq i\leq n} |W_{ni}|$ converges to zero in probability; so is $\max_{1\leq i\leq n} |R(W_{ni})|$. Therefore,

$$\log \prod_{i=1}^{n} \frac{p_n}{p}(X_i) = \sum_{i=1}^{n} W_{ni} - \frac{1}{4} E[g^2] + b_n,$$

where $b_n \to_p 0$. Combining all the results, we obtain

$$\log \prod_{i=1}^n \frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h' \dot{l}_{\theta_0}(X_i) - \frac{1}{2} h' I_{\theta_0} h + r_n,$$

where $r_n \to_{p_n} 0$.

Step II. Let Q_n be the probability measure with density $\prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}(x_i)$ and P_n be the probability measure with $\prod_{i=1}^n p_{\theta_0}(x_i)$. Define

$$S_n = \sqrt{n}(T_n - \theta_0), \quad \Delta_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_{\theta_0}(X_i).$$

By the assumptions, S_n weakly converges to some distribution and so is Δ_n under P_n; thus, (S_n, Δ_n) is tight under P_n. By the Prohorov's theorem, for any subsequence, there exists a further subsequence such that $(S_n, \Delta_n) \to_d (S, \Delta)$ under P_n. From Step I, we immediately obtain that under P_n,

$$(S_n, \log \frac{dQ_n}{dP_n}) \to_d (S, h'\Delta - \frac{1}{2} h' I(\theta_0) h).$$

Since under P_n, dQ_n/dP_n weakly converges to $N(-h'I(\theta_0)h/2, h'I(\theta_0)h)$, Corollary 4.3 gives that $\{Q_n\} \triangleleft \{P_n\}$. Then from the Le Cam's third lemma, under Q_n, $S_n = \sqrt{n}(T_n - \theta_0)$ converges in distribution to a distribution G_h. Clearly, G_h is the same as distribution with $Z + h$.

Step III. We show $Z = Z_0 + \Delta_0$ where $Z_0 \sim N(0, I(\theta_0)^{-1})$ is independent of Δ_0. From Step II, we have

$$E_{\theta_0+h/\sqrt{n}}[\exp\{it'S_n\}] \to \exp\{it'h\} E[\exp\{it'Z\}].$$

On the other hand,

$$E_{\theta_0+h/\sqrt{n}}[\exp\{it'S_n\}] = E_{\theta_0}[\exp\{it'S_n + \log \frac{dQ_n}{dP_n}\}] + o(1) \to E_{\theta_0}[\exp\{it'Z + h'\Delta - \frac{1}{2} h' I(\theta_0) h\}].$$

We have

$$E_{\theta_0}[\exp\{it'Z + h'\Delta - \frac{1}{2} h' I(\theta_0) h\}] = \exp\{it'h\} E_{\theta_0}[\exp\{it'Z\}]$$

and it should hold for any complex number t and h. We let $h = -i(t' - s')I(\theta_0)^{-1}$ and obtain

$$E_{\theta_0}[\exp\{it'(Z - I(\theta_0)^{-1}\Delta) + is'I(\theta_0)^{-1}\Delta\}] = E_{\theta_0}[\exp\{it'Z + \frac{1}{2} t' I(\theta_0)^{-1} t\}] \exp\{-\frac{1}{2} s' I(\theta_0)^{-1} s\}.$$

This implies that $\Delta_0 = (Z - I(\theta_0)^{-1}\Delta)$ is independent of $Z_0 = I(\theta_0)^{-1}\Delta$ and Z_0 has the characteristics function $\exp\{-s'I(\theta_0)^{-1}s/2\}$, meaning $Z_0 \sim N(0, I(\theta_0)^{-1})$. Then $Z = Z_0 + \Delta_0$. †

The convolution theorem indicates that if $\{T_n\}$ is locally regular and the model $\mathcal{P}$ is the Hellinger differentiable and LAN, then the Cramér-Rao bound is also the asymptotic lower bound. We have shown that the result holds for estimating θ. In fact, the same procedure applies to estimating $q(\theta)$ where q is differentiable at θ_0. Then the local regularity condition is that under $P_{\theta_0+h/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0 + h/\sqrt{n})) \to_d Z,$$

where Z is independent of h. The result in Theorem 4.4 then becomes that $Z = Z_0 + \Delta_0$ where $Z_0 \sim N(0, \dot{q}(\theta_0)' I(\theta_0)^{-1} q(\theta_0))$ is independent of Δ_0.

4.4 Sufficient conditions for Hellinger-differentiability and local regularity

Checking the conditions of the local regularity and the Hellinger-differentiability and may be easy in practice. The following propositions give some sufficient conditions for the Hellinger differentiability and the local regularity.

Proposition 4.6. For every θ in an open subset of R^k let p_θ be a μ-probability density. Assume that the map $\theta \mapsto s_\theta(x) = \sqrt{p_\theta(x)}$ is continuously differentiable for every x. If the elements of the matrix $I(\theta) = E[(\dot{p}_\theta/p_\theta)(\dot{p}_\theta/p_\theta)']$ are well defined and continuous at θ. Then the map $\theta \to \sqrt{p_\theta}$ is Hellinger differentiable with $\dot{l}_\theta$ given by $\dot{p}_\theta/p_\theta$. †

Proof The map $\theta \mapsto p_\theta = s_\theta^2$ is differentiable. We have $\dot{p}_\theta = 2s_\theta\dot{s}_\theta$ so conclude $\dot{s}_\theta$ is zero whenever $\dot{p}_\theta = 0$. We can write $\dot{s}_\theta = (\dot{p}_\theta/p_\theta)\sqrt{p_\theta}/2$.

On the other hand,

$$\int \left\{ \frac{s_{\theta+th_t} - s_\theta}{t} \right\}^2 d\mu = \int \left\{ \int_0^1 (h_t)'\dot{s}_{\theta+uth_t} du \right\}^2 d\mu$$

$$\leq \int \int_0^1 ((h_t)'\dot{s}_{\theta+uth_t})^2 du d\mu = \frac{1}{2} \int_0^1 h_t' I(\theta + uth_t) h_t du.$$

As $h_t \to h$, the right side converges to $\int (h'\dot{s}_\theta)^2 d\mu$ by the continuity of I_θ. Since

$$\frac{s_{\theta+th_t} - s_\theta}{t} - h'\dot{s}_\theta$$

converges to zero almost surely, following the same proof as Theorem 3.1 (E) of Chapter 3, we obtain

$$\int \left[\frac{s_{\theta+th_t} - s_\theta}{t} - h'\dot{s}_\theta \right]^2 d\mu \to 0.$$

†

Proposition 4.7 If $\{T_n\}$ is an estimator sequence of $q(\theta)$ such that

$$\sqrt{n}(T_n - q(\theta)) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{q}_\theta I(\theta)^{-1} \dot{l}_\theta(X_i) \to_p 0,$$

where q is differentiable at θ, then T_n is the efficient and regular estimator for $q(\theta)$. †

Proof " $\Rightarrow$" Let $\Delta_{n,\theta} = n^{-1/2} \sum_{i=1}^n \dot{l}_\theta(X_i)$. Then $\Delta_{n,\theta}$ converges in distribution to a vector $\Delta_\theta \sim N(0, I(\theta))$. From Step I in proving Theorem 4.4, $\log dQ_n/dP_n$ is equivalent to $h'\Delta_{n,\theta} - h'I(\theta)h/2$ asymptotically. Thus, the Slutsky's theorem gives that under P_θ

$$\left(\sqrt{n}(T_n - q(\theta)), \log \frac{dQ_n}{dP_n} \right) \to_d (\dot{q}_\theta I(\theta)^{-1} \Delta_\theta, h'\Delta_\theta - h'I(\theta)h/2)$$

$$\sim N\left(\begin{pmatrix}0\\ -h'I(\theta)h/2\end{pmatrix}, \begin{pmatrix}\dot{q}_\theta' I(\theta)^{-1}\dot{q}_\theta & \dot{q}_\theta' h\\ \dot{q}_\theta h' & h'I(\theta)h\end{pmatrix}\right).$$

Then from the Le Cam's third lemma, under $P_{\theta+h/\sqrt{n}}$, $\sqrt{n}(T_n - q(\theta))$ converges in distribution to a normal distribution with mean $\dot{q}_\theta h$ and covariance matrix $\dot{q}_\theta' I(\theta)^{-1}\dot{q}_\theta$. Thus, under $P_{\theta+h/\sqrt{n}}$, $\sqrt{n}(T_n - q(\theta+h/\sqrt{n}))$ converges in distribution to $N(0, \dot{q}_\theta I(\theta)'\dot{q}_\theta')$. We obtain that T_n is regular. †

Definition 4.5 If a sequence of estimator $\{T_n\}$ has the expansion

$$\sqrt{n}(T_n - q(\theta)) = n^{-1/2}\sum_{i=1}^n \Gamma(X_i) + r_n,$$

where r_n converges to zero in probability, then T_n is called an *asymptotically linear estimator* for $q(\theta)$ with *influence function* Γ. Note that Γ depends on θ. †

For asymptotically linear estimator, the following result holds.

Proposition 4.8 Suppose T_n is an asymptotically linear estimator of $\nu = q(\theta)$ with influence function Γ. Then
A. T_n is Gaussian regular at θ_0 if and only if $q(\theta)$ is differentiable at θ_0 with derivative $\dot{q}_\theta$ and, with $\tilde{l}_\nu = \tilde{l}(\cdot, P_{\theta_0}|q(\theta), \mathcal{P})$ being the efficient influence function for $q(\theta)$, $E_{\theta_0}[(\Gamma - \tilde{l}_\nu)\dot{l}] = 0$ for any score $\dot{l}$ of $\mathcal{P}$.
B. Suppose $q(\theta)$ is differentiable and T_n is regular. Then $\Gamma \in [\dot{l}]$ if and only if $\Gamma = \tilde{l}_\nu$. †

Proof A. By asymptotic linearity of T_n, it follows that

$$\begin{pmatrix}\sqrt{n}(T_n - q(\theta_0))\\ L_n(\theta_0 + t_n/\sqrt{n}) - L_n(\theta_0)\end{pmatrix} \to_d N\left\{\begin{pmatrix}0\\ -t'I(\theta_0)t\end{pmatrix}, \begin{pmatrix}E_{\theta_0}[\Gamma\Gamma'] & E_{\theta_0}[\Gamma\dot{l}']t\\ E_{\theta_0}[\dot{l}\Gamma']t & t'I(\theta_0)t\end{pmatrix}\right\}.$$

From the Le Cam's third lemma, we obtain that under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0)) \to_d N(E_{\theta_0}[\Gamma'\dot{l}]t, E_{\theta_0}[\Gamma\Gamma']).$$

If T_n is regular, we have that under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0 + t_n/\sqrt{n})) \to_d N(0, E_{\theta_0}[\Gamma\Gamma']).$$

Comparing with the above convergence, we obtain

$$\sqrt{n}(q(\theta_0 + t_n/\sqrt{n}) - q(\theta_0)) \to E_{\theta_0}[\Gamma'\dot{l}]t.$$

This implies q is differentiable with $\dot{q}_\theta = E_\theta[\Gamma'\dot{l}]$. Since $E_{\theta_0}[\tilde{l}_\nu'\dot{l}] = \dot{q}_\theta$, the direction “ $\Rightarrow$″ holds.

To prove the other direction, since $q(\theta)$ is differentiable and under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0)) \to_d N(E_{\theta_0}[\Gamma'\dot{l}]t, E[\Gamma\Gamma'])$$

from the Le Cam's third lemma, we obtain under $P_{\theta_0+t_n/\sqrt{n}}$,

$$\sqrt{n}(T_n - q(\theta_0 + t_n/\sqrt{n})) \to_d N(0, E[\Gamma\Gamma']).$$

Thus, T_n is Gaussian regular.

B. If T_n is regular, from A, we obtain $\Gamma - \tilde{l}_\nu$ is orthogonal to any score in $\mathcal{P}$. Thus, $\Gamma \in [\dot{l}]$ implies that $\Gamma = \tilde{l}_\nu$. The converse is obvious. †

Remark 4.2 We have discussed the efficiency bound for real parameters. In fact, these results can be generalized (though non-trivial) to the situation where θ contains infinite dimensional parameter in semiparametric model. This generalization includes semiparametric efficiency bound, efficient score function, efficient influence function, locally regular estimator, Hellinger differentiability, LAN and the H$\acute{a}$jek convolution result.

READING MATERIALS: You should read Lehmann and Casella, Sections 1.6, 2.1, 2.2, 2.3, 2.5, 2.6, 6.1, 6.2, Ferguson, Chapter 19 and Chapter 20

PROBLEMS

1. Let $X_1, ..., X_n$ be i.i.d according to $Poisson(\lambda)$. Find the UMVU estimator of λ^k for any positive integer k.

2. Let $X_i, i = 1, ..., n$, be independently distributed as $N(\alpha + \beta t_i, \sigma^2)$ where α, β and σ^2 are unknown, and the t's are known constants that are not all equal. Find the least square estimators of α and β and show that they are also the UMVU estimators of α and β.

3. If X has the distribution $Poisson(\theta)$, show that $1/\theta$ does not have an unbiased estimator.

4. Suppose that we want to model the survival of twins with a common genetic defect, but with one of the two twins receiving some treatment. Let X represent the survival time of the untreated twin and let Y represent the survival time of the treated twin. One (overly simple) preliminary model might be to assume that X and Y are independent with Exponential(η) and Exponential($\theta\eta$) distributions, respectively:

$$f_{\theta,\eta}(x, y) = \eta e^{-\eta x} \eta\theta e^{-\eta\theta y} I(x > 0, y > 0).$$

   (a) On crude approach to estimation in this problem is to reduce the data to $W = X/Y$. Find the distribution of W and compute the Cram$\acute{e}$r-Rao lower bound for unbiased estimators of θ based on W.

   (b) Find the information bound for estimating θ based on observation of (X, Y) pairs when η is known and unknown.

   (c) Compare the bounds you computed in (a) and (b) and discuss the pros and cons of reducing to estimation based on the W.

5. This is a continuation of the preceding problem. A more realistic model involves assuming that the common parameter η for the two wins varies across sets of twins. There are several different ways of modeling this: one approach involves supposing that each pair of twins observed (X_i, Y_i) has its own fixed parameters $\eta_i, i = 1, .., n$. In this model we observe (X_i, Y_i) with density f_{θ,η_i} for $i = 1, ..., n$; i.e.,

$$f_{\theta,\eta_i}(x, y) = \eta_i e^{-\eta_i x_i} \eta_i \theta e^{-\eta_i \theta y_i} I(x_i > 0, y_i > 0).$$

   This is sometimes called a functional model (or model with incidental nuisance parameters).
   Another approach is to assume that $\eta \equiv Z$ has a distribution, and that our observations are from the mixture distribution. Assuming (for simplicity) that $Z = \eta \sim Gamma(a, 1/b)$ (a and b are known) with density

$$g_{a,b}(\eta) = \frac{b^a \eta^{a-1}}{\Gamma(a)} \exp\{-b\eta\} I(\eta > 0),$$

   it follows that the (marginal) distribution of (X, Y) is

$$p_{\theta,a,b}(x, y) \;=\; \int_0^\infty f_{\theta,z}(x, y) g_{a,b}(z) dz.$$

   This is sometimes called a "structural model" (or mixture model).

   (a) Find the information bound for θ in the functional model based on $(X_i, Y_i), i = 1, ..., n$.

   (b) Find the information bound for θ in the structural model based on $(X_i, Y_i), i = 1, ..., n$.

   (c) Compare the information bounds you computed in (a) and (b). When is the information for θ in the functional model larger than the information for θ in the structural model?

6. Suppose that $X \sim Gamma(\alpha, 1/\beta)$; i.e., X has density p_θ given by

$$p_\theta(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-\beta x\} I(x > 0), \quad \theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty).$$

   Consider estimation of $q(\theta) = E_\theta[X]$.

   (a) Compute the Fisher information matrix $I(\theta)$.

   (b) Derive the efficient score function, the efficient influence function and the efficient information bound for α.

   (c) Compute $\dot{q}(\theta)$ and find the efficient influence functions for estimation of $q(\theta)$. Compare the efficient influence functions you find in (c) with the influence function of the natural estimator $\bar{X}_n$.

7. Compute the score for location, $-(f'/f)(x)$, and the Fisher information when:

(a) $f(x) = \phi(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}$, (normal or Gaussian);

(b) $f(x) = \exp\{-x\}/(1+\exp\{-x\})^2$, (logistic);

(c) $f(x) = \exp\{-|x|\}/2$, (double exponential);

(d) $f(x) = t_k$, the t-distribution with k degrees of freedom;

(e) $f(x) = \exp\{-x\}\exp\{-\exp(-x)\}$, (Gumbel or extreme value).

8. Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}, \Theta \subset R^k$ is a parametric model satisfying the hypotheses of the multiparameter Craméer-Rao inequality. Partition θ as $\theta = (\nu, \eta)$, where $\nu \in R^m$ and $\eta \in R^{k-m}$ and $1 \le m < k$. Let $\dot{l} = \dot{l}_\theta = (\dot{l}_1, \dot{l}_2)$ be the corresponding partition of the scores and with $\tilde{l} = I^{-1}(\theta)\dot{l}$, the efficient influence function for θ, let $\tilde{l} = (\tilde{l}_1, \tilde{l}_2)$ be the corresponding partition of $\tilde{l}$. In both cases, $\dot{l}_1, \tilde{l}_1$ are m-vectors of functions and $\dot{l}_2, \tilde{l}_2$ are $k-m$ vectors. Partition $I(\theta)$ and $I^{-1}(\theta)$ correspondingly as

$$I(\theta) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix},$$

where I_{11} is $m \times m$, I_{12} is $m \times (k-m)$, I_{21} is $(k-m) \times m$, I_{22} is $(k-m) \times (k-m)$. also write

$$I^{-1}(\theta) = [I^{ij}]_{i,j=1,2}.$$

Verify that

(a) $I^{11} = I_{11\cdot 2}^{-1}$ where $I_{11\cdot 2} = I_{11} - I_{12}I_{22}^{-1}I_{21}$, $I^{22} = I_{22\cdot 1}^{-1}$ where $I_{22\cdot 1} = I_{22} - I_{21}I_{11}^{-1}I_{12}$, $I^{12} = -I_{11\cdot 2}^{-1}I_{12}I_{22}^{-1}$, $I^{21} = -I22 \cdot 1^{-1}I_{21}I_{11}^{-1}$..

(b) Verify that $\tilde{l}_1 = I^{11}\dot{l}_1 + I^{12}\dot{l}_2 = I_{11\cdot 2}^{-1}(\dot{l}_1 - I_{12}I_{22}^{-1}\dot{l}_2)$, and $\tilde{l}_2 = I^{21}\dot{l}_1 + I^{22}\dot{l}_2 = I_{22\cdot 1}^{-1}(\dot{l}_2 - I_{21}I_{11}^{-1}\dot{l}_1)$.

9. Let T_n be the Hodges superefficient estimator of θ.

(a) Show that T_n is not a regular estimator of θ at $\theta = 0$, but that it is regular at every $\theta \ne 0$. If $\theta_n = t/\sqrt{n}$, find the limiting distribution of $\sqrt{n}(T_n - \theta_n)$ under P_{θ_n}.

(b) For $\theta_n = t/\sqrt{n}$ show that

$$R_n(\theta_n) = nE_{\theta_n}[(T_n - \theta_n)^2] \to a^2 + t^2(1-a)^2.$$

This is larger than 1 if $t^2 > (1+a)/(1-a)$, and hence supper efficiency also entails worse risks in a local neighborhood of the points where the asymptotic variance is smaller.

10. Suppose that $(Y|Z) \sim Weibull(\lambda^{-1}\exp\{-\gamma Z\}, \beta)$ and $Z \sim G_\eta$ on R with density g_η with respect to some dominating measure μ. Thus the conditional cumulative hazards function $\Lambda(t|z)$ is given by

$$\Lambda_{\gamma,\lambda,\beta}(t|z) = (\lambda e^{\gamma z}t)^\beta = \lambda^\beta e^{\beta\gamma z}t^\beta$$

and hence

$$\lambda_{\gamma,\lambda,\beta}(t|z) = \lambda^\beta e^{\beta\gamma z}\beta t^{\beta-1}.$$

(Recall that $\lambda(t) = f(t)/(1 - F(t))$ and $\Lambda(t) = -\log(1 - F(t))$ if F is continuous). Thus it makes sense to reparameterize by defining $\theta_1 = \beta\gamma$ (this the parameter of interest since it reflects the effect of the covariate Z), $\theta_2 = \lambda^\beta$ and $\theta_2 = \beta$. This yields

$$\lambda_\theta(t|z) = \theta_2\theta_3 \exp\{\theta_1 z\} t^{\theta_3 - 1}.$$

You may assume that $a(z) = (\partial/\partial z) \log g_\eta(z)$ exists and $E[a(Z)^2] < \infty$. Thus Z is a "covariate" or "predictor variable", θ_1 is a "regression parameter" which affects the intensity the (conditionally) Exponential variable Y, and $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ where $\theta_4 = \eta$.

(a) Derive the joint density $p_\theta(y, z)$ of (Y, Z) for the reparameterized model.

(b) Find the information matrix for θ. What does the structure of this matrix say about the effect of $\eta = \theta_4$ being known or unknown about the estimation of $\theta_1, \theta_2, \theta_3$?

(c) Find the information and information bound for θ_1 if the parameter θ_2 and θ_3 are known.

(d) What is the information for θ_1 if just θ_3 is known to be equal to 1?

(e) Find the efficient score function and the efficient influence function for estimation of θ_1 when θ_3 is known.

(f) Find the information $I_{11\cdot(2,3)}$ and information bound for θ_1 if the parameters θ_2 and θ_3 are unknown.

(g) Find the efficient score function and the efficient influence function for estimation of θ_1 when θ_2 and θ_3 are unknown.

(h) Specialize the calculation in (d)-(g) to the case when $Z \sim Bernoulli(\theta_4)$ and compare the information bounds.

11. Lehmann and Casella, page 72, problems 6.33, 6.34, 6.35

12. Lehmann and Casella, pages 129-137, problems 1.1-3.30

13. Lehamann and Casella, pages 138-143, problems 5.1-6.12

14. Lehmann and Casella, pages 496-501, problems 1.1-2.14

15. Ferguson, pages 131-132, problems 2-5

16. Ferguson, page 139, problems 1-4

CHAPTER 5 EFFICIENT ESTIMATION: MAXIMUM LIKELIHOOD APPROACH

In the previous chapter, we have discussed the asymptotic lower bound (efficiency bound) for all the regular estimators. Then a natural question is what estimator can achieve this bound; equivalently, what estimator can be asymptotically efficient. In this chapter, we will focus on the most commonly-used estimator, maximum likelihood estimator. We will show that under some regularity conditions, the maximum likelihood estimator is asymptotically efficient.

Suppose $X_1, ..., X_n$ are i.i.d from P_{θ_0} in the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We assume

(A0). $\theta \neq \theta^*$ implies $P_\theta \neq P_{\theta^*}$ (identifiability).
(A1). P_θ has a density function p_θ with respect to a dominating σ-finite measure μ.
(A2). The set $\{x : p_\theta(x) > 0\}$ does not depend on θ.

Furthermore, we denote

$$L_n(\theta) = \prod_{i=1}^n p_\theta(X_i), \quad l_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i).$$

$L_n(\theta)$ and $l_n(\theta)$ are called the *likelihood function* and the *log-likelihood function* of θ, respectively. An estimator $\hat{\theta}_n$ of θ_0 is the maximum likelihood estimator (MLE) of θ_0 if it maximizes the likelihood function $L_n(\theta)$, equivalently, $l_n(\theta)$.

Some cautions should be taken in the maximization: first, the maximum likelihood estimator may not exist; second, even if the maximum likelihood estimator exists, it may not be unique; third, the definition of the maximum likelihood estimator depends on the parameterization of p_θ so different parameterization may lead to the different estimators.

5.1 Ad Hoc Arguments of MLE Efficiency

In the following, we explain the intuition why the maximum likelihood estimator is the efficient estimator; while we leave rigorous conditions and arguments to the subsequent sections. First, to see the consistency of the maximum likelihood estimator, we introduce the definition of the Kullback-Leibler information as follows.

Definition 5.1 Let P be a probability measure and let Q be another measure on $(\Omega, \mathcal{A})$ with densities p and q with respect to a σ-finite measure μ ($\mu = P + Q$ always works). $P(\Omega) = 1$ and $Q(\Omega) \leq 1$. Then the *Kullback-Leibler information* $K(P, Q)$ is

$$K(P, Q) = E_P[\log \frac{p(X)}{q(X)}].$$

†

Immediately, we obtain the following result.

Proposition 5.1 $K(P, Q)$ is well-defined, and $K(P, Q) \geq 0$. $K(P, Q) = 0$ if and only if $P = Q$.
†

Proof By the Jensen's inequality,

$$K(P,Q) = E_P[-\log\frac{q(X)}{p(X)}] \ge -\log E_P[\frac{q(X)}{p(X)}] = -\log Q(\Omega) \ge 0.$$

The equality holds if and only if $p(x) = Mq(x)$ almost surely with respect P and $Q(\Omega) = 1$. Thus, $M = 1$ and $P = Q$. †

Now that $\hat{\theta}_n$ maximizes $l_n(\theta)$,

$$\frac{1}{n}\sum_{i=1}^{n} p_{\hat{\theta}_n}(X_i) \ge \frac{1}{n}\sum_{i=1}^{n} p_{\theta_0}(X_i).$$

Suppose $\hat{\theta}_n \to \theta^*$. Then we would expect to the both sides converge to

$$E_{\theta_0}[p_{\theta^*}(X)] \ge E_{\theta_0}[p_{\theta_0}(X)],$$

which implies $K(P_{\theta_0}, P_{\theta^*}) \le 0$. From Proposition 5.1, $P_{\theta_0} = P_{\theta^*}$. From (A0), $\theta^* = \theta_0$ (the model identifiability condition is used here). That is, $\hat{\theta}_n$ converges to θ_0. Note in this argument, three conditions are essential: (i) $\hat{\theta}_n \to \theta^*$ (compactness of $\hat{\theta}_n$); (ii) the convergence of $n^{-1}l_n(\hat{\theta}_n)$ (locally uniform convergence); (iii) $P_{\theta_0} = P_{\theta^*}$ implies $\theta_0 = \theta^*$ (identifiability).

Next, we give an ad hoc discussion on the efficiency of the maximum likelihood estimator. Suppose $\hat{\theta}_n \to \theta_0$. If $\hat{\theta}_n$ is in the interior of Θ, $\hat{\theta}_n$ solves the following likelihood (or score) equations

$$\dot{l}_n(\hat{\theta}_n) = \sum_{i=1}^{n} \dot{l}_{\hat{\theta}_n}(X_i) = 0.$$

Suppose $\dot{l}_\theta(X)$ is twice-differentiable with respect to θ. We apply the Taylor expansion to $\dot{l}_{\hat{\theta}_n}(X_i)$ at θ_0 and obtain

$$-\sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) = \sum_{i=1}^{n} \ddot{l}_{\theta^*}(X_i)(\hat{\theta} - \theta_0),$$

where θ^* is between θ_0 and $\hat{\theta}$. This gives that

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{1}{\sqrt{n}}\left\{n^{-1}\sum_{i=1}^{n} \ddot{l}_{\theta^*}(X_i)\right\}^{-1}\left\{\sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i)\right\}.$$

By the law of large number, we can see $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(\theta_0)^{-1}\dot{l}_{\theta_0}(X_i).$$

Then $\hat{\theta}_n$ is an asymptotically linear estimator of θ_0 with the influence function $I(\theta_0)^{-1}\dot{l}_{\theta_0} = \tilde{l}(\cdot, P_{\theta_0}|\theta, \mathcal{P})$. This shows that $\hat{\theta}_n$ is the efficient estimator of θ_0 and the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ attains the efficiency bound, which was defined in the previous chapter. Again, the above arguments require a few conditions to go through.

As mentioned before, in the following sections we will rigorously prove the consistency and the asymptotic efficiency of the maximum likelihood estimator. Moreover, we will discuss the computation of the maximum likelihood estimators and some alternative efficient estimation approaches.

5.2 Consistency of Maximum Likelihood Estimator

We provide some sufficient conditions for obtaining the consistency of maximum likelihood estimator.

Theorem 5.1 Suppose that
(a) Θ is compact.
(b) $\log p_\theta(x)$ is continuous in θ for all x.
(c) There exists a function $F(x)$ such that $E_{\theta_0}[F(X)] < \infty$ and $|\log p_\theta(x)| \leq F(x)$ for all x and θ.
Then $\hat{\theta}_n \rightarrow_{a.s.} \theta_0$. †

Proof For any sample $\omega \in \Omega$, $\hat{\theta}_n$ is compact. Thus, be choosing a subsequence, we assume $\hat{\theta}_n \rightarrow \theta^*$. Suppose we can show that

$$\frac{1}{n}\sum_{i=1}^n l_{\hat{\theta}_n}(X_i) \rightarrow E_{\theta_0}[l_{\theta^*}(X)].$$

Then since

$$\frac{1}{n}\sum_{i=1}^n l_{\hat{\theta}_n}(X_i) \geq \frac{1}{n}\sum_{i=1}^n l_{\theta_0}(X_i),$$

we have

$$E_{\theta_0}[l_{\theta^*}(X)] \geq E_{\theta_0}[l_{\theta_0}(X)].$$

Thus Proposition 5.1 plus the identifiability gives $\theta^* = \theta_0$. That is, any subsequence of $\hat{\theta}_n$ converges to θ_0. We conclude that $\hat{\theta}_n \rightarrow_{a.s.} \theta_0$.

It remains to show

$$\mathbf{P}_n[l_{\hat{\theta}_n}(X)] \equiv \frac{1}{n}\sum_{i=1}^n l_{\hat{\theta}_n}(X_i) \rightarrow E_{\theta_0}[l_{\theta^*}(X)].$$

Since

$$E_{\theta_0}[l_{\hat{\theta}_n}(X)] \rightarrow E_{\theta_0}[l_{\theta^*}(X)]$$

by the dominated convergence theorem, it suffices to show

$$|\mathbf{P}_n[l_{\hat{\theta}_n}(X)] - E_{\theta_0}[l_{\hat{\theta}_n}(X)]| \rightarrow 0.$$

We can even prove the following uniform convergence result

$$\sup_{\theta\in\Theta} |\mathbf{P}_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]| \rightarrow 0.$$

To see this, we define

$$\psi(x,\theta,\rho) = \sup_{|\theta'-\theta|<\rho} (l_{\theta'}(x) - E_{\theta_0}[l_{\theta'}(X)]).$$

Since l_θ is continuous, $\psi(x,\theta,\rho)$ is measurable and by the dominance convergence theorem, $E_{\theta_0}[\psi(X,\theta,\rho)]$ decreases to $E_{\theta_0}[l_\theta(x) - E_{\theta_0}[l_\theta(X)]] = 0$. Thus, for $\epsilon > 0$, for any $\theta \in \Theta$, there exists a ρ_θ such that

$$E_{\theta_0}[\psi(X,\theta,\rho_\theta)] < \epsilon.$$

The union of $\{\theta' : |\theta'-\theta| < \rho_\theta\}$ covers Θ. By the compactness of Θ, there exists a finite number of $\theta_1, ..., \theta_m$ such that

$$\Theta \subset \cup_{i=1}^m \{\theta' : |\theta' - \theta_i| < \rho_{\theta_i}\}.$$

Therefore,

$$\sup_{\theta\in\Theta} \{\mathbf{P}_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]\} \quad \leq \quad \sup_{1\leq i\leq m} \mathbf{P}_n[\psi(X, \theta_i, \rho_{\theta_i})].$$

We obtain

$$\limsup_n \sup_{\theta\in\Theta} \{\mathbf{P}_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]\} \leq \sup_{1\leq i\leq m} \mathbf{P}_\theta[\psi(X, \theta_i, \rho_{\theta_i})] \leq \epsilon.$$

Thus, $\limsup_n \sup_{\theta\in\Theta} \{\mathbf{P}_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]\} \leq 0$. We apply the similar arguments to $\{-l(X, \theta)\}$ and obtain $\limsup_n \sup_{\theta\in\Theta} \{-\mathbf{P}_n[l_\theta(X)] + E_{\theta_0}[l_\theta(X)]\} \leq 0$. Thus,

$$\lim_n \sup_{\theta\in\Theta} |\mathbf{P}_n[l_\theta(X)] - E_{\theta_0}[l_\theta(X)]| \to 0.$$

†

As a note, condition (c) in Theorem 5.1 is necessary. Ferguson (2002) page 116 gives an interesting counterexample showing that if (c) is invalid, the maximum likelihood estimator converges to a fixed constant whatever true parameter is.

Another type of consistency result is the classical Wald's consistency result.

Theorem 5.2 (Wald's Consistency) Θ is compact. Suppose $\theta \mapsto l_\theta(x) = \log p_\theta(x)$ is upper-semicontinuous for all x, in the sense

$$\limsup_{\theta'\to\theta} l_{\theta'}(x) \leq l_\theta(x).$$

Suppose for every sufficient small ball $U \subset \Theta$,

$$E_{\theta_0}[\sup_{\theta'\in U} l_{\theta'}(X)] < \infty.$$

Then $\hat{\theta}_n \to_p \theta_0$. †

Proof Since $E_{\theta_0}[l_{\theta_0}(X)] > E_{\theta_0}[l_{\theta'}(X)]$ for any $\theta' \neq \theta_0$, there exists a ball $U_{\theta'}$ containing θ' such that

$$E_{\theta_0}[l_{\theta_0}(X)] > E_{\theta_0}[\sup_{\theta^*\in U_{\theta'}} l_{\theta^*}(X)].$$

Otherwise, there exists a sequence $\theta_m^* \to \theta'$ but $E_{\theta_0}[l_{\theta_0}(X)] \leq E_{\theta_0}[l_{\theta_m^*}(X)]$. Since $l_{\theta_m^*}(x) \leq \sup_{U'} l_{\theta'}(X)$ where U' is the ball satisfying the condition, we obtain

$$\limsup_m E_{\theta_0}[l_{\theta_m^*}(X)] \leq E_{\theta_0}[\limsup_m l_{\theta_m^*}(X)] \leq E_{\theta_0}[l_{\theta'}(X)].$$

We then obtain $E_{\theta_0}[l_{\theta_0}(X)] \leq E_{\theta_0}[l_{\theta'}(X)]$ and this is a contradiction.

For any ϵ, the balls $\cup_{\theta'} U_{\theta'}$ covers the compact set $\Theta \cap \{|\theta' - \theta_0| > \epsilon\}$ so there exists a finite covering balls, $U_1, ..., U_m$. Then

$$P(|\hat{\theta}_n - \theta_0| > \epsilon) \le P(\sup_{|\theta'-\theta_0|>\epsilon} \mathbf{P}_n[l_{\theta'}(X)] \ge \mathbf{P}_n[l_{\theta_0}(X)]) \le P(\max_{1\le i\le m} \mathbf{P}_n[\sup_{\theta'\in U_i} l_{\theta'}(X)] \ge \mathbf{P}_n[l_{\theta_0}(X)])$$

$$\le \sum_{i=1}^m P(\mathbf{P}_n[\sup_{\theta'\in U_i} l_{\theta'}(X)] \ge \mathbf{P}_n[l_{\theta_0}(X)]).$$

Since

$$\mathbf{P}_n[\sup_{\theta'\in U_i} l_{\theta'}(X)] \to_{a.s.} E_{\theta_0}[\sup_{\theta'\in U_i} l_{\theta'}(X)] < E_{\theta_0}[l_{\theta_0}(X)],$$

the right-hand side converges to zero. Thus, $\hat{\theta}_n \to_p \theta_0$. †

5.3. Asymptotic Efficiency of Maximum Likelihood Estimator

The following theorem gives some regular conditions so that the maximum likelihood estimator attains asymptotic efficiency bound.

Theorem 5.3 Suppose that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is Hellinger differentiable at an inner point θ_0 of $\Theta \subset R^k$. Furthermore, suppose that there exists a measurable function $F(X)$ with $E_{\theta_0}[F(X)^2] < \infty$ such that for every θ_1 and θ_2 in a neighborhood of θ_0,

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \le F(x)|\theta_1 - \theta_2|.$$

If the Fisher information matrix $I(\theta_0)$ is nonsingular and $\hat{\theta}_n$ is consistent, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n I(\theta_0)^{-1} \dot{l}_{\theta_0}(X_i) + o_p(1).$$

In particular, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $I(\theta_0)^{-1}$.†

Proof For any $h_n \to h$, by the Hellinger differentiability,

$$W_n = 2\left(\sqrt{\frac{p_{\theta_0+h_n/\sqrt{n}}}{p_{\theta_0}}} - 1\right) \to h'\dot{l}_{\theta_0}, \ \text{ in } L_2(P_{\theta_0}).$$

We obtain

$$\sqrt{n}(\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}) = 2\sqrt{n}\log(1 + W_n/2) \to_p h'\dot{l}_{\theta_0}.$$

Using the Lipschitz continuity of $\log p_\theta$ and the dominate convergence theorem, we can show

$$E_{\theta_0}\left[\sqrt{n}(\mathbf{P}_n - P)[\sqrt{n}(\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}) - h'\dot{l}_{\theta_0}]\right] \to 0$$

and

$$Var_{\theta_0}\left[\sqrt{n}(\mathbf{P}_n - P)[\sqrt{n}(\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}) - h'\dot{l}_{\theta_0}]\right] \to 0.$$

Thus,

$$\sqrt{n}(\mathbf{P}_n - P)[\sqrt{n}(\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}) - h'\dot{l}_{\theta_0}] \to_p 0,$$

where $\sqrt{n}(\mathbf{P}_n - P)[g(X)]$ is defined as

$$n^{-1/2}\left[\sum_{i=1}^n \{g(X_i) - E_{\theta_0}[g(X)]\}\right].$$

From Step I in proving Theorem 4.4, we know

$$\log\prod_{i=1}^n \frac{\log p_{\theta_0+h_n/\sqrt{n}}}{\log p_{\theta_0}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n h'\dot{l}_{\theta_0}(X_i) - \frac{1}{2}h'I(\theta_0)h + o_p(1).$$

We obtain

$$nE_{\theta_0}[\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}] \to -h'I(\theta_0)h/2.$$

Hence the map $\theta \mapsto E_{\theta_0}[\log p_\theta]$ is twice-differentiable with second derivative matrix $-I(\theta_0)$.

Furthermore, we obtain

$$n\mathbf{P}_n[\log p_{\theta_0+h_n/\sqrt{n}} - \log p_{\theta_0}] = -\frac{1}{2}h_n'I(\theta_0)h_n + h_n'\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}] + o_p(1).$$

We choose $h_n = \sqrt{n}(\hat{\theta}_n - \theta_0)$ and $h_n = I(\theta_0)^{-1}\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}]$. It gives that

$$n\mathbf{P}_n[\log p_{\hat{\theta}_n} - \log p_{\theta_0}] = -\frac{n}{2}(\hat{\theta}_n - \theta_0)'I(\theta_0)(\hat{\theta} - \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0)\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}] + o_p(1),$$

$$n\mathbf{P}_n[\log p_{\theta_0+I(\theta_0)^{-1}\sqrt{n}(\mathbf{P}_n-P)[\dot{l}_{\theta_0}]/\sqrt{n}} - \log p_{\theta_0}]$$
$$= \frac{1}{2}\{\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}]\}'I(\theta_0)^{-1}\{\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}]\} + o_p(1).$$

Since the left-hand side of the fist equation is larger than the left-hand side of the second equation, after simple algebra, we obtain

$$-\frac{1}{2}\left\{\sqrt{n}(\hat{\theta}_n - \theta_0) - I(\theta_0)^{-1}\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}]\right\}' I(\theta_0)\left\{\sqrt{n}(\hat{\theta}_n - \theta_0) - I(\theta_0)^{-1}\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}]\right\}$$
$$+o_p(1) \geq 0.$$

Thus,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1}\sqrt{n}(\mathbf{P}_n - P)[\dot{l}_{\theta_0}] + o_p(1).$$

†

A classical condition for the asymptotic normality for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is the following theorem.

Theorem 5.4 For each θ in an open subset of Euclidean space. Let $\theta \mapsto \dot{l}_\theta(x) = \log p_\theta(x)$ be twice continuously differentiable for every x. Suppose $E_{\theta_0}[\dot{l}_{\theta_0}\dot{l}_{\theta_0}'] < \infty$ and $E[\ddot{l}_{\theta_0}]$ exists and is nonsingular. Assume that the second partial derivative of $\dot{l}_\theta(x)$ is dominated by a fixed integrable function $F(x)$ for every θ in a neighborhood of θ_0. Suppose $\hat{\theta}_n \rightarrow_p \theta_0$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(E_{\theta_0}[\ddot{l}_{\theta_0}])^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}_{\theta_0}(X_i) + o_p(1).$$

†

Proof $\hat{\theta}_n$ solves the equation

$$0 = \sum_{i=1}^n \dot{l}_{\hat{\theta}}(X_i).$$

After the Taylor expansion, we obtain

$$0 = \sum_{i=1}^n \dot{l}_{\theta_0}(X_i) + \sum_{i=1}^n \ddot{l}_{\theta_0}(X_i)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)' \left\{ \sum_{i=1}^n l^{(3)}_{\tilde{\theta}_n}(X_i) \right\} (\hat{\theta}_n - \theta_0),$$

where $\tilde{\theta}_n$ is between $\hat{\theta}_n$ and θ_0. Thus,

$$| \left\{ \frac{1}{n} \sum_{i=1}^n \ddot{l}_{\theta_0}(X_i) \right\} (\hat{\theta}_n - \theta_0) + \frac{1}{n} \sum_{i=1}^n \dot{l}_{\theta_0}(X_i)| \le \frac{1}{n} \sum_{i=1}^n |F(X_i)||\hat{\theta}_n - \theta_0|^2.$$

We obtain $(\hat{\theta}_n - \theta_0) = o_p(1/\sqrt{n})$. Then it holds

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \left\{ \frac{1}{n} \sum_{i=1}^n \ddot{l}_{\theta_0}(X_i) + o_p(1) \right\} = -\frac{1}{\sqrt{n}} \sum_{i=1}^n l_{\theta_0}(X_i).$$

The result holds. † .

5.4 Computation of Maximum Likelihood Estimate

A variety of methods can be used to compute the maximum likelihood estimate. Since the maximum likelihood estimate, $\hat{\theta}_n$, solves the likelihood equation

$$\sum_{i=1}^n \dot{l}_\theta(X_i) = 0,$$

one numerical method for the calculation is via the *Newton-Raphson iteration*: at kth iteration,

$$\theta^{(k+1)} = \theta^{(k)} - \left\{ \frac{1}{n} \sum_{i=1}^n \ddot{l}_{\theta^{(k)}}(X_i) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^n \dot{l}_{\theta^{(k)}}(X_i) \right\}.$$

Sometimes, calculating $\ddot{l}_\theta$ may be complicated. Note the

$$-\frac{1}{n} \sum_{i=1}^n \ddot{l}_{\theta^{(k)}}(X_i) \approx I(\theta^{(k)}).$$

Then a *Fisher scoring algorithm* is via the following iteration

$$\theta^{(k+1)} = \theta^{(k)} + I(\theta^{(k)})^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n} \dot{l}_{\theta^{(k)}}(X_i)\right\}.$$

An alternative method to find the maximum likelihood estimate is by optimum search algorithm. Note that the objective function is $L_n(\theta)$. Then a simple search method is grid search by evaluating the $L_n(\theta)$ along a number of θ's in the parameter space. Clearly, such a method is only feasible with very low-dimensional θ. Other efficient methods include quasi-Newton search (gradient-decent search) where at each θ, we search along the direction of $\dot{L}_n(\theta)$. Recent development has seen many Bayesian computation methods, including MCMC, simulation annealing etc.

In this section, we particularly focus on the calculation of the maximum likelihood estimate when part of data are missing or some mis-measured data are observed. In such calculation, a useful algorithm is called the *expectation-maximization* (EM) algorithm. We will describe this algorithm in detail and explain why the EM algorithm may give the maximum likelihood estimate. A few examples are given for illustration.

5.4.1 EM framework

Suppose Y denotes the vector of statistics from n subjects. In many practical problems, Y can not be fully observed due to data missingness; instead, partial data or a function of Y is observed. For simplicity, suppose $Y = (Y_{mis}, Y_{obs})$, where Y_{obs} is the part of Y which is observed and Y_{mis} is the part of Y which is not observed. Furthermore, we introduce R as a vector of 0/1 indicating which subjects are missing/not missing. Then the observed data include (Y_{obs}, R).

Assume Y has a density function $f(Y;\theta)$ where $\theta \in \Theta$. Then the density function for the observed data (Y_{obs}, R)

$$\int_{Y_{mis}} f(Y;\theta)P(R|Y)dY_{mis},$$

where $P(R|Y)$ denotes the conditional probability of R given Y. One additional assumption is that $P(R|Y) = P(R|Y_{obs})$ and $P(R|Y)$ does not depend on θ; i.e., the missing probability only depends on the observed data and it is non-informative about θ. Such an assumption is called the *missing at random* (MAR) and is often assumed for missing data problem. Under the MAR, the density function for the observed data is equal

$$\int_{Y_{mis}} f(Y;\theta)dY_{mis}P(R|Y).$$

Hence, if we wish to calculate the maximum likelihood estimator for θ, we can ignore the part of $P(R|Y)$ but simply maximize the part of $\int_{Y_{mis}} f(Y;\theta)dY_{mis}$. Note the latter is exactly the marginal density of Y_{obs}, denoted by $f(Y_{obs};\theta)$.

The way of the EM algorithm is as follows: we start from any initial value of $\theta^{(1)}$ and use the following iterations. The kth iteration consists both E-step and M-step:

E-step. We evaluate the conditional expectation

$$E\left[\log f(Y;\theta)|Y_{obs}, \theta^{(k)}\right].$$

Here, $E[\cdot|Y_{obs}, \theta^k]$ is the conditional expectation given the observed data and the current value of θ. That is,

$$E\left[\log f(Y;\theta)|Y_{obs}, \theta^{(k)}\right] = \frac{\int_{Y_{mis}}[\log f(Y;\theta)]f(Y;\theta^{(k)})dY_{mis}}{\int_{Y_{mis}} f(Y;\theta^{(k)})dY_{mis}}.$$

Such an expectation can often be evaluated using simple numerical calculation, as will be seen in the later examples.

M-step. We obtain $\theta^{(k+1)}$ by maximizing

$$E\left[\log f(Y;\theta)|Y_{obs}, \theta^{(k)}\right].$$

We then iterate till the convergence of θ; i.e., the difference between $\theta^{(k+1)}$ and $\theta^{(k)}$ is less than a given criteria.

The reason why the EM algorithm may give the maximum likelihood estimator is the following result.

Theorem 5.5 At each iteration of the EM algorithm, $\log f(Y_{obs};\theta^{(k+1)}) \geq \log f(Y_{obs};\theta^{(k)})$ and the equality holds if and only if $\theta^{(k+1)} = \theta^{(k)}$. †

Proof From the EM algorithm, we see

$$E\left[\log f(Y;\theta^{(k+1)})|Y_{obs}, \theta^{(k)}\right] \geq E\left[\log f(Y;\theta^{(k)})|Y_{obs}, \theta^{(k)}\right].$$

Since

$$\log f(Y;\theta) = \log f(Y_{obs};\theta) + \log f(Y_{mis}|Y_{obs}, \theta),$$

we obtain

$$E\left[\log f(Y_{mis}|Y_{obs}, \theta^{(k+1)})|Y_{obs}, \theta^{(k)}\right] + \log f(Y_{obs};\theta^{(k+1)})$$
$$\geq E\left[\log f(Y_{mis}|Y_{obs}, \theta^{(k)})|Y_{obs}, \theta^{(k)}\right] + \log f(Y_{obs};\theta^{(k)}).$$

On the other hand, since

$$E\left[\log f(Y_{mis}|Y_{obs}, \theta^{(k+1)})|Y_{obs}, \theta^{(k)}\right] \leq E\left[\log f(Y_{mis}|Y_{obs}, \theta^{(k)})|Y_{obs}, \theta^{(k)}\right]$$

by the non-negativity of the Kullback-Leibler information, we conclude that $\log f(Y_{obs};\theta^{(k+1)}) \geq \log f(Y_{obs}, \theta^{(k)})$. The equality implies

$$\log f(Y_{mis}|Y_{obs}, \theta^{(k+1)}) = \log f(Y_{mis}|Y_{obs}, \theta^{(k)}),$$

and hence $\log f(Y;\theta^{(k+1)}) = \log f(Y;\theta^{(k)})$, thus $\theta^{(k+1)} = \theta^{(k)}$. †

From Theorem 5.5, we conclude that each iteration of the EM algorithm increases the observed likelihood function. Thus, it is expected that $\theta^{(k)}$ will eventually converge to the maximum likelihood estimate. If the initial value of the EM algorithm is chosen close to the maximum likelihood estimate (though we never know) and the objective function is concave in the neighborhood of the maximum likelihood estimate, then the maximization in the M-step can be replaced by the Newton-Raphson iteration. Correspondingly, an alternative way to the EM algorithm is given by:

E-step. We evaluate the conditional expectation

$$E\left[\frac{\partial}{\partial\theta}\log f(Y;\theta)|Y_{obs},\theta^{(k)}\right]$$

and

$$E\left[\frac{\partial^2}{\partial\theta^2}\log f(Y;\theta)|Y_{obs},\theta^{(k)}\right]$$

M-step. We obtain $\theta^{(k+1)}$ by solving

$$0=E\left[\frac{\partial}{\partial\theta}\log f(Y;\theta)|Y_{obs},\theta^{(k)}\right]$$

using one-step Newton-Raphson iteration:

$$\theta^{(k+1)}=\theta^{(k)}-\left\{E\left[\frac{\partial^2}{\partial\theta^2}\log f(Y;\theta)|Y_{obs},\theta^{(k)}\right]\right\}^{-1}E\left[\frac{\partial}{\partial\theta}\log f(Y;\theta)|Y_{obs},\theta^{(k)}\right]\Bigg|_{\theta=\theta^{(k)}}.$$

We note that in the second form of the EM algorithm, only one-step Newton-Raphson iteration is used in the M-step since it still ensures that the iteration will increase the likelihood function.

5.4.2 Examples of using EM algorithm

Example 5.1 Suppose a random vector Y has a multinomial distribution with $n=197$ and

$$p=(\frac{1}{2}+\frac{\theta}{4},\frac{1-\theta}{4},\frac{1-\theta}{4},\frac{\theta}{4}).$$

Then the probability for $Y=(y_1,y_2,y_3,y_4)$ is given by

$$\frac{n!}{y_1!y_2!y_3!y_4!}(\frac{1}{2}+\frac{\theta}{4})^{y_1}(\frac{1-\theta}{4})^{y_2}(\frac{1-\theta}{4})^{y_3}(\frac{\theta}{4})^{y_4}.$$

If we use the Newton-Raphson iteration to calculate the maximum likelihood estimator for θ, then after calculating the first and the second derivative of the log-likelihood function, we iterate using

$$\theta^{(k+1)}=\theta^{(k)}+\left\{Y_1\frac{1/16}{(1/2+\theta^{(k)}/4)^2}+(Y_2+Y_3)\frac{1}{(1-\theta^{(k)})^2}+Y_4\frac{1}{\theta^{(k)2}}\right\}^{-1}$$

$$\times\left\{Y_1\frac{1/4}{1/2+\theta^{(k)}/4}-(Y_2+Y_3)\frac{1}{1-\theta^{(k)}}+Y_4\frac{1}{\theta^{(k)}}\right\}.$$

Suppose we observe $Y=(125,18,20,34)$. If we start with $\theta^{(1)}=0.5$, after the convergence, we obtain $\theta^{(k)}=0.6268215$. We can use the EM algorithm to calculate the maximum likelihood estimator. Suppose the full data is X which has a multivariate normal distribution with n and the $p = (1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$. Then Y can be treated as an incomplete data of X by $Y = (X_1 + X_2, X_3, X_4, X_5)$. The score equation for the complete data X is simple

$$0 = \frac{X_2 + X_5}{\theta} - \frac{X_3 + X_4}{1-\theta}.$$

Thus we note the M-step of the EM algorithm needs to solve the equation

$$0 = E\left[\frac{X_2 + X_5}{\theta} - \frac{X_3 + X_4}{1-\theta} | Y, \theta^{(k)}\right];$$

while the E-step evaluates the above expectation. By simple calculation,

$$E[X|Y, \theta^{(k)}] = (Y_1 \frac{1/2}{1/2 + \theta^{(k)}/4}, Y_1 \frac{\theta^{(k)}/4}{1/2 + \theta^{(k)}/4}, Y_2, Y_3, Y_4).$$

Then we obtain

$$\theta^{(k+1)} = \frac{E[X_2 + X_5|Y, \theta^{(k)}]}{E[X_2 + X_5 + X_3 + X_4|Y, \theta^{(k)}]} = \frac{Y_1 \frac{\theta^{(k)}/4}{1/2+\theta^{(k)}/4} + Y_4}{Y_1 \frac{\theta^{(k)}/4}{1/2+\theta^{(k)}/4} + Y_2 + Y_3 + Y_4}.$$

We start form $\theta^{(1)} = 0.5$. The following table gives the results from iterations:

k	$\theta^{(k+1)}$	$\theta^{(k+1)} - \theta^{(k)}$	$\frac{\theta^{(k+1)} - \hat{\theta}_n}{\theta^{(k)} - \hat{\theta}_n}$
0	.500000000	.126821498	.1465
1	.608247423	.018574075	.1346
2	.624321051	.002500447	.1330
3	.626488879	.000332619	.1328
4	.626777323	.000044176	.1328
5	.626815632	.000005866	.1328
6	.626820719	.000000779	
7	.626821395	.000000104	
8	.626821484	.000000014	

From the table, we find the EM converges and the result agrees with what is obtained form the Newton-Raphson iteration. We also note the the convergence is linear as $(\theta^{(k+1)} - \hat{\theta}_n)/(\theta^{(k)} - \hat{\theta}_n)$ becomes a constant when convergence; comparatively, the convergence in the Newton-Raphson iteration is quadratic in the sense $(\theta^{(k+1)} - \hat{\theta}_n)/(\theta^{(k)} - \hat{\theta}_n)^2$ becomes a constant when convergence. Thus, the Newton-Raphon iteration converges much faster than the EM algorithm; however, we have already seen the calculation of the EM is much less complex than the Newton-Raphson iteration and this is the advantage of using the EM algorithm.

Example 5.2 We consider the example of exponential mixture model. Suppose $Y \sim P_\theta$ where P_θ has density

$$p_\theta(y) = \left\{p\lambda e^{-\lambda y} + (1-p)\mu e^{-\mu y}\right\} I(y > 0)$$

and $\theta = (p, \lambda, \mu) \in (0,1) \times (0, \infty) \times (0, \infty)$. Consider estimation of θ based on $Y_1, ..., Y_n$ i.i.d $p_\theta(y)$. Solving the likelihood equation using the Newton-Raphson is much computation involved. We take an approach based on the EM algorithm. We introduce the complete data $X = (Y, \Delta) \sim p_\theta(x)$ where

$$p_\theta(x) = p_\theta(y, \delta) = (pye^{-\lambda y})^\delta ((1-p)\mu e^{-\mu y})^{1-\delta}.$$

This is natural from the following mechanism: Δ is a Bernoulli variable with $P(\Delta = 1) = p$ and we generate Y from $\text{Exp}(\lambda)$ if $\Delta = 1$ and from $\text{Exp}(\mu)$ if $\Delta = 0$. Thus, Δ is missing. The score equation for θ based on X is equal to

$$0 = \dot{l}_p(X_1, ..., X_n) = \sum_{i=1}^n \left\{ \frac{\Delta_i}{p} - \frac{1-\Delta_i}{1-p} \right\},$$

$$0 = \dot{l}_\lambda(X_1, ..., X_n) = \sum_{i=1}^n \Delta_i (\frac{1}{\lambda} - Y_i),$$

$$0 = \dot{l}_\mu(X_1, ..., X_n) = \sum_{i=1}^n (1-\Delta_i)(\frac{1}{\mu} - Y_i).$$

Thus, the M-step of the EM algorithm is to solve the following equations

$$0 = \sum_{i=1}^n E\left[\left\{ \frac{\Delta_i}{p} - \frac{1-\Delta_i}{1-p} \right\} |Y_1, ..., Y_n, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right] = \sum_{i=1}^n E\left[\left\{ \frac{\Delta_i}{p} - \frac{1-\Delta_i}{1-p} \right\} |Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right],$$

$$0 = \sum_{i=1}^n E\left[\Delta_i (\frac{1}{\lambda} - Y_i) |Y_1, ..., Y_n, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right] = \sum_{i=1}^n E\left[\Delta_i (\frac{1}{\lambda} - Y_i) |Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right],$$

$$0 = \sum_{i=1}^n E\left[(1-\Delta_i)(\frac{1}{\mu} - Y_i) |Y_1, ..., Y_n, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right] = \sum_{i=1}^n E\left[(1-\Delta_i)(\frac{1}{\mu} - Y_i) |Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)} \right].$$

This immediately gives

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^n E[\Delta_i | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}],$$

$$\lambda^{(k+1)} = \frac{\sum_{i=1}^n E[\Delta_i | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}{\sum_{i=1}^n Y_i E[\Delta_i | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]},$$

$$\mu^{(k+1)} = \frac{\sum_{i=1}^n E[(1-\Delta_i) | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}{\sum_{i=1}^n Y_i E[(1-\Delta_i) | Y_i, p^{(k)}, \lambda^{(k)}, \mu^{(k)}]}.$$

The conditional expectation

$$E[\Delta | Y, \theta] = \frac{p\lambda e^{-\lambda Y}}{p\lambda e^{-\lambda Y} + (1-p)\mu e^{-\mu Y}}.$$

As seen above, the EM algorithm facilitates the computation.

5.4.3 Information calculation in EM algorithm

We now consider the information of θ in the missing data. Denote $\dot{l}_c$ as the score function for θ in the full data and denote $\dot{l}_{mis|obs}$ as the score for θ in the conditional distribution of Y_{mis} given Y_{obs} and $\dot{l}_{obs}$ as the the score for θ in the distribution of Y_{obs}. Then it is clear that $\dot{l}_c = \dot{l}_{mis|obs} + \dot{l}_{obs}$. Using the formula

$$Var(U) = Var(E[U|V]) + E[Var(U|V)],$$

we obtain

$$Var(\dot{l}_c) = Var(E[\dot{l}_c|Y_{obs}]) + E[Var(\dot{l}_c|Y_{obs})].$$

Since

$$E[\dot{l}_c|Y_{obs}] = \dot{l}_{obs} + E[\dot{l}_{mis|obs}|Y_{obs}] = \dot{l}_{obs}$$

and

$$Var(\dot{l}_c|Y_{obs}) = Var(\dot{l}_{mis|obs}|Y_{obs}),$$

we obtain

$$Var(\dot{l}_c) = Var(\dot{l}_{obs}) + E[Var(\dot{l}_{mis|obs}|Y_{obs})].$$

Note that $Var(\dot{l}_c)$ is the information for θ based the complete data Y, denote by $I_c(\theta)$, $Var(\dot{l}_{obs})$ is the information for θ based on the observed data Y_{obs}, denote by $I_{obs}(\theta)$, and the $Var(\dot{l}_{mis|obs}|Y_{obs})$ is the conditional information for θ based on Y_{mis} given Y_{obs}, denoted by $I_{mis|obs}(\theta; Y_{obs})$. We obtain the following Louis formula

$$I_c(\theta) = I_{obs}(\theta) + E[I_{mis|obs}(\theta, Y_{obs})].$$

Thus, the complete information is the summation of the observed information and the missing information. One can even show when the EM converges, the convergence linear rate, denote as $(\theta^{(k+1)} - \hat{\theta}_n)/(\theta^{(k)} - \hat{\theta}_n)$ approximates the $1 - I_{obs}(\hat{\theta}_n)/I_c(\hat{\theta}_n)$.

The EM algorithms can be applied to not only missing data but also data with measurement error. Recently, the algorithms have been extended to the estimation in missing data in many semiparametric models.

5.5 Nonparametric Maximum Likelihood Estimation

In the previous section, we have studied the maximum likelihood estimation for parametric models. The maximum likelihood estimation can also be applied to many semiparametric or nonparametric models and this approach has been received more and more attention in recent years. We illustrate through some examples how such an estimation approach is used in the semiparametric or nonparametric model. Since obtaining the consistency and the asymptotic properties of the maximum likelihood estimators require both advanced probability theory in metric space and semiparametric efficiency theory, we would rather not get into details of these theories.

Example 5.3 Let $X_1, ..., X_n$ be i.i.d random variables with common distribution F, where F is any unknown distribution function. One may be interested in estimating F. This model is a nonparametric model. We consider maximizing the likelihood function to estimate F. The likelihood function for F is given by

$$L_n(F) = \prod_{i=1}^{n} f(X_i),$$

where $f(X_i)$ is the density function of F with respect to some dominating measure. However, the maximum of $L_n(F)$ does not exists since one can always choose a continuous f such that $f(X_1) \to \infty$. To avoid this problem, instead, we maximize an alternative function

$$\tilde{L}_n(F) = \prod_{i=1}^{n} F\{X_i\},$$

where $F\{X_i\}$ denotes the value $F(X_i)-F(X_i-)$. It is clear that $\tilde{L}_n(F) \le 1$ and if $\hat{F}_n$ maximizes $\tilde{L}_n(F)$, $\hat{F}_n$ must be a distribution function with point masses only at $X_1, ..., X_n$. We denote $q_i = F\{X_i\}$ and $q_i = q_j$ if $X_i = X_j$. Then maximizing $\tilde{L}_n(F)$ is equivalent to maximizing

$$\prod_{i=1}^{n} q_i \text{ subject to } \sum_{\text{distinct } q_i} q_i = 1.$$

The maximization with the Lagrange-Multiplier gives that

$$q_i = \frac{1}{n} \sum_{j=1}^{n} I(X_j = X_i).$$

Then

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_n \le x) = F_n(x).$$

In other words, the maximum likelihood estimator for F is the empirical distribution function F_n. It can be shown that F_n converges to F almost surely uniformly in x and $\sqrt{n}(F_n - F)$ converges in distribution to a Brownian bridge process. F_n is called the *nonparametric maximum likelihood estimator* of F.

Example 5.4 Suppose $X_1, ..., X_n$ are i.i.d F and $Y_1, ..., Y_n$ are i.i.d G. We observe i.i.d pairs $(Z_1, \Delta_1), ..., (Z_n, \Delta_n)$, where $Z_i = \min(X_i, Y_i)$ and $\Delta_i = I(X_i \le Y_i)$. We consider X_i as survival time and Y_i as censoring time. Then it is easy to calculate the joint distributions for (Z_i, Δ_i), $i = 1, ..., n$, is equal to

$$L_n(F, G) = \prod_{i=1}^{n} \{f(Z_i)(1 - G(Z_i))\}^{\Delta_i} \{(1 - F(Z_i))g(Z_i)\}^{1-\Delta_i}.$$

Similarly, $L_n(F, G)$ does not have the maximum so we consider an alternative function

$$\tilde{L}_n(F, G) = \prod_{i=1}^{n} \{F\{Z_i\}(1 - G(Z_i))\}^{\Delta_i} \{(1 - F(Z_i))G\{Z_i\}\}^{1-\Delta_i}.$$

$\tilde{L}_n(F, G) \leq 1$ and maximizing $\tilde{L}_n(F, G)$ is equivalent to maximizing

$$\prod_{i=1}^{n} \{p_i(1-Q_i)\}^{\Delta_i} \{q_i(1-P_i)\}^{1-\Delta_i},$$

subject to the constraint $\sum_i p_i = \sum_j q_j = 1$, where $p_i = F\{Z_i\}, q_i = G\{Z_i\}$, and $P_i = \sum_{Y_j \leq Y_i} p_j, Q_i = \sum_{Y_j \leq Y_i} q_j$. However, this maximization may not be easy. Instead, we will take a different approach by considering a new parameterization. Define the hazard functions $\lambda_X(t)$ and $\lambda_Y(t)$ as

$$\lambda_X(t) = f(t)/(1-F(t-)), \quad \lambda_Y(t) = g(t)/(1-G(t-))$$

and the cumulative hazard functions $\Lambda_X(t)$ and $\Lambda_Y(t)$ as

$$\Lambda_X(t) = \int_0^t \lambda_X(s)ds, \quad \Lambda_Y(t) = \int_0^t \lambda_Y(s)ds.$$

The derivation of F and G from Λ_X and Λ_Y is based on the following product-limit form:

$$1 - F(t) = \prod_{s \leq t}(1 - d\Lambda_X) \equiv \lim_{\max_{i=1}^m |t_i - t_{i-1}| \to 0} \prod_{0=t_0<t_1<\ldots<t_m=t} \{1 - (\Lambda_X(t_i) - \Lambda_X(t_{i-1}))\},$$

$$1 - G(t) = \prod_{s \leq t}(1 - d\Lambda_Y) \equiv \lim_{\max_{i=1}^m |t_i - t_{i-1}| \to 0} \prod_{0=t_0<t_1<\ldots<t_m=t} \{1 - (\Lambda_Y(t_i) - \Lambda_Y(t_{i-1}))\}.$$

Under the new parameterization, the likelihood function for $(Z_i, \Delta_i), i = 1, ..., n$, is given by

$$\prod_{i=1}^{n} \left[\lambda_X(Z_i)^{\Delta_i} \exp\{-\Lambda_X(Z_i)\} \lambda_Y(Z_i)^{1-\Delta_i} \exp\{-\Lambda_Y(Z_i)\}\right].$$

Again, we maximize a modified function

$$\prod_{i=1}^{n} \left[\Lambda_X\{Z_i\}^{\Delta_i} \exp\{-\Lambda_X(Z_i)\} \Lambda_Y\{Z_i\}^{1-\Delta_i} \exp\{-\Lambda_Y(Z_i)\}\right],$$

where $\Lambda_X\{Z_i\}$ and $\Lambda_Y\{Z_i\}$ are the jump sizes of Λ_X and Λ_Y at Z_i. The maximization becomes maximizing

$$\prod_{i=1}^{n} \left[a_i^{\Delta_i} \exp\{-A_i\} b_i^{1-\Delta_i} \exp\{-B_i\}\right],$$

where $A_i = \sum_{Z_j \leq Z_i} a_j$ and $B_i = \sum_{Z_j \leq Z_i} b_j$. Simple calculation gives that

$$a_i = \frac{\Delta_i}{R_i}, \quad b_i = \frac{(1-\Delta_i)}{R_i}, \quad R_i = \sum_{Y_j \geq Y_i} 1.$$

Thus, the NPMLE's for Λ_X and Λ_Y are given by

$$\hat{\Lambda}_X(t) = \sum_{Y_i \leq t} \frac{\Delta_i}{R_i}, \quad \hat{\Lambda}_Y(t) = \sum_{Y_i \leq t} \frac{1-\Delta_i}{R_i}.$$

As a result of the product-limit formula, we obtain the NPMLE's for F and G are

$$\hat{F}_n = 1 - \prod_{Y_i \le t} \left\{ 1 - \frac{\Delta_i}{R_i} \right\}, \quad \hat{G}_n = 1 - \prod_{Y_i \le t} \left\{ 1 - \frac{1-\Delta_i}{R_i} \right\}.$$

Both $1 - \hat{F}_n$ and $1 - \hat{G}_n$ are called the Kaplan-Meier estimates of the survival functions for the survival time and the censoring time respectively. The results based on counting process theory show that $\hat{F}_n$ and $\hat{G}_n$ are uniformly consistent and both $\sqrt{n}(\hat{F}_n - F)$ and $\sqrt{n}(\hat{G}_n - G)$ are asymptotically Gaussian.

Example 5.5 Suppose T is survival time and Z is covariate. Assume that the conditional distribution of T given Z has a conditional hazard function

$$\lambda(t|Z) = \lambda(t)e^{\theta' Z}.$$

Then the likelihood function from n i.i.d $(T_i, Z_i), i = 1, ..., n$ is given by

$$L_n(\theta, \Lambda) = \prod_{i=1}^n \left\{ \lambda(T_i) \exp\{-\Lambda(T_i)e^{\theta' Z_i}\} f(Z_i) \right\}.$$

Note $f(Z_i)$ is not informative about θ and λ so we can discard it from the likelihood function. Again, we replace $\lambda\{T_i\}$ by $\Lambda\{T_i\}$ and obtain a modified function

$$\tilde{L}_n(\theta, \Lambda) = \prod_{i=1}^n \left\{ \Lambda\{T_i\} \exp\{-\Lambda(T_i)e^{\theta' Z_i}\} \right\}.$$

Let $p_i = \Lambda\{T_i\}$ we maximize

$$\prod_{i=1}^n \left\{ p_i \exp\{-(\sum_{Y_j \le Y_i} p_j) e^{\theta' Z_i}\} \right\}$$

or its logarithm as

$$\sum_{i=1}^n \left\{ \theta' Z_i - \exp\{\theta' Z_i\} \sum_{Y_j \le Y_i} p_j + \log p_j \right\}.$$

We obtain

$$\hat{p}_i = \frac{1}{\sum_{Y_j \ge Y_i} \exp\{\theta' Z_j\}}$$

by differentiating with respect to p_i. After substituting it back into the $\log \tilde{L}_n(\theta, \Lambda)$, we find $\hat{\theta}_n$ maximizes the function

$$\log \left\{ \prod_{i=1}^n \frac{\exp\{\theta' Z_i\}}{\sum_{Y_j \ge Y_i} \exp\{\theta' Z_j\}} \right\}.$$

The function inside the logarithm is called the Cox's partial likelihood for θ. The consistency and the asymptotic efficiency for $\hat{\theta}_n$ have been well studied since the Cox (1972) proposed this estimation, with help from the martingale process theory.

Example 5.6 We consider $X_1, ..., X_n$ are i.i.d F and $Y_1, ..., Y_n$ are i.i.d G. We only observe (Y_i, Δ_i) where $\Delta_i = I(X_i \le Y_i)$ for $i = 1, ..., n$. This data is one type of interval censored data (or current status data). The likelihood for the observations is

$$\prod_{i=1}^n \left\{ F(Y_i)^{\Delta_i}(1 - F(Y_i))^{1-\Delta_i} g(Y_i) \right\}.$$

To derive the NPMLE for F and G, we instead maximize

$$\prod_{i=1}^n \left\{ P_i^{\Delta_i}(1 - P_i)^{1-\Delta_i} q_i \right\},$$

subject to the constraint that $\sum q_i = 1$ and $0 \le P_i \le 1$ increases with Y_i. Clearly, $\hat{q}_i = 1/n$ (suppose Y_i are all different). This constrained maximization turns out to be solved by the following steps:
(i) Plot the points $(i, \sum_{Y_j \le Y_i} \Delta_j), i = 1, ..., n$. This is called the cumulative sum diagram.
(ii) Form the $H^*(t)$, the greatest the convex minorant of the cumulative sum diagram.
(iii) Let $\hat{P}_i$ be the left derivative of H^* at i.
Then $(\hat{P}_1, ..., \hat{P}_n)$ maximizes the object function. Groeneboom and Wellner (1992) shows that if $f(t), g(t) > 0$,

$$n^{1/3}(\hat{F}_n(t) - F(t)) \to_d \left(\frac{F(t)(1 - F(t))f(t)}{2g(t)} \right)^{1/3} (2Z),$$

where Z is the location the maximum of the process $\{B(t) - t^2 : t \in R\}$ where $B(t)$ is standard Brownian motion starting from 0.

In summary, the NPMLE is a generalization of the maximum likelihood estimation to the semiparametric or nonparametric models. We have seen that in such a generalization, we often replace the functional parameter by an empirical function with jumps only at observed data and maximize a modified likelihood function. However, both computation of the NPMLE and the asymptotic property of the NPMLE can be difficult and vary for different specific problems.

5.6 Alternative Efficient Estimation

Although the maximum likelihood estimation is the most popular way of obtaining an asymptotically efficient estimator, there are alternative ways of deriving efficient estimation. Among them, one-step efficient estimation is the simplest.

In one-step efficient estimation, we assume that a strongly consistent estimator for parameter θ, denoted by $\tilde{\theta}_n$, is given. Moreover $|\tilde{\theta}_n - \theta_0| = O_p(n^{-1/2})$. One-step procedure is essentially a one-step Newton-Raphson iteration in solving the likelihood score equation; that is, we define

$$\hat{\theta}_n = \tilde{\theta}_n - \left\{ \ddot{l}_n(\tilde{\theta}_n) \right\}^{-1} \dot{l}_n(\tilde{\theta}_n),$$

where $\dot{l}_n(\theta)$ is the sore function of the observed log-likelihood function and $\ddot{l}_n(\theta)$ is the derivative of $\dot{l}_n(\theta)$. The next theorem shows that $\hat{\theta}_n$ is an asymptotically efficient estimator.

Theorem 5.6 Let $l_\theta(X)$ be the log-likelihood function of θ. Assume that there exists a neighborhood of θ_0 such that in this neighborhood, $|l_\theta^{(3)}(X)| \le F(X)$ with $E[F(X)] < \infty$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, I(\theta_0)^{-1}),$$

where $I(\theta_0)$ is the Fisher information. †

Proof Since $\tilde{\theta}_n \to_{a.s.} \theta_0$, we perform the Taylor expansion on the right-hand side of the one-step equation and obtain

$$\hat{\theta}_n = \tilde{\theta}_n - \left\{\ddot{l}_n(\tilde{\theta}_n)\right\}\left\{\dot{l}_n(\theta_0) + \ddot{l}_n(\theta^*)(\tilde{\theta}_n - \theta_0)\right\}$$

where θ^* is between $\tilde{\theta}_n$ and θ_0. Therefore,

$$\hat{\theta}_n - \theta_0 = \left[I - \left\{\ddot{l}_n(\tilde{\theta}_n)\right\}^{-1}\ddot{l}_n(\theta^*)\right](\tilde{\theta}_n - \theta_0) - \left\{\ddot{l}_n(\tilde{\theta}_n)\right\}\dot{l}_n(\theta_0).$$

On the other hand, by the condition that $|l_\theta^{(3)}(X)| \le F(X)$ with $E[F(X)] < \infty$, we know

$$\frac{1}{n}\ddot{l}_n(\theta^*) \to_{a.s.} E[\ddot{l}_{\theta_0}(X)], \quad \frac{1}{n}\ddot{l}_n(\tilde{\theta}_n) \to_{a.s.} E[\ddot{l}_{\theta_0}(X)].$$

Thus,

$$\hat{\theta}_n - \theta_0 = o_p(|\tilde{\theta}_n - \theta_0|) - \left\{E[\ddot{l}_{\theta_0}(X)] + o_p(1)\right\}^{-1}\frac{1}{n}\dot{l}_n(\theta_0)$$

so

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = o_p(1) - \left\{E[\ddot{l}_{\theta_0}(X)] + o_p(1)\right\}^{-1}\frac{1}{\sqrt{n}}\dot{l}_n(\theta_0) \to_d N(0, I(\theta_0)^{-1}).$$

We have proved that $\hat{\theta}_n$ is asymptotically efficient. †

Remark 5.1 Many different conditions from Theorem 5.6 can be used to ensure the asymptotic efficiency of $\hat{\theta}_n$ and here we have presented a simple one. Additionally, in the one-step estimation, since $\ddot{l}_n(\tilde{\theta}_n)$ approximates $-I(\theta_0)$ and the latter can be estimated by $-I(\tilde{\theta}_n)$, we sometimes use a slightly different one-step update:

$$\hat{\theta}_n = \tilde{\theta}_n + I(\tilde{\theta}_n)^{-1}\dot{l}(\tilde{\theta}_n).$$

One can recognize that this estimation is in fact one-step iteration in the Fisher scoring algorithm. Another efficient estimation arises from the Bayesian estimation method, where it can be shown that under regular condition of prior distribution, the posterior mode is equivalent to the maximum likelihood estimator. We will not pursue this method here.

In summary, efficient estimation is one of the most important goals in statistical inference. The maximum likelihood approach provides a natural and simple way of deriving an efficient estimator. However, when the maximum likelihood approach is not feasible, for example, the maximum likelihood estimator does not exist or the computation is difficult, other estimation approaches may be considered such as one-step estimation, Bayesian estimation etc. So far, we only focus on parametric models. When model is given semiparametrically or nonparametrically, the maximum likelihood estimator or the Bayesian estimator usually does not exist because of the presence of some infinite dimensional parameters. In this case, some approximated likelihood approaches have been developed, one of which is the nonparametric maximum likelihood approach (sometimes called empirical likelihood approach) as given in Section 5.5. Other approaches include partial likelihood approach, sieve likelihood approach, and penalized likelihood approach etc. These topics need another full text to describe and will be deferred to some future course.

READING MATERIALS: You should read Ferguson, Sections 16-20, Lehmann and Casella, Sections 6.2-6.7

PROBLEMS

We need the following definitions to answer the given problems.

Definition 5.2. $\{T_n\}$ and $\{\tilde{T}_n\}$ are two sequences of estimators for θ. Suppose

$$\sqrt{n}(T_n-\theta)\to_d N(0,\sigma^2),\quad \sqrt{n}(\tilde{T}_n-\theta)\to_d N(0,\tilde{\sigma}^2).$$

The asymptotic relative efficiency (ARE) of $\{T_n\}$ with respect to $\{\tilde{T}_n\}$ is defined as $r=\tilde{\sigma}^2/\sigma^2$. Intuitively, r can be understood as: to achieve the same accuracy in estimating θ, using the estimator T_n needs approximately $1/r$ times as many observations as using the estimator $\tilde{T}_n$. Thus, if $r>1$, T_n is more efficient than $\tilde{T}_n$; vice versa.

Definition 5.3. If δ_0 and δ_1 are statistics, then the random interval (δ_0,δ_1) is called a $(1-\alpha)$-*confidence interval* for $g(\theta)$ if

$$P_\theta(g(\theta)\in(\delta_0,\delta_1))\geq 1-\alpha.$$

Intuitively, the above inequality says: however data are generated, there is at least $(1-\alpha)$ probability that the interval contains the true value $g(\theta)$. Also, a random set $\mathcal{S}$ constructed from data is called a $(1-\alpha)$-*confidence region* for $g(\theta)$ if

$$P_\theta(g(\theta)\in\mathcal{S})\geq 1-\alpha.$$

If (δ_0,δ_1) and $\mathcal{S}$ change with sample size n and the above inequalities hold at the limit, then (δ_0,δ_1) and $\mathcal{S}$ are approximately $(1-\alpha)$-confidence interval and confidence region respectively.

1. Suppose that (X_1,Y_1),...,(X_n,Y_n) are i.i.d. with bivariate normal distribution $N_2(\mu,\Sigma)$ where $\mu=(\mu_1,\mu_2)'\in R^2$ and
$$\Sigma=\begin{pmatrix}\sigma^2 & \sigma\tau\rho\\ \sigma\tau\rho & \tau^2\end{pmatrix}$$
where $\sigma^2>0$, $\tau^2>0$, and $\rho\in(-1,1)$.

(a) If we assume that $\mu_1 = \mu_2 = \theta$ and Σ is known, what is the maximum likelihood estimator of θ?

(b) If we assume that μ is known and $\sigma^2 = \tau^2 = \theta$, what is the maximum likelihood estimator of (θ, ρ)?

(c) What is the asymptotic distribution of the estimator you found in (b)?

2. Let $X_1, ..., X_n$ be i.i.d. with common density

$$f_\theta(x) = \frac{\theta}{(1+x)^{\theta+1}} I(x > 0), \quad \theta > 0.$$

(a) Find the maximum likelihood estimator of θ, denoted as $\hat{\theta}_n$. Give the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$.

(b) Find a function g such that, regardless the value of θ, $\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to_d N(0, 1)$.

(c) Construct an approximately $1 - \alpha$ confidence interval based on (b).

3. Suppose X has a standard exponential distribution with density $f(x) = e^{-x} I(x > 0)$. Given $X = x$, Y has a Poisson distribution with mean λx.

(a) Determine the marginal mass function of Y. Find $E[Y]$ and $Var(Y)$ without using the mass function of Y.

(b) Give a lower bound for the variance of an unbiased estimator of λ based on X and Y.

(c) Suppose $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d., with each pair having the same joint distribution as X and Y. Let $\hat{\lambda}_n$ be the maximum likelihood estimator based on these data, and let $\tilde{\lambda}_n$ be the maximum likelihood estimator based on $Y_1, ..., Y_n$. Determine the asymptotic relative efficiency of $\tilde{\lambda}_n$ with respect to $\hat{\lambda}_n$.

4. Suppose that $X_1, ..., X_n$ are i.i.d. with density function $p_\theta(x)$, $\theta \in \Theta \subset R^k$. Denote $l_\theta(x) = \log p_\theta(x)$. Assume $l_\theta(x)$ is three times differentiable with respect to θ and its third derivatives are bounded by $M(x)$, where $\sup_\theta E_\theta[M(X)] < \infty$. Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ and assume $\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, I_\theta^{-1})$, where I_θ denotes the Fisher information at θ and is assumed to be non-singular.

(a) To estimate the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$, one proposes an estimator $\hat{I}_n^{-1}$, where

$$\hat{I}_n = -\frac{1}{n} \sum_{i=1}^n \ddot{l}_{\hat{\theta}_n}(X_i).$$

Prove that $\hat{I}_n^{-1}$ is a consistent estimator of I_θ^{-1}.

(b) Show

$$\sqrt{n}\hat{I}_n^{1/2}(\hat{\theta}_n - \theta) \to_d N(0, I_{k \times k}),$$

where $\hat{I}_n^{1/2}$ is the square root matrix of $\hat{I}_n$ and $I_{k\times k}$ is k-by-k identity matrix. From this approximation, construct an approximate $(1 - \alpha)$-confidence region for θ.

(c) Let $l_n(\theta) = \sum_{i=1}^n l_\theta(X_i)$. Perform Taylor expansion on $-2(l_n(\theta) - l_n(\hat\theta_n))$ (called likelihood ratio statistic) at $\hat\theta_n$ and show

$$-2(l_n(\theta) - l_n(\hat\theta_n)) \to_d \chi_k^2.$$

From this result, construct an approximate $1-\alpha$ confidence region for θ.

5. Human beings can be classified into one of four blood groups (phenotypes) O,A,B,AB. The inheritance of blood groups is controlled by three genes, O, A, B, of which O is recessive to A and B. If r, p, q are the gene probabilities in the population of O,A,B respectively $(r + p + q = 1)$, the probabilities of the six possible combinations (genotypes) in random mating (where two individuals draw at random from the population contribute one gene each) are shown in the following tables:

Phenotype	Genotype	probability
O	OO	r^2
A	AA	p^2
A	AO	$2rp$
B	BB	q^2
B	BO	$2rq$
AB	AB	$2pq$

We observe among N individuals that the phenotype frequencies N_O, N_A, N_B, N_{AB} and wish to estimate the gene probabilities from such data. A simple approach is to regard the observations as incomplete, the complete data set being the genotype frequencies N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}.

(a) Derive the EM algorithm for estimation of (p, q, r).

(b) Suppose that we observe $N_O = 176$, $N_A = 182$, $N_B = 60$, $N_{AB} = 17$. Use the EM algorithm to calculate the maximum likelihood estimator of (p, q, r), with starting value $p = q = r = 1/3$ and stopping iteration once the maximal difference between the new estimates and the previous one is less than 10^{-4}.

6. Suppose that X has a density function $f(x)$ and given $X = x$, $Y \sim N(\beta x, \sigma^2)$. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. observations with the same distribution as (X, Y). However, in many applications, not all X's are observable and we assume that $X_{m+1}, ..., X_n$ are missing for some $1 < m < n$ and that the missingness satisfies MAR assumption. Then the observed likelihood function is

$$\prod_{i=1}^m \left[f(X_i) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(Y_i - \beta X_i)^2}{2\sigma^2}\} \right] \times \prod_{i=m+1}^n \int_x \left[f(x) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(Y_i - \beta x)^2}{2\sigma^2}\} \right] dx.$$

Suppose that the observed values for X's are distinct. We want to calculate the NPMLE for β and σ^2. To do that, we "assume" that X only has point mass $p_i > 0$ at the observed data $X_i = x_i$ for $i = 1, ..., m$.

(a) Rewrite the likelihood function using β, σ^2 and $p_1, ..., p_m$.

(b) Write out the score equations for all the parameters.

(c) A simple approach to calculate the NPMLE is to use the EM algorithm, where $X_{m+1}, ..., X_n$ are missing data. Derive the EM algorithm. *Hint:* $X_i, i = m+1, ..., n$, can only have values $x_1, ..., x_m$ with probabilities $p_1, ..., p_m$.

7. Ferguson, pages 117-118, problems 1-3

8. Ferguson, pages 124-125, problems 1-7

9. Ferguson, page 131, problem 1

10. Ferguson, page 139, problems 1-4

11. Lehmann and Casella, pages 501-514, problems 3.1-7.34