# Chapter 3. Asymptotic Methods

## 1 Modes of Convergence of A Sequence of Random Variables

Due to the difficulty of making exact calculation, we make use of asymptotic results. For example, we experience the approximation of probabilities for computing significance levels and setting confidence. In this process, we use the following facts: Law of Large Numbers, Central Limit Theorem, and the Approximation of Binomial Distribution by Normal Distribution or Poisson Distribution and etc.. The essence of asymptotic methods is approximation. We approximate functions, random variables, probability distributions, means, variances, and covariance. However, we need to understand what kind of approximation we are using. The strong law of large numbers and the central limit theorem illustrate the two main types of limit theorems in probability.
**Strong limit theorems.** Given a sequence of functions $X_1(w), X_2(w), \ldots$ there is a limit function $X(w)$ such that $P(w : \lim_n X_n(w) = X(w)) = 1$.
**Weak limit theorems.** Given a sequence of functions $X_1(w), X_2(w), \ldots$ show that $\lim_n P(w : X_n(w) < x)$ exists for every $x$.

There is a great difference between strong and weak theorems which will become more apparent. A more dramatic example of this is: on $([0,1), \mathcal{B}_1([0,1)))$ with $P$ being Lebesgue measure, define

$$X_n(w) = \begin{cases} 0, & w < \frac{1}{2}, \\ 1, & \frac{1}{2} \le w < 1, \end{cases}$$

for $n$ even. For $n$ odd,

$$X_n(w) = \begin{cases} 1, & w < \frac{1}{2}, \\ 0, & \frac{1}{2} \le w < 1. \end{cases}$$

For all $n$, $P(w : X_n(w) < x) = P(w : X_1(w) < x)$. But for every $w \in [0,1)$

$$\limsup_n X_n(w) = 1, \quad \liminf_n X_n(w) = 0.$$

In this chapter, we will attempt to understand these asymptotic calculation.

### 1.1 The O, o Notation

Before the discussion of the concept of convergence for random variable, we will give a quick review of ways of comparing the magnitude of two sequences. A

notation that is especially useful for keeping track of the order of an approximation is the "big O, little o." Let $\{a_n\}$ and $\{\beta_n\}$ be two sequences of real numbers. We have the following three concept of comparison:

$a_n = O(\beta_n)$ if the ratio $a_n/\beta_n$ is bounded for large $n$, if there exists a number $K$ and an integer $n(K)$ such that if $n \geq K$, then $|a_n| < K|b_n|$.

$a_n = o(\beta_n)$ if the ratio $a_n/\beta_n$ converges to 0, as $n \to \infty$.

$a_n \sim \beta_n$ iff $a_n/\beta_n = c + o(1)$, $c \neq 0$, as $n \to \infty$.

**Fact**. (1) $O(a_n)O(\beta_n) = O(a_n\beta_n)$, (2) $O(a_n)o(\beta_n) = o(a_n\beta_n)$, (3) $o(a_n)o(\beta_n) = o(a_n\beta_n)$,
(4) $o(1) + O(n^{-1/2}) + O(n^{-1}) = o(1)$. The order of magnitude of a *finite* sum is the largest order of magnitude of the summands.
**Example**. Taylor expansion of a function $f(\cdot)$ about the value $c$ can be stated as

$$f(x) = f(c) + (x - c)f'(c) + o(|x - c|) \quad \text{as } x \to c.$$

In general,

**Theorem 1** *(Taylor). Let the function $f$ have a finite nth derivatives $f^{(n)}$ everywhere in the open interval $(a, b)$ and $(n-1)$th derivative $f^{(n-1)}$ continuous in the closed interval $[a, b]$. Let $x \in [a, b]$. For each point $y \in [a, b]$, $y \neq x$, there exists a point $z$ interior to the interval joining $x$ and $y$ such that*

$$f(y) = f(x) + \sum_{k=1}^{n-1} \frac{f^{(k)}(x)}{k!}(y - x)^k + \frac{f^{(n)}(z)}{n!}(y - x)^n.$$

*or*

$$f(y) = f(x) + \sum_{k=1}^{n-1} \frac{f^{(k)}(x)}{k!}(y - x)^k + o(|y - x|^{n-1}) \quad \text{as } y \to x.$$

### 1.2 Convergence of Stochastic Sequences

Now we consider probabilistic version of these *order of magnitude* relations. Let $A_n$ and $B_n$ be sequences of real random variables. Then

$A_n = O_p(B_n)$ iff for every $\epsilon > 0$, there exists a constant $M(\epsilon)$ and an integer $N(\epsilon)$ such that if $n \geq N(\epsilon)$, then

$$P\{|A_n/B_n| \leq M(\epsilon)\} \geq 1 - \epsilon.$$

$A_n = o_p(B_n)$ iff for every $\epsilon > 0$, $\lim_{n\to\infty} P\{|A_n/B_n| \leq \epsilon\} = 1$.

$A_n \approx B_n$ iff $A_n = B_n + o_p(B_n)$.

If $\mathbf{X}_n$ is a vector, we say that $\mathbf{X}_n = o_p(\beta_n)$ if $\|\mathbf{X}_n\| = o_p(\beta_n)$. Here $\|\mathbf{X}_n\|$ denotes the length of the vector $\mathbf{X}_n$.

Let $X_1, X_2, \ldots$ and $X$ be random variables on a probability space $(\Omega, \mathcal{A}, P)$. As an example, we take a measurement from an experiment in a laboratory. Usually the outcome of the experiment cannot be predicted with certainty. To handle this situation, we introduce the probability of $A$ (a collection of possible outcomes) to be the fraction of times that the outcome of the experiment results in $A$ in a large number of trials of the experiment. The set of all outcomes of an experiment are called *elementary events*. Here, $\Omega$ is the set of all elementary events which is also called the *sample space*. $\mathcal{A}$ is a class of subsets of $\Omega$ to which we can assign probability. For each set $A \in \mathcal{A}$ we assign a value $P(A)$ to be called the *probability* of $A$. Note that $P$ is a set function over the members of $\mathcal{A}$.

What kind of $\mathcal{A}$ would suffice our need? From our experience, four kinds of operations on sets, which are *intersection*, *complement*, *union* and *set difference*, are convenient and useful tools describing events. It is then quite natural to require that $\mathcal{A}$ contains the event formed by such operations. Such a class of sets is called a *Boolean* field. Based on the need, we also like to consider unions of all *countable* sequences of sets (events). We therefore require that $\mathcal{A}$ to be a *Borel field* or a $\sigma$-field. It means that it contains unions of all countable sequences of sets (and therefore countable intersections) and complementation.

For example, $\Omega$ can be a set of numbers or a subinterval of the real line. The context that is necessary for the strong limit theorems we want to prove is:

**Definition** A probability space consists of a triple $(\Omega, \mathcal{F}, P)$ where

(i) $\Omega$ is a space of points $w$, called the sample space and sample points. It is a nonempty set that represents the collection of all possible outcomes of an experiment.

(ii) $\mathcal{F}$ is a $\sigma$-field of subsets of $\Omega$. It includes the empty set as well as the set $\Omega$ and is closed under the set operations of complements and finite or countable unions and intersections. The elements of $\mathcal{F}$ are called *measurable events*, or simply *events*.

(iii) $P(\cdot)$ is a probability measure on $\mathcal{F}$; henceforth refer to $P$ as simply a probability. It is an assignment of probabilities to events in $\mathcal{F}$ that is subject to the conditions that

1. $0 \leq P(F) \leq 1$, for each $F \in \mathcal{F}$,

2. $P(\emptyset) = 0$, $P(\Omega) = 1$,
3. $P(\cup_i F_i) = \sum_i P(F_i)$ for any finite or countable sequence of *mutually exclusive* events $F_i$, $i = 1, 2, \ldots$, belonging to $\mathcal{F}$.

## 2 Remarks on measure and integration

A pair $(\Omega, \mathcal{F})$ consisting of a set $\Omega$ and a $\sigma$-field $\mathcal{F}$ of subsets of $\Omega$ is called a *measurable space*. For any given $\Omega$, there is one trivial $\sigma$-field which is the collection containing exactly two elements, empty set and $\Omega$. However, this field cannot be useful in applications.

Consider the set of real numbers $R$, which is uncountably infinite. We define the *Lebesgue measure* of intervals in $R$ to be their length. This definition and the properties of measure determine the Lebesgue measure of many, but not all, subsets of $R$. The collection of subsets of $R$ we consider, and for which Lebesgue measure is defined, is the collection of Borel sets defined below.
Let $\mathcal{C}$ be the collection of all finite open intervals on $R$. Then $\mathcal{B} = \sigma(\mathcal{C})$ is called the *Borel* $\sigma$-field. The elements of $\mathcal{B}$ are called *Borel sets*.

- All intervals (finite or infinite), open sets, and closed sets are Borel sets. These can be shown easily by the following.

  $$(a, \infty) = \cup_{n=1}^{\infty}(a, a+n), \quad (-\infty, a) = \cup_{n=1}^{\infty}(a-n, a), \quad [a, b] = ((-\infty, a) \cup (b, \infty))^c,$$
  $$[a, \infty) = \cup_{n=1}^{\infty}[a, a+n), \quad (-\infty, a] = \cup_{n=1}^{\infty}[a-n, a), \quad (a, b] = (-\infty, b] \cap (a, \infty),$$
  $$\{a\} = \cap\left(a - \frac{1}{n}, a + \frac{1}{n}\right).$$

  This means that every set containing countably infinitely many numbers is Borel; if $A = \{a_1, a_2, \ldots\}$, then

  $$A = \cup_{k=1}^{\infty}\{a_k\}.$$

  Hence the set of rational numbers is Borel, as is its complement, the set of irrational numbers. There are, however, sets which are not Borel. We have just seen that any non-Borel set must have uncountably many points.

- $\mathcal{B} = \sigma(\mathcal{O})$, where $\mathcal{O}$ is the collection of all open sets.

- The Borel $\sigma$-field $\mathcal{B}^k$ on the $k$-dimensional Enclidean space $R^k$ can be similarly defined.

- Let $C \subset R^k$ be a Borel set and let $\mathcal{B}_C = \{C \cap B : B \in \mathcal{B}^k\}$. Then $(C, \mathcal{B}_C)$ is a measurable space and $\mathcal{B}_C$ is called the Borel $\sigma$-field on $C$. (In statistics, it is quite often that we need to consider conditional probability and etc.)

The closure properties of $\mathcal{F}$ ensure that the usual applications of set operations in representing events do not lead to *nonmeasurable* events for which no (consistent) assignment of probability is possible.

The required *countable additivity* property (3) gives probabilities a sufficiently rich structure for doing calculations and approximations involving limits. Two immediate consequences of (3) are the following so-called *continuity properties*: if $A_1 \subset A_2 \subset \cdots$ is a *nondecreasing* sequence of events in $\mathcal{F}$ then, thinking of $\cup_{n=1}^{\infty} A_n$ as the *limiting event* for such sequences,

$$P(\cup_{n=1}^{\infty} A_n) = \lim_n P(A_n).$$

To prove this, disjointify $\{A_n\}$ by $B_n = A_n - A_{n-1}$, $n \geq 1$, $A_0 = \emptyset$, and apply (iii) to $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n$. By considering complements, one gets for decreasing measurable events $A_1 \supset A_2 \supset \cdots$ that

$$P(\cup_{n=1}^{\infty} A_n) = \lim_n P(A_n).$$

Example 1. Suppose that $\{X_t : 0 \leq t < \infty\}$ is a continuous-time Markov chains with a finite or countable state space $S$. The Markov property here refers to the property that the conditional distribution of the future, given past and present states of the process, does not depend on the past. The conditional probabilities $p_{ij}(s,t) = P(X_t = j | X_s = i)$, $0 \leq s < t$, are collectively referred to as the *transition probability law* for the process. In the case $p_{ij}(s,t)$ is a function of $t - s$, the transition law is called *time-homogeneous*, and we write $p_{ij}(s,t) = p_{ij}(t-s)$. Write $\mathbf{p}(t_0) = ((p_{ij}(t_0)))$, where $p_{ij}(t_0)$ gives the probability that the process will be in state $j$ at time $t_0$ if it is initially at state $i$. We assume that $\lim_{t \to 0} \mathbf{p}(t) = \mathbf{I}$, where $\mathbf{I}$ is the *identity matrix*. It means that with probability 1, the process spends a positive (but variable) amount of time in the initial state $i$ before moving to a different state $j$. Set

$$q_{ij} = \lim_{t \to 0} \frac{p_{ij}(t) - p_{ij}(0)}{t} = \lim_{t \to 0} \frac{p_{ij}(t) - \delta}{t}$$

which is being referred to as the *infinitesimal transition rates*. Write $\mathbf{Q} = ((q_{ij}))$, the *infinitesimal generator*.

Assume the Markov chain have the initial state $i$ and let $T_0 = \inf\{t > 0 : X_t \neq i\}$. An important question is finding the distribution of $T_0$.
Let $A$ denote the event that $\{T_0 > t\}$. Choose and fix $t > 0$. For each integer $n \geq 1$ define the finite-dimensional event

$$A_n = \{X_{(m/2^n)t} = i \text{ for } m = 0, 1, \ldots, 2^n\}.$$

The events $A_n$ are decreasing as $n$ increases and

$$\begin{aligned} A &= \lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n \\ &= \{X_u = i \text{ for all } u \text{ in } [0,t] \text{ which is a binary rational multiple of } t\} \\ &= \{T_0 > t\}. \end{aligned}$$

Since there is some $u$ of the form $u = (m/2^n)t \leq t$ in every nondegenerate interval, it follows that $T_0$ has an exponential distribution with parameter $-q_{ii}$.

### 2.0.1 The Homogeneous Poisson Process, the Poisson Distribution and the Exponential Distribution

In survival analysis, we are often interested in studying whether a particular event occurs or not. In this case, we can think in terms of death (denoted by state 1) and survive (denoted by state 0) using the language of Markov chain. We just describe a very special chain with two states and state 1 is an absorbing state. As an illustration, we now consider a simplest chain in which there are only two states, 0 and 1. Usually, we would like to know the sojourn time of staying at state 0. Denote the sojourn time of staying at state 0 by $T$. We know that

$$P(T < t + \delta | T \geq t) \approx \lambda(t)\delta,$$

where $\lambda(t)$ is the hazard function of $T$. Let $T_0$ denote a fix time and $\delta = T_0/n$ where $n \in \mathcal{N}$. Using Markov property, we have

$$\begin{aligned} P(T \geq T_0) &= P\left(T \geq (n-1)\frac{T_0}{\delta}\right) P\left(T \geq T_0 \middle| T \geq (n-1)\frac{T_0}{\delta}\right) \\ &= P\left(T \geq (n-2)\frac{T_0}{\delta}\right) P\left(T \geq (n-1)\frac{T_0}{\delta} \middle| T \geq (n-2)\frac{T_0}{\delta}\right) \\ &\quad \cdot P\left(T \geq T_0 \middle| T \geq (n-1)\frac{T_0}{\delta}\right). \end{aligned}$$

Continue in this fashion, we have

$$\begin{aligned} P(T \geq T_0) &\approx \prod_i \left[1 - \lambda\left(i\frac{T_0}{\delta}\right)\right] = \exp\left\{\sum_i \ln\left[1 - \lambda\left(i\frac{T_0}{\delta}\right)\right]\right\} \\ &\approx \exp\left\{\sum_i \left[1 - \lambda\left(i\frac{T_0}{\delta}\right)\right]\right\} \\ &\to \exp\left[-\int_0^{T_0} \lambda(t)dt\right]. \end{aligned}$$

This is the commonly seen form of survival function written in terms of the hazard function. If $\lambda(t) = \lambda_0$, $T$ is exponential distributed random variable.

Now we consider a different kind of chains with no absorbing states. This is usually seen in terms of *Poisson Processes* and *Queues*. The occurrences of

a sequence of discrete events can often be realistically modelled as a Poisson process. The defining characteristic of such a process is that the time intervals between successive events are exponentially distributed. Now it is still a multi-state Markov chain with no *absorbing* state. For the purpose of illustration, we describe discrete-time Markov chains. In such a chain, it is often being discussed in terms of *finite*, *aperiodic*, and *irreducible*. Finiteness means that there is a finite number of possible states. The aperiodicity assumption is that there is no state such that a return to, that state is possible only at $t_0, 2t_0, 3t_0, \ldots$ transitions later, where $t_0$ is an integer exceeding 1. If the transition matrix of a Markov chain with states $E_l, E_2, E_3, E_4$ is, for example,

$$P = \begin{bmatrix} 0 & 0 & 0.6 & 0.4 \\ 0 & 0 & 0.3 & 0.7 \\ 0.5 & 0.5 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \end{bmatrix},$$

then the Markov chain is periodic. If the Markovian random variable starts (at time 0) in $E_1$ then at time 1 it must be either in $E_3$ or $E_4$, at time 2 it must be in either $E_l$ or $E_2$, and in general it can visit only $E_l$ at times $2, 4, 6, \ldots$. It is therefore periodic. The irreducibility assumption implies that any state can eventually be reached from any other state, if not in one step then after several steps except for the case of Markov chains with absorbing states.

Now we come back to the chain associated with Poisson process. Given a sequence of discrete events occurring at times $t_0, t_1, t_2, t_3, \ldots$ the intervals between successive events are $\triangle t_1 = (t_1 - t_0)$, $\triangle t_2 = (t_2 - t_1)$, $\triangle t_3 = (t_3 - t_2), \ldots$, and so on. Assume the transition law is time-homogeneous. By the above argument, $\triangle t_i$ is again exponentially distributed. Due to the definition of Markov chain, these intervals $\triangle t_i$ are treated as independent random variables drawn from an exponentially distributed population, i.e., a population with the density function $f(x) = \lambda \exp(-\lambda x)$ for some fixed constant $\lambda$.

Now we state the fundamental properties that define a Poisson process, and from these properties we derive the Poisson distribution. Suppose that a sequence of random events occur during some time interval. These events form a homogeneous Poisson process if the following two conditions are met:

**(1)** The occurrence of any event in the time interval $(a, b)$ is independent of the occurrence of any event in the time interval $(c, d)$, while $(a, b)$ and $(c, d)$ do not overlap.

**(2)** There is a constant $\lambda > 0$ such that for any sufficiently small time interval,

$(t, t+h)$, $h > 0$, the probability that one event occurs in $(t, t+h)$, is independent of $t$, and is $\lambda h + o(h)$, and the probability that more than one event occurs in the interval $(t, t+h)$ is $o(h)$.

Condition 2 has two implications. The first is *time homogeneity*: The probability of an event in the time interval $(t, t+h)$ is independent of $t$. Second, this condition means that the probability of an event occurring in a small time interval is (up to a small order term) proportional to the length of the interval (with fixed proportionality constant $\lambda$). Thus the probability of no events in the interval $(t, t+h)$ is $1 - \lambda h + o(h)$, and the probability of one or more events in the interval $(t, t+h)$ is $\lambda h + o(h)$.

Various naturally occurring phenomena follow, or very nearly follow, these two conditions. Suppose a cellular protein degrades spontaneously, and the quantity of this protein in the cell is maintained at a constant level by the continual generation of new proteins at approximately the degradation rate. The number of proteins that degrade in any given time interval approximately satisfies conditions 1 and 2. The justification that condition 1 can be assumed in the model is that the number of proteins in the cell is essentially constant and that the spontaneous nature of the degradation process makes the independence assumption reasonable. Through *time division* and using Bernoulli random variable to indicate whether such an event occurs in $(t, t+h)$, Condition 2 also follows when $np$ is small, the probability of at least one success in $n$ Bernoulli trials is approximately $np$.

We now show that under conditions 1 and 2, the number $N$ of events that occur up to any arbitrary time $t$ has a Poisson distribution with parameter $\lambda t$. At time 0 the value of $N$ is necessarily 0, and at any later time $t$, the possible values of $N$ are $0, 1, 2, 3, \ldots$. We denote the probability that $N = j$ at any given time $t$ by $P_j(t)$. We would like to assess how $P_j(t)$ behaves as a function of $j$ and $t$.

The event that $N = 0$ at time $t+h$ occurs only if no events occur in $(0, t)$ and also no events occur in $(t, t+h)$. Thus for small $h$,

$$P_0(t+h) = P_0(t)(1 - \lambda h + o(h)) = P_0(t)(l - \lambda h) + o(h).$$

This equality follows from conditions 1 and 2.

The event that $N = 1$ at time $t+h$ can occur - in two ways. The first is that $N = 1$ at time $t$ and that no event occurs in the time interval $(t, t+h)$, the second is that $N = 0$ at time $t$ and that exactly one event occurs in the

time interval $(t, t+h)$. This gives

$$P_1(t+h) = P_0(t)(\lambda h) + P_1(t)(1-\lambda h) + o(h),$$

where the term $o(h)$ is the sum of two terms, both of which are $o(h)$. Finally, for $j = 2, 3, \ldots$ the event that $N = j$ at time $t+h$ can occur in three different ways. The first is that $N = j$ at time $t$ and that no event occurs in the time interval $(t, t+h)$. The second is that $N = j-1$ at time $t$ and that exactly one event occurs in $(t, t+h)$. The final possibility is that $N < j-2$ at time $t$ and that two or more events occur in $(t, t+h)$. Thus, for $j = 2, 3, \ldots$,

$$P_j(t+h) = P_{j-l}(t)(\lambda h) + P_j(t)(l-\lambda h) + o(h).$$

The above discussion leads to

$$\begin{aligned} \frac{P_0(t+h) - P_0(t)}{h} &= -\frac{P_0(t)(\lambda h) + o(h)}{h} \\ \frac{P_j(t+h) - P_j(t)}{h} &= -\frac{P_{j-1}(t)(\lambda h) - P_j(t(\lambda h)) + o(h)}{h}, \end{aligned}$$

$j = 1, 2, 3, \ldots$. Letting $h \to 0$, we get,

$$\frac{d}{dt}P_0(t) = -\lambda P_0(t),$$

and

$$\frac{d}{dt}P_j(t) = \lambda P_{j-l}(t) - \lambda P_j(t), \quad j = 1, 2, 3, \ldots.$$

The $P_j(t)$ are subject to the conditions

$$P_0(0) = 1, \quad P_j(0) = 0, \quad j = 1, 2, 3, \ldots.$$

The probability of the system still being in state 0 at time $t$, $P_0(t) = \exp(-\lambda t)$, which can be obtained easily. Note that $P_0(t) + P_1(t) = 1$. We could replace $P_0$ with $1 - P_1$ and write this as

$$\frac{1}{\lambda}\frac{dP_1(t)}{dt} + P_1(t) = 1.$$

From this

$$\frac{d}{dt}(P_1(t)\exp(\lambda t)) = \lambda.$$

We have

$$P_1(t) = e^{-\lambda t}\lambda t.$$

By induction, the probability of the $n$th state at time $t$ is

$$P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

This is the probability distribution for a simple Poisson *counting* process, representing the probability that exactly $n$ events will have occurred by the time $t$. Obviously the sum of these probabilities for $n = 1$ to $\infty$ equals 1, because the exponential $\exp(-\lambda t)$ factors out of the sum, and the sum of the remaining factors is just the power series expansion of $\exp(-\lambda t)$.

It's worth noting that since the distribution of intervals between successive occurrences is exponential, the Poisson distribution is stationary, meaning that any time can be taken as the initial time $t = 0$, which implies that the probability of $n$ occurrences in an interval of time depends only on the length of the interval, not on when the interval occurs. The expected number of occurrences by the time $t$ is given by the integral

$$E(n, t) = \sum_{i=0}^{\infty} iP_i(t) = \lambda t.$$

Since the distribution of the time between successive events is given by the exponential distribution. Thus the (random) time until the $k$th event occurs is the sum of $k$ independent exponentially distributed times. Let $t_0$ be some fixed value of $t$. Then if the time until the $k$th evenn occurs exceeds $t_0$, the number of events occurring before time $t_0$ is less than $k$, and conversely. This means that the probability that $k - 1$ or fewer-events occur before time $t_0$ must be identical to the probability that the time until the $k$th event occurs exceeds $t_0$. In other words it must be true that

$$e^{-\lambda t_0}\left(1 + (\lambda t_0) + \frac{(\lambda t_0)^2}{2!} + \cdots + \frac{(\lambda t_0)^{k-1}}{(k-1)!}\right) = \frac{\lambda^k}{\Gamma(k)}\int_{t_0}^{\infty} x^{k-1} exp(-\lambda x)dx.$$

This equation can also be established by repeated integration by parts of the right-hand side.

## 2.1 Counting measure and Lebesgue measure

First, we consider the counting measure in which $\Omega$ is a finite or countable set. Then probabilities are defined for all subsets $F$ of $\Omega$ once they are specified for singletons, so $\mathcal{F}$ is the collection of all subsets of $\Omega$. Thus, if $f$ is a *probability mass function* (p.m.f.) for singletons, i.e., $f(w) \geq 0$ for all $w \in \Omega$ and $\sum_w f(w) = 1$, then one may define $P(F) = \sum_{w \in F} f(w)$. The function $P$ so defined on the class of all subsets of $\Omega$ is *countably additive*, i.e., $P$ satisfies (3). So $(\Omega, \mathcal{F}, P)$ is easily seen to be a probability space. In this case the probability measure $P$ is determined by the probabilities of singletons $\{w\}$.

In the case $\Omega$ is not finite or countable, e.g., when $\Omega$ is the real line or the space of all infinite sequences of 0's and 1's, then the counting measure

formulation is no longer possible in general. We consider Lebesgue measure. The Lebesgue measure $\mu_0$ of a set containing only one point must be zero. In fact, since

$$\{a\} \subset \left(a - \frac{1}{n}, a + \frac{1}{n}\right)$$

for every positive integer $n$, we must have $\mu_0(\{a\}) = 0$. Hence, the Lebesgue measure of a set containing countably many points must also be zero. Instead, for example in the case $\Omega = R^1$, one is often given a piecewise continuous *probability density function* (p.d.f.) $f$, i.e., $f$ is nonnegative, integrable, and $\int_{-\infty}^{\infty} f(x)dx = 1$. For an interval $I = (a, b)$ or $(b, \infty)$, $-\infty \leq a < b \leq \infty$, one then assigns the probability $P(I) = \int_a^b f(x)dx$, by a Riemann integral. The Lebesgue measure of a set containing uncountably many points can be either zero, positive and finite, or infinite. We may not compute the Lebesgue measure of an uncountable set by adding up the Lebesgue measure of its individual members, because there is no way to add up uncountably many numbers.

This set function $P$ may be extended to the class $\mathcal{C}$ comprising all finite unions $F = \cup_j I_j$ of pairwise disjoint intervals $I_j$ by setting $P(F) = \sum_j P(I_j)$. The class $\mathcal{C}$ is a *field*, i.e., the empty set and $\Omega$ belong to $\mathcal{C}$ and it is closed under complements and finite intersection (and therefore finite unions). But, since $\mathcal{C}$ is not a $\sigma$ field, usual sequentially applied operations on events may lead to events outside of $\mathcal{C}$ for which probabilities have not been defined. But a theorem from measure theory, the *Caratheodory Extension Theorem*, asserts that there is a unique countably additive extension of $P$ from a field $\mathcal{C}$ to the smallest $\sigma$ field that contains $\mathcal{C}$. In the case of $\mathcal{C}$ above, this $\sigma$ field is called the *Borel* $\sigma$ field $\mathcal{B}^1$ on $R^1$ and its sets are called *Borel sets* of $R^1$.

In general, such an extension of $P$ to the *power set* $\sigma$-field, that is the collection of all subsets of $R^1$, is not possible. The same considerations apply to all *measures* (i.e., countably additive nonnegative set functions $\mu$ defined on a $\sigma$-field with $\mu(O) = 0$), whether the measure of $\Omega$ is 1 or not. The measure $\mu_0 = m$, which is defined first for each interval $I$ and the *length* of the interval, and then extend uniquely to $\mathcal{B}^1$, is called *Lebesgue measure* on $R^1$. Similarly, one defines the *Lebesgue measure* on $R^k$ ($k \geq 2$) whose Borel $\sigma$-field $\mathcal{B}^k$ is the smallest $\sigma$ field that contains all $k$-dimensional rectangles $I = I_1 \times I_2 \times \cdots \times I_k$, with $I_j$ a one-dimensional rectangle (interval) of the previous type. The Lebesgue measure of a rectangle is the product of the lengths of its sides, i.e., its volume. Lebesgue measure on $R^k$ has the property that the space can be decomposed into a countable union measurable sets of finite Lebesgue measure; such measures are said to be *sigma-finite*. All measures

referred to in this note are sigma-finite.

### 2.2 Extension

A finitely additive measure $\mu$ on a field F is a real-valued (including $\infty$), non-negative function with domain F such that for $A, B \in \mathcal{F}$, $A \cap B = \emptyset$,

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

The extension problem for measures is: Given a finitely additive measure $\mu_0$ on a field $\mathcal{F}_0$, when does there exist a measure $\mu$ on $\mathcal{F}(\mathcal{F}_0)$ agreeing with $\mu_0$ on $\mathcal{F}_0$? A measure has certain continuity properties:

**Theorem 2** *Let $\mu$ be a measure on the $\sigma$-field F. If $A_n \downarrow A$, $A_n \in \mathcal{F}$, and if $\mu(A_n) < \infty$ for some $n$, then*

$$\lim_n \mu(A_n) = \mu(A).$$

*Also, if $A_n \uparrow A$, $A_n \in \mathcal{F}$, then*

$$\lim_n \mu(A_n) = \mu(A).$$

This is called continuity from above and below. Certainly, if $\mu_0$ is to be extended, then the minimum requirement needed is that $\mu_0$ be continuous on its domain. Call $\mu_0$ *continuous from above at* $\emptyset$ if whenever $A_n \in \mathcal{F}_0$, $A_n \downarrow \emptyset$, and $\mu_0(A_n) < \infty$ for some $n$, then

$$\lim_n \mu_0(A_n) = 0.$$

Consider the example that

$$A_1 = [1, \infty), A_2 = [2, \infty), A_3 = [3, \infty), \ldots.$$

Then $\cap_{k=1}^{\infty} A_k = \emptyset$, so $\mu(\cap_{k=1}^{\infty} A_k) = 0$, but $\lim_{n\to\infty} \mu(A_n) = \infty$.

**Caratheodory Extension Theorem.** If $\mu_0$ on $\mathcal{F}_0$ is continuous from above at $\emptyset$, then there is a unique measure $\mu$ on $\mathcal{F}(\mathcal{F}_0)$ agreeing with $\mu_0$ on $\mathcal{F}_0$ (see Halmos, p. 54).

The extension of a measure $\mu$ from a field $\mathcal{C}$, as provided by the Caratheodory Extension Theorem stated above, is *unique* and may be expressed by the formula

$$\mu(F) = \inf \sum_n \mu(C_n), \quad (F \in \mathcal{F}),$$

where the summation is over a finite collection $C_1, C_2, \ldots$ of sets in $\mathcal{C}$ whose union contains $F$ and the infimum is over all such collections.

As suggested by the construction of measures on $\mathcal{B}^k$ outlined above, starting from their specifications on a class of rectangles, if two measures $\mu_1$ and $\mu_2$ on a sigmafield $\mathcal{F}$ agree on subclass $\mathcal{A} \subset \mathcal{F}$ closed under finite intersections and $\Omega \in \mathcal{A}$, then they agree on the *smallest sigmafield*, denoted $\sigma(\mathcal{A})$, that contains $\mathcal{A}$. The sigmafield $\sigma(\mathcal{A})$ is called the *$\sigma$-field generated by $\mathcal{A}$*. On a metric space $S$ the $\sigma$-field $\mathcal{B} = \mathcal{B}(S)$ generated by the class of all open sets is called the *Borel $\sigma$ field*.

### 2.3 Lebesgue integral

An indicator function $g$ from $R$ to $R$ is a function which takes only the values 0 and 1. We call

$$A = \{x \in R; g(x) = 1\}$$

the set indicated by $g$. We define the Lebesgue integral of $g$ to be

$$\int_R g d\mu = \mu_0(A).$$

A simple function $h$ from $R$ to $R$ is a linear combination of indicators, i.e., a function of the form $h(x) = \sum_{k=1}^n c_k g_k(x)$, where each $g_k$ is of the form

$$g_k(x) = \begin{cases} 1 & \text{if } x \in A_k \\ 0 & \text{if } x \notin A_k \end{cases}$$

and each $c_k$ is a real number. We define the Lebesgue integral of $h$ to be $\sum_{k=1}^n c_k \mu(A_k)$. Let $f$ be a nonnegative function defined on $R$, possibly taking the value $\infty$ at some points. We define the *Lebesgue integral* of $f$ to be

$$\int_R f d\mu_0 = \sup \left\{ \int_R h d\mu_0; h \text{ is simple and } h(x) \leq f(x) \text{ for every } x \in R \right\}.$$

It is possible that this integral is infinite. If it is finite, we say that $f$ is *integrable*.

Finally, let $f$ be a function defined on $R$, possibly taking the value $\infty$ at some points and the value $-\infty$ at other points. We define the *positive* and *negative parts* of $f$ to be

$$f^+(x) = \max\{f(x), 0\}, \quad f^-(x) = \max\{-f(x), 0\},$$

respectively, and we define the Lebesgue integral of $f$ to be

$$\int_R f d\mu_0 = \int_R f^+ d\mu_0 - \int_R f^- d\mu_0,$$

provided the right-hand side is not of the form $\infty - \infty$. If both $\int_R f^+ d\mu_0$ and $\int_R f^- d\mu_)$ are finite (or equivalently, $\int_R |f| d\mu_0 < \infty$, since $|f| = f^+ + f^-$), we say that $f$ is *integrable*.

Let $f$ be a function defined on $R$, possibly taking the value $\infty$ at some points and the value $-\infty$ at other points. Let $A$ be a subset of $R$. We define $\int_A f d\mu_0 = \int_R 1_A f d\mu_0$.

The Lebesgue integral just defined is related to the Riemann integral in one very important way: if the Riemann integral $\int_a^b f(x)dx$ is defined, then the Lebesgue integral $\int_{[a,b]} f d\mu_0$ agrees with the Riemann integral. The Lebesgue integral has two important advantages over the Riemann integral. The first is that the Lebesgue integral is defined for more functions, as we show in the following examples.

Example 2. Let $Q$ be the set of rational numbers in $[0,1]$ and consider $f = 1_Q$. Being a countable set, $Q$ has Lebesgue measure zero, and so the Lebesgue integral of $f$ over $[0,1]$ is $\int_{[0,1]} f d\mu_0 = 0$. To compute the Riemann integral $\int_0^1 f(x)dx$, we choose partition points $0 = x_0 < x_1 < \cdots < x_n = 1$ and and divide the interval $[0,1]$ into subintervals $[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]$. In each subinterval $[x_{k-1}, x_k]$ there is a rational point $q_k$, where $f(q_k) = 1$, and also an irrational point $r_k$, where $f(r_k) = 0$. We approximate the Riemann integral from above by the upper sum 1 and also approximate it from below by the lower sum 0. No matter how fine we take the partition of $[0,1]$, the upper sum is always 1 and the lower sum is always 0. Since these two do not converge to a common value as the partition becomes finer, the Riemann integral is not defined.

Example 3. Consider the function

$$f(x) = \begin{cases} \infty, & \text{if } x = 0, \\ 0, & \text{if } x \neq 0. \end{cases}$$

Every simple function which lies between 0 and $f$ is of the form

$$h(x) = \begin{cases} y, & \text{if } x = 0, \\ 0, & \text{if } x \neq 0. \end{cases}$$

for some $y \in [0, \infty)$, and thus has Lebesgue integral

$$\int_R h d\mu_0 = y\mu(\{0\}).$$

It follows that $\int_R f d\mu_0 = 0$. Now consider the Riemann integral $\int_{-\infty}^{\infty} f(x)dx$, which for this function $f$ is the same as the Riemann integral $\int_{-1}^{1} f(x)dx$. When we partition $[-1,1]$ into subintervals, one of these will contain the point 0, and when we compute the upper approximating sum for $\int_{-1}^{1} f(x)dx$, this point will contribute $\infty$ times the length of the subinterval containing it. Thus the upper approximating sum is $\infty$. On the other hand, the lower approximating sum is 0, and again the Riemann integral does not exist.

The Lebesgue integral has all linearity and comparison properties one would expect of an integral. In particular, for any two functions $f$ and $g$ and any real constant $c$,

$$\begin{aligned}
\int_R (f+g)d\mu_0 &= \int_R f d\mu_0 + \int_R g d\mu_0, \\
\int_R cf d\mu_0 &= c\int_R f d\mu_0, \\
\int_R f d\mu_0 &\leq \int_R g d\mu_0, \text{ when } f(x) \leq g(x) \\
\int_{A\cup B} f d\mu_0 &= \int_A f d\mu_0 + \int_B f d\mu_0.
\end{aligned}$$

There are three *convergence* theorems satisfied by the Lebesgue integral. In each of these the situation is that there is a sequence of functions $f_n$, $n = 1, 2, \ldots$ converging pointwise to a limiting function $f$. *Pointwise convergence* just means that

$$\lim_{n\to\infty} f_n(x) = f(x) \text{ for every } x \in R.$$

There are no such theorems for the Riemann integral, because the Riemann integral of the limiting function $f$ is too often not defined. Before we state the theorems, we given two examples of pointwise convergence which arise in probability theory.

Example 4. Consider a sequence of normal densities, each with variance 1 and the $n$-th having mean $n$:

$$f_n(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-n)^2}{2}\right).$$

These converge pointwise to the zero function. We have $\int_R f_n d\mu_0 = 1$ for every $n$ but $\int_R f d\mu_) = 0$.

Example 5. Consider a sequence of normal densities, each with mean 0 and the $n$-th having variance $1/n$:

$$f_n(x) = \frac{n}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2n^{-1}}\right).$$

These converge pointwise to the function

$$f(x) = \begin{cases} \infty, & \text{if } x = 0, \\ 0, & \text{if } x \neq 0. \end{cases}$$

We have $\int_R f_n d\mu_0 = 1$ for every $n$ but $\int_R f d\mu_0 = 0$.

**Theorem 3** *(Fatous Lemma) Let $f_n$, $n = 1, 2, \ldots$ be a sequence of nonnegative functions converging pointwise to a function $f$. Then*

$$\int_R f d\mu_0 \leq \liminf_{n\to\infty} \int_R f_n d\mu_0.$$

The key assumption in Fatou's Lemma is that all the functions take only nonnegative values. Fatou's Lemma does not assume much but it is is not very satisfying because it does not conclude that

$$\int_R f d\mu_0 = \lim_{n\to\infty} \int_R f_n d\mu_0,$$

There are two sets of assumptions which permit this stronger conclusion.

**Theorem 4** *(Monotone Convergence Theorem) Let $f_n$, $n = 1, 2, \ldots$ be a sequence of functions converging pointwise to a function $f$. Assume that*

$$0 \le f_1(x) \le f_2(x) \le \cdots \text{ for every } x \in R.$$

*Then*

$$\int_R f d\mu_0 = \lim_{n\to\infty} \int_R f_n d\mu_0,$$

*where both sides are allowed to be $\infty$.*

**Theorem 5** *(Dominated Convergence Theorem) Let $f_n$, $n = 1, 2, \ldots$ be a sequence of functions converging pointwise to a function $f$. Assume that there is a nonnegative integrable function $g$ (i.e., $\int_R g d\mu_0 < \infty$) such that*

$$|f_n(x)| \le g(x) \text{ for every } x \in R \text{ for every } n.$$

*Then*

$$\int_R f d\mu_0 = \lim_{n\to\infty} \int_R f_n d\mu_0,$$

*and both sides will be finite.*

### 2.4 Related results in probability theory

**Theorem 6** *(Bounded Convergence Theorem) Suppose that $X_n$ converges to $X$ in probability and that there exists a constant $M$ such that $P(|X_n| \le M) = 1$. Then $E(X_n) \to E(X)$.*

**Proof**. Let $\{x_i\}$ be a partition of $R$ such that $F_X$ is continuous at each $x_i$. Then

$$\sum_i x_i P\{x_i < X_n \le x_{i+1}\} \le E(X_n) \le \sum_i x_{i+1} P\{x_i < X_n \le x_{i+1}\}$$

and taking limits we have

$$\begin{aligned} \sum_i x_i P\{x_i < X_n \le x_{i+1}\} &\le \underline{\lim} E(X_n) \\ &\le \overline{\lim} E(X_n) \le \sum_i x_{i+1} P\{x_i < X_n \le x_{i+1}\}. \end{aligned}$$

As $\max |x_{i+1} - x_i| \to 0$, the left and right sides converges to $E(X)$ giving the theorem.

**Theorem 7** *(Monotone Convergence Theorem) Suppose* $0 \leq X_n \leq X$ *and* $X_n$ *converges to* $X$ *in probability. Then* $\lim_{n\to\infty} E(X_n) = E(X)$.

**Proof**. For $M > 0$

$$E(X) \geq E(X_n) \geq E(X_n \wedge M) \to E(X \wedge M)$$

where the convergence on the right follows from the bounded convergence theorem. It follows that

$$E(X \bigwedge M) \leq \liminf_{n\to} E(X_n) \leq \limsup_{n\to} E(X_n) \leq E(X).$$

**Theorem 8** *(Dominated Convergence Theorem) Assume* $X_n$ *and* $Y_n$ *converge to* $X$ *and* $Y$*, respectively, in probability. Also,* $|X_n| \leq Y_n$ *and* $E(Y_n) \to E(Y) < \infty$*. Then* $\lim_{n\to\infty} E(X_n) = E(X)$.

Its proof follows from Fatou Lemma.

# 3 Mode of Convergence

On $\Omega$ there is defined a sequence of real-valued functions $X_1(w), X_2(w), \ldots$ which are random variables in the sense of the following definition.
**Definition** A function $X(w)$ defined on $\Omega$ is called a random variable if for every Borel set $B$ in the real line $R$, the set is $\{w : X(w) \in B\}$ is in $\mathcal{F}$. ($X(w)$ is a measurable function on $(\Omega, \mathcal{F})$. )

## 3.1 Convergence in Distribution

Suppose we flip a fair coin 400 times and want to find out the probability of getting heads between 190 and 210. A standard practice is to invoke the *Central Limit Theorem* to get an approximation of the above probability. Let $S_{400}$ denote the number of heads in the 400 flips. For this particular problem, our major concern is $P(190 \leq S_{400} \leq 210)$ or whether this probability can be approximated well by $P(-1.05 \leq Z \leq 1.05)$. Here $Z$ is a standard normal random variable. In this example, we need the concept of *converges in distribution*. Consider distribution functions $F_1(\cdot), F_2(\cdot), \ldots$ and $F(\cdot)$. Let $X_1, X_2, \ldots$ and $X$ denote random variables (not necessarily on a common probability space) having these distributions, respectively. We say that $X_n$ *converges in distribution* (or *in law*) to $X$ if

$$\lim_{n\to\infty} F_n(t) = F(t), \qquad \text{for all } t \text{ which are continuity points of } F.$$

This is written $X_n \xrightarrow{d} X$ or $X_n \xrightarrow{\mathcal{L}} X$ or $F_n \xrightarrow{w} F$. What are convergent here are not the values of the random variables themselves, but the probabilities with which the random variables assume certain values. If $X_n \xrightarrow{d} X$, then the distribution of $X_n$ can be well approximated for large $n$ by the distribution of $X$. This observation is extremely useful since $F_X$ is often easier to compute than $F_{X_n}$.

In general, we would like to say that the distribution of the random variables $X_n$ converes to the distribution of $X$ if $F_n(x) = P(X_n < x) \to F(x) = P(X < x)$ for every $x \in R$. But this is a bit too strong. We now use an example to illustrate why we require the convergence only occurs at the continuity points of $F$? Consider random variables $X_n$ which take values $1 - n^{-1}$ or $1 + n^{-1}$ with probabilities $1/2$. Heuristically, we would want the values of $X_n$ to be more and more concentrated about 1. Note that the distribution of $X_n$ is

$$F_n(x) = \begin{cases} 0, & x < 1 - n^{-1} \\ 1/2, & 1 - n^{-1} \le x < 1 + n^{-1} \\ 1, & x \ge 1 + n^{-1}. \end{cases}$$

By calculation, we have $F_n(x) \to F^*(x)$ as $n \to \infty$ where

$$F^*(x) = \begin{cases} 0, & x < 1 \\ 1/2, & x = 1 \\ 1, & 1 < x. \end{cases}$$

On the other hand, for the random variable $X$ taking value 1 with probability 1. The distribution of $X$ is

$$F(x) = \begin{cases} 0, & x < 1 \\ 1, & x \ge 1. \end{cases}$$

Apparently, not much should be assumed about what happens for $x$ at a discontinuity point of $F(x)$. Therefore, we can only consider *convergence in distribution* at continuity points of $F$. Read Example 14.3-2(pp467) of Bishop, Feinberg and Holland (1975) for direct verification that $F_n \xrightarrow{w} F$. Another important tool for establishing *convergence in distribution* is to use moment-generating function or characteristic function. Read Example 14.3-3(pp467) of Bishop, Feinberg and Holland (1975). In later section, we will use this tool to prove the central limit theorem (Chung[1974], Theorem 6.4.4).

When we talk about *convergence in distribution*, $w$ never come into the picture. As an example, flip a fair coin once. Let $X = 1$ if we get head and $X = 0$, otherwise. On the other hand, set $Y = 1 - X$. It is obvious that $X$

and $Y$ have the same distribution. As a remark, the random variable $X$ is a function of $w$ but we can never observe $w$.

### 3.2 Convergence with Probability 1

Next, we discuss *convergence with probability* 1 (or strongly, almost surely, almost everywhere, etc.) which is closely related to the convergence of sequences of functions in advanced calculus. This criterion of convergence is of particular importance in the probability limit theorems known as the *laws of large numbers*. This is defined in terms of the entire sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$. Regarding such a sequence as a new random variable with realized value $x_1, x_2, \ldots, x_n, \ldots$, we may say that this realized sequence either does or does not converge in the ordinary sense to a limit $x$. If the probability that it does so is unity, then we say that $X_n \to X$ almost certainly. Consider random variables $X_1, X_2, \cdots$ and $X$, we say that $X_n$ *converges with probability* 1 (or almost surely) to $X$ if

$$P(w : \lim_{n\to\infty} X_n(w) = X(w)) = 1.$$

This is written $X_n \overset{wp1}{\to} X$, $n \to \infty$. To be better understanding this convergence, we give the following equivalent condition:

$$\lim_{n\to\infty} P(|X_m - X_n| < \epsilon, \text{for all } m \geq n) = 1, \qquad \text{for every } \epsilon > 0.$$

Suppose we have to deal with questions of convergence when no limit is in evidence. For convergence almost surely, this is immediately reducible to the numerical case where the Cauchy criterion is applicable. Specifically, $\{X_n\}$ converges a.s. if and only if there exists a null set $\mathcal{N}$ such that for every $w \in \Omega - \mathcal{N}$ and every $\epsilon > 0$, there exists $m(w, \epsilon)$ such that

$$n' > n \geq m(w, \epsilon) \to |X_n(w) - X_{n'}(w)| \leq \epsilon.$$

Or, for any positive $\epsilon$ and $\eta$, there is an $n_0$ such that

$$P\{|X_n - X_m| > \epsilon \text{ for at least one } m \geq n\} < \eta$$

for all $n \geq n_0$. As almost surely convergence depends on the *simultaneous* behavior of $X_n$ for all $n \geq n_0$, it is obviously more difficult to handle, but the following sufficient criterion is useful. If $\sum_{n=1}^{\infty} E\{|X_n - X|^p\} < \infty$ for some $p > 0$, then $X_n \to X$ almost surely. This criterion follows from the observation:

$$\begin{aligned} P(|X_m - X| > \epsilon \text{ for some } m \geq n) &= P\left(\cup_{m=n}^{\infty}\{|X_m - X| > \epsilon\}\right) \\ &\leq \sum_{m=n}^{\infty} P(|X_m - X| > \epsilon). \end{aligned}$$

### 3.2.1 Consistency of the Empirical Distribution Function

Let $X_1, \ldots, X_n$ be independent identically distributed random variables on $\mathcal{R}$ with distribution function $F(x) = P(X \leq x)$. The nonparametric maximum-likelihood estimate of $F$ is the *sample distribution function* or *empirical distribution function* defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i,\infty)}(x).$$

Thus, $F_n(x)$ is the proportion of the observations that fall less than or equal to $x$. For each fixed $x$, the strong law of large numbers implies that $F_n(x) \xrightarrow{a.s.} F(x)$, because we may consider $I_{[X_i,\infty)}(x)$ as i.i.d. random variables with mean $F(x)$. Thus, $F_n(x)$ is a strongly consistent estimate of $F(x)$ for every $x$.

The following corollary improves on this observation in two ways. First, the set of probability one on which convergence takes place may be chosen to be independent of $x$. Second, the convergence is uniform in $x$. This assertion, that the empirical distribution function converges uniformly almost surely to the true distribution function, is known as the Glivenko-Cantelli Theorem.

**COROLLARY.** $P\{\sup_x |F_n(x) - F(x)| \to 0\} = 1$.

**Proof.** Let $\varepsilon > 0$. Find an integer $k > 1/\varepsilon$ and numbers $-\infty = x_0 < x_1 \leq x_2 \leq \cdots \leq x_{k-1} < x_k = \infty$, such that

$$F(x_j^-) \leq j/k \leq F(x_j)$$

for $j = 1, \ldots, k-1$. [$F(x_j^-)$ may be considered notation for $P(X < x_j)$.] Note that if $x_{j-1} < x_j$ then $F(x_j^-) - F(x_{j-1}) \leq \varepsilon$. From the strong law of large numbers, $F_n(x_j) \xrightarrow{a.s.} F(x_j)$ and $F_n(x_j^-) \xrightarrow{a.s.} F(x_j^-)$ for $j = 1, \ldots, k-1$. Hence,

$$\triangle_n = \max(|F_n(x_j) - F(x_j)|, |F_n(x_j^-) - F(x_j^-)|, j = 1, \ldots, k-1) \xrightarrow{a.s.} 0.$$

Let $x$ be arbitrary and find $j$ such that $x_{j-1} < x \leq x_j$. Then,

$$F_n(x) - F(x) \leq F_n(x_j^-) - F(x_{j-1}) \leq F_n(x_j^-) - F(x_j^-) + \epsilon,$$

and

$$F_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_j^-) \geq F_n(xj-1) - F(x_{j-1}) - \epsilon.$$

This implies that

$$\sup_x |F_n(x) - F(x)| \leq \triangle_n + \epsilon \xrightarrow{a.s.} \epsilon.$$

Since this holds for all $\epsilon > 0$, the corollary follows.

### 3.2.2 Law of Large Numbers

The weak (strong) law of large numbers states the sample mean is a weakly (strongly) consistent estimate of the population mean. The weak law of large numbers says that if $X_1, \ldots, X_n$ are i.i.d. random variables with finite first moment, $\mu$, then for every $\epsilon > 0$ we have

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0$$

as $n \to \infty$. The argument of using Chebyschev inequality with finite second moment shows that

$$P(|\hat{X}_n - \mu| > \epsilon) \to 0$$

at rate $1/n$. On the other hand, we can show that $\hat{X}_n$ converges to $\mu$ weakly (strongly) as long as $E|X| < \infty$.

## 3.3 Convergence in Probability

We say that $X_n$ *converges in probability* to $X$ as $n \to \infty$ if, for any positive $\epsilon$,

$$\lim_{n\to\infty} P(w : |X_n(w) - X(w)| > \epsilon) = 0.$$

This is written $X_n \xrightarrow{P} X$, as $n \to \infty$. A necessary and sufficient condition for such convergence is that for any positive $\epsilon$ and $\eta$ there is an $n_0$ such that

$$P(w : |X_n(w) - X(w)| > \epsilon) < \eta \qquad \text{for all } n \geq n_0.$$

A numerical constant $c$ can always be viewed as a degenerate random variable $C$ whose distribution has all of its probability concentrated on the single value $c$. As an example, the weak law of large numbers states that the random variable *sample mean* converges in probability to a population mean (a constant).

Now we try to use the following theorem and the example to illustrate the difference between *converegence with probability* 1 and *convergence in probability*. For *convergence in probability*, one needs for every $\epsilon > 0$ that the probability that $X_n$ is within $\epsilon$ of $X$ tends to one. For *convergence almost surely*, one needs for every $\epsilon > 0$ that the probability that $X_n$ stays within $\epsilon$ of $X$ for all $k \geq n$ tends to one as $n$ tends to infinity.

**Theorem 9** *The sequence* $\{X_n\}$ *of random variables converges to a random variable* $X$ *with probability* 1 *if and only if*

$$\lim_{n\to\infty} P\left\{\cup_{m=n}^{\infty}(|X_m - X| \geq \epsilon)\right\} = 0$$

*for every* $\epsilon > 0$.

By the above theorem, convergence in probability is weaker than *convergence with probability* 1. The following example is used to illustrate the difference.

Example 6. Let $\Omega = [0,1]$, and let S be the class of all Borel sets on $\Omega$. Let $P$ be the Lebesgue measure. For any positive integer $n$, choose integer $m$ with $2^m \leq n < 2^{m+1}$. Clearly, $n \to \infty$ if and only if $m \to \infty$. We can write $n \geq 1$ as $n = 2^m + k$, $k = 0, 1, \ldots, 2^m - 1$. Let us define $X_n$ on $\Omega$ by

$$X_n(w) = \begin{cases} 1 & \text{if } w \in \left[\frac{k}{2^m}, \frac{k+1}{2^m}\right], \\ 0 & \text{otherwise,} \end{cases}$$

if $n = 2^m + k$. Then $X_n$ is a random variable which satisfies

$$P\{|X_n(w)| \geq \epsilon\} = \begin{cases} \frac{1}{2^m} & \text{if } 0 < \epsilon < 1, \\ 0 & \text{if } \epsilon \geq 1, \end{cases}$$

so that $X_n \xrightarrow{P} 0$. However, $X_n$ does not converge to 0 with probability 1. In fact, for any $w \in [0,1]$, there are an infinite number of intervals of the form $[k/2^m, (k+1)/2^m]$ which contain $w$. Such a sequence of intervals depends on $w$. Let us denote it by

$$\left\{ \left[ \frac{k}{2^m}, \frac{k+1}{2^m} \right], m = 1, 2, \ldots \right\},$$

and let $n_m = 2^m + k_m$. Then $X_{n_m}(w) = 1$, but $X_n(w) = 0$ if $n \neq n_m$. It follows that $\{X_n\}$ does not converge at $w$. Since $w$ is arbitrary, $X_n$ does not converge with probability 1 to any random variable.

### 3.3.1 Borel-Cantelli Lemma

First, we give an example to illustrate the difference between *convergence in probability* and *convergence in distribution*. Consider $\{X_n\}$ where $X_n$ is uniformly distributed on the set of points $\{1/n, 2/n, \ldots, 1\}$. It can be shown easily that $X_n \xrightarrow{L} X$ where $X$ is uniformly distributed over $(0,1)$. Can we answer the question whether $X_n \xrightarrow{P} X$?

Next, we give the Borel-Cantelli Lemma and the concept of *infinitely often* which are often used in proving strong law of large number. For events $A_j$, $j = 0, 1, \ldots$, the event $\{A_j \text{ i.o.}\}$ (read $A_j$ *infinitely often*), stands for the event that infinitely many $A_j$ occur.

**THE BOREL-CANTELLI LEMMA.** If $\sum_{j=1}^{\infty} P(A_j) < \infty$, then $P\{A_j \text{ i.o.}\} = 0$. Conversely, if the $A_j$ are independent and $\sum_{j=1}^{\infty} P(A_j) = \infty$, then $P\{A_j \text{ i.o.}\} = 1$.

**Proof.** (The general half) If infinitely many of the $A_j$ occur, then for all $n$, at least one $A_j$ with $j \geq n$ occurs. Hence,

$$P\{A_j \text{ i.o.}\} \leq P\left\{\bigcup_{j=n}^{\infty} A_j\right\} \leq \sum_{j=n}^{\infty} P(A_j) \to 0.$$

The proof of the converse can be found in standard probability textbook.

A typical example of the use of the Borel-Cantelli Lemma occurs in coin tossing. Let $X_1, X_2, \ldots$ be a sequence of independent Bernoulli trials with probability of success on the $n$th trial equal to $p_n$. What is the probability of an infinite number of successes? Or, equivalently, what is $P\{X_n = 1 \text{ i.o.}\}$? From the Borel-Cantelli Lemma and its converse, this probability is zero or one depending on whether $\sum p_n < \infty$ or not. If $p_n = 1/n^2$, for example, then $P\{X_n = 1 \text{ i.o.}\} = 0$. If $p_n = 1/n$, then $P\{X_n = 1 \text{ i.o.}\} = 1$.

The Borel-Cantelli Lemma is useful in dealing with problems involving almost sure convergence because $X_n \xrightarrow{a.s.} X$ is equivalent to

$$P\{|X_n - X| > \epsilon \text{ i.o.}\} = 0, \quad \text{for all } \epsilon > 0.$$

### 3.4 Convergence in $r$th Mean

We say that $X_n$ *converges in rth mean* to $X$ if

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

This is written $X_n \xrightarrow{rth} X$, $n \to \infty$. We say that $X$ is dominated by $Y$ if $|X| \leq Y$ almost surely, and that the sequence $\{X_n\}$ is dominated by $Y$ iff this is true for each $X_n$ with the same $Y$. We say that $X$ or $\{X_n\}$ is *uniformly bounded* iff the $Y$ above may be taken to be a constant. Observe that

$$E|X_n - X|^r = E|X_n - X|^r 1_{\{|X_n - X| < \epsilon\}} + E|X_n - X|^r 1_{\{|X_n - X| > \epsilon\}} \leq \epsilon^r + EY^r 1_{\{|X_n - X| > \epsilon\}}.$$

We then conclude that $X_n \xrightarrow{rth} X$ if $X_n \xrightarrow{P} X$ and $\{X_n\}$ is dominated by some $Y$ that belongs to $L^p$.

We now use a Chebyshev type of "weak laws of large numbers" to demonstrate a method for determining the large sample behavior of linear combination of random variables.

**Theorem** (Chebyshev). Let $X_1, X_2, \ldots$ be uncorrelated with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. If $\sum_{i=1}^n \sigma_i^2 = o(n^2)$, $n \to \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i \xrightarrow{P} 0.$$

**Proof.** By the Chebyschev's inequality, we see that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right|>\epsilon\right)\leq\frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\to 0.$$

By the definition of convergence in probability, we prove that this theorem holds.

**Remark**. The method we just use can be viewed as an application of the fact of **Convergence in rth Mean Implies Convergence in Probability**.

**Theorem 10** *If $X_n$ converges to $X$ in $r$th mean, then it converges to $X$ in probability. The converse is true provided that $\{X_n\}$ is dominated by some $Y$ with $E|Y|^r<\infty$.*

**Proof.** For any $\epsilon>0$,

$$E|X_n-X|^r\geq E\{|X_n-X|^r I(|X_n-X|>\epsilon)\}\geq\epsilon^r P(|X_n-X|>\epsilon)$$

and thus

$$P(|X_n-X|>\epsilon)\leq\epsilon^{-r}E|X_n-X|^r\to 0,\ \ n\to\infty.$$

As a further demonstration of **Convergence in $r$th Mean Implies Convergence in Probability**, consider the following two examples.

Example 7. Let $X_1,X_2,\ldots$ be $n$ independent random variables with mean $\mu$, common variance $\sigma^2$, and common third and fourth moments about their mean, $\mu_3$ and $\mu_4$, respectively (that is $\mu_r=E(X_i-\mu_i)^r$). Show $s^2=(n-1)^{-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ converges to $\sigma^2$ in probability.

**Solution 1**: Calculate $E(s^2-\sigma^2)^2$ and employ the fact that **Convergence in $r$th Mean Implies Convergence in Probability**.

**Fact**: Let $X_1,X_2,\ldots,X_n$ be $n$ independent random variables with means $\theta_1,\theta_2,\ldots,\theta_n$, common variance $\sigma^2$, and common third and fourth moments about their mean, $\mu_3$ and $\mu_4$, respectively. If $\mathbf{A}$ is any $n\times n$ symmetric matrix, and $\mathbf{a}$ is the column vector of the diagonal elements of $\mathbf{A}$, then

$$Var[\boldsymbol{X^t AX}]=(\mu_4-3\sigma^4)\boldsymbol{a}^t\boldsymbol{a}+2\sigma^4 tr\boldsymbol{A}^2+4\sigma^2\boldsymbol{\theta}^t\boldsymbol{A}^2\boldsymbol{\theta}+4\mu_3\boldsymbol{\theta}^t\boldsymbol{A}\boldsymbol{a}.$$

Observe that $E(s^2)=\sigma^2$ and $s^2$ can be written as $\boldsymbol{X}^t\boldsymbol{A}\boldsymbol{X}$, where $\mathbf{A}$ is a projection matrix with diagonal elements $1-n^{-1}$ and off-diagonal elements $-n^{-1}$. By calculation, we have $Var(s^2)=(\mu_4-\frac{n-3}{n-1}\sigma^4)/n$. Hence, $s^2\xrightarrow{P}\sigma^2$.

**Solution 2**: Should we use such a complicate calculation? Write

$$s^2=\frac{1}{n-1}\left[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right]=\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-\frac{n}{n-1}\bar{X}^2.$$

Using WLLN and the definition of *convergence in probability*, we have $s^2 \xrightarrow{P} \sigma^2$ since $X_n + Y_n \xrightarrow{P} c + d$ when $X_n \xrightarrow{P} c$ and $Y_n \xrightarrow{P} d$.

Example 8. (Lehmann[1983], p333) Consider a two-state Markov chain. The variables $X_1, X_2, \ldots$ each take on the values 0 and 1, with the joint distribution determined by the initial probability $P(X_1 = 1) = p_1$, and the transition probabilities

$$P(X_{i+1} = 1|X_i = 0) = \pi_0, \qquad P(X_{i+1} = 1|X_i = 1) = \pi_1,$$

of which we shall assume $0 < \pi_0, \pi_1 < 1$. For such a chain, the probability $p_k = P(X_k = 1)$ typically depends on $k$ and the initial probability $p_1$. However, as $k \to \infty$, $p_k$ tends to a limit $p$, which is independent of $p_1$. It is easy to see what the value of $p$ must be. For consider the recurrence relation

$$p_{k+1} = p_k\pi_1 + (1 - p_k)\pi_0 = p_k(\pi_1 - \pi_0) + \pi_0. \tag{1}$$

If $p_k \to p$, this implies $p = \pi_0/(1 - \pi_1 + \pi_0)$. Since $p_k = (p_1 - p)(\pi_1 - \pi_0)^{k-1} + p$ by the iteration of (1) with $k = 1$, we have $p_k \to p$ by $|\pi_1 - \pi_0| < 1$.

Now, the question is how to estimate $p$. Recall that $p_k = (p_1 - p)(\pi_1 - \pi_0)^{k-1} + p$. We then expect that $p_k \approx p$ when $k$ is large. After $n$ trials, a natural estimator is $\bar{X}_n$ which is the frequency of ones in these trials. We now prove that $\bar{X}_n$ is a consistent estimator of $p$. Observe that $E(\bar{X}_n) = (p_1 + \cdots + p_n)/n$. Since $p_n \to p$, we have $E(\bar{X}_n) \to p$ or $\bar{X}_n$ is asymptotically unbiased. Note that $E\bar{X}_n - p$ is the bias of proposed estimate $\bar{X}$. Consistency of $\bar{X}_n$ will therefore follow if we can show that $Var(\bar{X}_n) \to 0$. Now

$$Var(\bar{X}_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)/n^2$$

and for $i < j$

$$Cov(X_i, X_j) = P(X_i = 1, X_j = 1) - p_ip_j = P(X_i = 1)P(X_j = 1|X_i = 1) - p_ip_j.$$

Define the probability $p_{j,i} = P(X_j = 1|X_i = 1)$ for $i < j$. Then we have

$$\begin{aligned} p_{j,i} &= \pi_1 p_{j-1,i} + \pi_0(1 - p_{j-1,i}) = p_{j-1,i}(\pi_1 - \pi_0) + \pi_0 \\ p_{i+1,i} &= \pi_1. \end{aligned}$$

Note that

$$\begin{aligned} p_ip_{j,i} - p_ip_j &= p_i\{[(\pi_1 - p)(\pi_1 - \pi_0)^{j-i-1} + p] - [(p_1 - p)(\pi_1 - \pi_0)^{j-1} + p]\} \\ &= (\pi_1 - \pi_0)^{j-i}\{(\pi_1 - p) - (p_1 - p)(\pi_1 - \pi_0)^{i-1}\}[(p_1 - p_0)(\pi_1 - \pi_0)^{i-1} + p]. \end{aligned}$$

Hence, we have

$$|Cov(X_i, X_j)| \leq M|\pi_1 - \pi_0|^{j-i}.$$

Therefore, $Var(\bar{X}_n)$ is of order $n^{-1}$ and hence that $\bar{X}_n$ is consistent.

### 3.5 Relationships Among The Modes of Convergence

**Theorem 1**. (a) If $X_n \overset{wp1}{\to} X$, then $X_n \overset{P}{\to} X$.
(b) If $X_n \overset{rth}{\to} X$, then $X_n \overset{P}{\to} X$.
(c) If $X_n \overset{P}{\to} X$, then $X_n \overset{d}{\to} X$.

**Proof of (c)**. For $\epsilon > 0$ and $x$ is a point of continuity of $F$, we have

$$\{X \leq x-\epsilon\} = \{X_n \leq x, X \leq x-\epsilon\} \cup \{X_n > x, X \leq x-\epsilon\} \subset \{X_n \leq x\} \cup \{|X_n - X| \geq \epsilon\}.$$

Hence,

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

Since $X_n \overset{P}{\to} X$, we have $P(|X_n - X| \geq \epsilon) \to 0$. Thus $F(x-\epsilon) \leq \liminf_{n\to\infty} F_n(x)$. By a similar argument, we have $\limsup_{n\to\infty} F_n(x) \leq F(x+\epsilon)$. Since $x$ is a point of continuity of $F$, we have

$$\liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = F(x).$$

## 4 Conditions for Existence of Moments of a Distribution

In this section, we discuss the relationship between moments and probability. In fact, we talk about *convergence in moments implies convergence in probability* before. We now attempt to answer when can we say that *convergence in probability implies convergence in moments*. Refer to Chapter 1 of Serfling (1980) for further details.

**Lemma 1** *For any random variable $X$, (a) $E|X| = \int_0^\infty P(|X| \geq t)dt$, $(\leq \infty)$ and*
*(b) if $E|X| < \infty$, then $P(|X| \geq t) = o(t^{-1})$, $t \to \infty$.*

**Proof**. Denote by $G$ the distribution of $|X|$ and let $c$ denote a (finite) continuity point of $G$. By integration by parts, we have

$$\int_0^c x dG(x) = \int_o^c [1 - G(x)]dx - c[1 - G(c)], \tag{2}$$

and hence also

$$\int_0^c x dG(x) \leq \int_o^c [1 - G(x)]dx. \tag{3}$$

Further, it is easily seen that

$$c[1 - G(c)] \leq \int_c^\infty x dG(x). \tag{4}$$

Now suppose that $E|X| = \infty$. Then (3) yields (a) for this case. On the other hand, suppose that $E|X| < \infty$. Then (4) yields (b). Also, making use of (4) in conjuction with (2), we obtain (a) for this case.

The lemma immediately yields its own generalization:

**Corollary** For any random variable $X$ and real number $r > 0$,
(a) $E|X|^r = r \int_0^\infty t^{r-1} P(|X| \geq t) dt$, and
(b) if $E|X^r| < \infty$, then $P(|X| \geq t) = o(t^{-r})$, $t \to \infty$.

If $X_n \xrightarrow{d} X$, we may also want to know if $E(X_n^r) \to E(X^r)$ for various choices of $r$, usually 1 and 2. Usually, we refer to $E(X^r)$ as the *asymptotic* $r$th moment of $X_n$, while $\lim_{n\to\infty} E(X_n^r)$ is the *limit of the $r$th moment* of $X_n$, if it exists. Recall Slutsky Theorem which states that $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{p} 0$ imply $X_n + Y_n \xrightarrow{d} X$. Moreover, $Y_n \xrightarrow{p} 0$ does not guarantee that $E(Y_n^r) \not\to 0$. Hence, *convergence in moments* will not imply *convergence in probability* in general. An implication is that the *asymptotic variance* is not necessary equal to the limit of the *variance*. What can we stated quite generally is that the limit of the variances is greater than or equal to the asymptotic variance which is given in the next theorem.

**Theorem 11** . *If $X_n \xrightarrow{d} X$ and if we let $Var(X_n)$ denote the variance of $X_n$ when it exists and set it equals to $\infty$ otherwise, then*

$$\liminf_{n\to\infty} Var(X_n) \geq Var(X).$$

To see this, let us first state the following lemma.

**Lemma 2** . *Let $Y_n$, $n = 1, 2, \ldots$ be a sequence of random variables tending in law to a random variable $Y$ with cdf $H$ and with $E(Y^2) = v^2$. Let $Y_{nA}$ be the random variable $Y_n$ truncated at $\pm A$, so that $Y_{nA} = Y_n$ if $|Y_n| \leq A$, and $Y_{nA} = A$ or $-A$ if $Y_n > A$ or $< -A$.*
*(i) Then*

$$\lim_{A\to\infty} \lim_{n\to\infty} E(Y_{nA}^2) = \lim_{A\to\infty} \lim_{n\to\infty} E\left[\min(Y_n^2, A^2)\right].$$

*exists and is equal to $v^2$.*
*(ii) If, in addition, $E(Y_n^2) \to w^2$ as $n \to \infty$, then $v^2 \leq w^2$.*

**Proof**. (i) Note that $Y_{nA}$ is a truncation of the random variable $Y_n$. Since $Y_n$ tends in law to $Y$, it follows that

$$\lim_{n\to\infty} E(Y_{nA}^2) = \int_{-A}^{A} y^2 dH(y) + A^2 P(|Y| > A),$$

and as $A \to \infty$, the right side tends to $v^2$.
(ii) It follows easily that

$$\lim_{A\to\infty} \lim_{n\to} E(Y_{nA}^2) \leq \lim_{n\to} \lim_{A\to\infty} E(Y_{nA}^2) \tag{5}$$

provided the indicated limits exist. Now

$$\lim_{A\to\infty} E(Y_{nA}^2) = E(Y_n^2)$$

so that the right side of (5) is $w^2$, while the left side is $v^2$ by (i).

Next, we give an example in which the asymptotic variance is equal to the limit of variance.
Example 9. Suppose that $X_n$ has the binomial distribution $\mathcal{B}(n,p)$. Let $\hat{p} = n^{-1}X_n$ and $U_n = \sqrt{n}(\hat{p}-p)$. Then we have

$$\sqrt{n}(\hat{p}^2 - p^2) = 2pU_n + \frac{U_n^2}{\sqrt{n}}$$

. Observe that

$$E\frac{U_n^2}{\sqrt{n}} = \sqrt{n}E(\hat{p}-p)^2 = \frac{p(1-p)}{\sqrt{n}}.$$

By Markov inequality, we have $U_n^2/\sqrt{n} \xrightarrow{P} 0$. It follows from Slutsky Theorem, $\sqrt{n}(\hat{p}^2 - p^2) \xrightarrow{d} V$, where $V$ has the normal distribution $\mathcal{N}(0, 4p^3(1-p))$. We calculate that the actual variance of $\sqrt{n}(\hat{p}_n^2 - p^2)$ is :

$$\begin{aligned} Var(\sqrt{n}(\hat{p}_n^2 - p^2)) &= nVar(\hat{p}_n^2) \\ &= 4p^3(1-p) + n^{-1}p^2(10p^2 - 16p + 6) - n^{-2}p(6p^3 - 12p^2 + 7p - 1) \\ &= 4p^3(1-p) + O(n^{-1}). \end{aligned}$$

Now we state a theorem which can be used to justify why the above example holds. (i.e., The reason is that $|\hat{p}| \leq 1$.) Quite often, we are interested in getting an approximation of moments of $h(\bar{X})$. Suppose that $\bar{X} \xrightarrow{P} \mu$. If $h$ is continuous, Taylor expansion gives us

$$h(x) = h(\mu) + h^{'}(\mu)(x-\mu) + \frac{1}{2}h^{''}(\mu)(x-\mu)^2 + R(x,\mu).$$

We then *expect* that

$$\begin{aligned} Eh(\bar{X}) &= h(\mu) + h^{'}(\mu)E(\bar{X}-\mu) + \frac{1}{2}h^{''}(\mu)var(\bar{X}) + R_n \\ &= h(\mu) + \frac{1}{2n}h^{''}(\mu)\sigma^2 + R_n \end{aligned}$$

where $R_n = E(R(\bar{X},\mu))$. The question is whether $R_n$ tends to zero fast enough. Refer to Chapter 1 of Bickel and Doksum (1977) for further reading.

**Theorem 12** *(Dominated Convergence Theorem) Suppose that* $X_n \xrightarrow{P} X$, $|X_n| \leq |Y|$ *with probability* 1 *(all* $n$*), and* $E|Y|^r < \infty$. *Then* $X_n \xrightarrow{rth} X$.

# 5 Further Discussion on Convergence in Distribution

The following theorem provides another methodology (characteristic function) for establishing convergence in distribution.

**Theorem 13** *(Serfling[1980], pp16). Let the distribution functions $F, F_1, F_2, \ldots$ possess respective characteristic functions $\phi, \phi_1, \phi_2, \ldots$. The following statements are equivalent:*

1. $F_n \xrightarrow{w} F$;
2. $\lim_n \phi_n(t) = \phi(t)$, *each real* $t$;
3. $\lim_n \int g dF_n = \int g dF$, *each bounded continuous function* $g$.

To demonstrate the above theorem, we will prove the most widely known versions of the Central Limit Theorem.

**Theorem 14** *(Lindeberg-Levy). Let $\{X_i\}$ be I.I.D. with mean $\mu$ and finite variance $\sigma^2 > 0$. Then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} N(0, \sigma^2).$$

**Proof**. We may suppose $\mu = 0$ by considering the r.v.'s $X_i - \mu$, whose second moment is $\sigma^2$. Note that

$$\begin{aligned} E\left(\exp\left(it\sum_{i=1}^{n}\frac{X_i}{\sigma\sqrt{n}}\right)\right) &= \left[\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \\ &= \left\{1 + \frac{i^2\sigma^2}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + o\left(\frac{|t|}{\sigma\sqrt{n}}\right)^2\right\}^n \\ &= \left\{1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right\}^n \to \exp(-t^2/2). \end{aligned}$$

The limit being the ch.f. of normal distribution, the proof is ended.

We now state a result due to Cramer and Wold which states that the distribution of a p-dimensional random variable is completely determined by the one-dimensional distributions of linear functions. This result allows the question of convergence of multivariate distribution functions to be reduced to that convergence of univariate distribution functions.

**Theorem**. In $R^k$, the random vector $\boldsymbol{X}_n$ converges in distribution to the random vector $\boldsymbol{X}$ if and only if each linear combination of the components of $\boldsymbol{X}_n$ converges in distribution to the same linear combination of the components of $\boldsymbol{X}$.

Next, we use this "Cramer-Wold device" to prove *asymptotic multivariate normality of cell frequency vectors*. Consider a squence of $n$ independent trials, with $k$ possible outcomes for each trial. Let $p_j$ denote the probability of occurrence of the $j$th outcome in any given trial ($\sum_1^k p_j = 1$). Let $n_j$ denote the number of occurrences of the $j$th outcome in the series of $n$ trials ($\sum_i^k n_j = n$). We call $(n_1, \ldots, n_k)$ the "cell frequency vector" associated with the $n$ trials.

The exact distribution of $(n_1, \ldots, n_k)$ is the multinomial distribution $\mathcal{M}(n, \boldsymbol{p})$ where $\boldsymbol{p} = (p_1, \ldots, p_k)$. Then $E(n_i) = np_i$, $Var(n_i) = np_i(1-p_i)$ and $Cov(n_i, n_j) = -np_ip_j$, so that $E(n_1, \ldots, n_k) = n\boldsymbol{p}$, $Cov((n_1, \ldots, n_k)) = n(\boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p})$, where $\boldsymbol{D}_p = diag(\boldsymbol{p})$. Let $\hat{\boldsymbol{p}} = n^{-1}(n_1, \ldots, n_k)$ be the vector of sample proportions, and set $\boldsymbol{U}_n = \sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p})$. Then $E(\boldsymbol{U}_n) = \boldsymbol{0}$, $Cov(\boldsymbol{U}_n) = \boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p}$.

**Theorem 15** *The random vector $\boldsymbol{U}_n$ converges in distribution to $k$-variate normal with mean $\boldsymbol{0}$ and covariance $\boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p}$.*

**Proof**. We compute the characteristic function of $E\exp(it\sum_{i=1}^n u_i)$ where $\boldsymbol{U}_n = (u_1, \ldots, u_k)$. Observe that

$$\begin{aligned}
E\left(\exp\left[it\sum_{j=1}^k \lambda_j u_j\right]\right) &= E\left(\exp\left[\sum_{j=1}^k it\lambda_j\left(\frac{n_j}{\sqrt{n}} - \sqrt{n}p_j\right)\right]\right) \\
&= \exp\left(-it\sqrt{n}\sum_{j=1}^k \lambda_j p_j\right)\cdot E\left(\exp\left[\frac{it}{\sqrt{n}}\sum_{j=1}^k \lambda_j n_j\right]\right) \\
&= \exp\left(-it\sqrt{n}\sum_{j=1}^k \lambda_j p_j\right)\cdot\left(\sum_{j=1}^k p_j \exp\left(\frac{it}{\sqrt{n}}\lambda_j\right)\right)^n \\
&= \left(\sum_{j=1}^k p_j \cdot \exp\left[\frac{it}{\sqrt{n}}\left(\lambda_j - \sum_{i=1}^k \lambda_i p_i\right)\right]\right)^n \\
&= \left\{\sum_{j=1}^k p_j\left[1 + \frac{it}{\sqrt{n}}(\lambda_j - \sum_{i=1}^k \lambda_i p_i) - \frac{t^2}{2n}(\lambda_j - \sum_{i=1}^k \lambda_i p_i)^2 + o(n^{-1})\right]\right\}^n \\
&= \left\{1 - \frac{t^2}{2n}\sum_{j=1}^k p_j\left(\lambda_j - \sum_{i=1}^k \lambda_i p_i\right)^2 + o(n^{-1})\right\}^n \\
&\to \exp\left(-\frac{t^2}{2}(\lambda_1, \ldots, \lambda_k)(\mathbf{D}_p - \mathbf{p}^t\mathbf{p})(\lambda_1, \ldots, \lambda_k)^t\right).
\end{aligned}$$

The limit being the ch.f. of the multivariate normal distribution with mean vector $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D}_p - \boldsymbol{p}^t\boldsymbol{p}$.

# 6 Convergence in Distribution for Perturbed Random Variables

A common situation in mathematical statistics is that the statistic of interest is a slight modification of a random variable having a known limit distribution. Usually, we can use the following theorem to derive the limit distribution for perturbed random variable.

**Theorem 16** *(Slutsky). Let* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$*, where* $c$ *is a finite constant. Then*
*(i)* $X_n + Y_n \xrightarrow{d} X + c$*;*
*(ii)* $X_n Y_n \xrightarrow{d} cX$*;*
*(iii)* $X_n/Y_n \xrightarrow{d} X/c$ *if* $c \neq 0$*.*

**Proof of (i)**. Choose and fix $t$ such that $t-c$ is a continuity point of $F_X$. Let $\epsilon > 0$ be such that $t - c + \epsilon$ and $t - c - \epsilon$ are also continuity points of $F_X$. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \leq t) \\ &\leq P(X_n + Y_n \leq t, |Y_n - c| < \epsilon) + P(|Y_n - c| \geq \epsilon) \\ &\leq P(X_n \leq t - c + \epsilon) + P(|Y_n - c| \geq \epsilon). \end{aligned}$$

Hence, by the hypotheses of the theorem, and by the choice of $t - c + \epsilon$,

$$\begin{aligned} \limsup_n F_{X_n+Y_n}(t) &\leq \limsup_n P(X_n \leq t - c + \epsilon) + \limsup_n P(|Y_n - c| \geq \epsilon) \\ &= F_X(t - c + \epsilon). \end{aligned}$$

Similarly,

$$P(X_n \leq t - c - \epsilon) \leq P(X_n + Y_n \leq t) + P(|Y_n - c| \geq \epsilon)$$

and thus

$$F_X(t - c - \epsilon) \leq \liminf_n F_{X_n+Y_n}(t).$$

Since $t - c$ is a continuity point of $F_X$, and since $\epsilon$ may be taken arbitrarily small, we have

$$\lim_n F_{X_n+Y_n}(t) = F_X(t - c) = F_{X+c}(t).$$

Now we derive the asymptotic distribution of sample variance to demonstrate Slustky Theorem.

**(Cont.) Example 7**. Let $X_1, X_2, \ldots$ be $n$ independent random variables with mean $\mu$, common variance $\sigma^2$, and common third and fourth moments about

their mean, $\mu_3$ and $\mu_4$, respectively. Find the asymptotic distribution of $s^2$.
**Solution**: It is well-known that $(n-1)s^2/\sigma^2$ follows a Chi-square distribution with degrees of freedom $n-1$ when the $X_i$'s are normally distributed. By the calculation of Example 1.1, it states that $var(s^2) = (\mu_4 - \frac{n-3}{n-1}\sigma^4)/n$. This result indicates that the distribution of $s^2$ will depend on the third and fourth moments of $X$ in general. This raises the following question:

> Can we look up the chi-square distribution table to derive confidence interval or to do hypotheses testing on $\sigma^2$ in general?

or

> Is the distribution of $s^2$ is robust against the assumption of *normality*?

To get some feeling on the above question, we will derive the distribution of $\sqrt{n}(s_n^2 - \sigma^2)$ without assuming that $X_i$'s are normally distributed. To make our life easier, we rely on an asymptotic analysis to derive the asymptotic distribution of $\sqrt{n}(s_n^2 - \sigma^2)$.

Observe that

$$\frac{n}{n-1}s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

By WLLN, we expects that $\bar{X} \approx EX$ or

$$\frac{n-1}{n}s_n^2 \approx \frac{1}{n}\sum_{i=1}^{n}(X_i - E(X))^2.$$

Then CLT can be applied or

$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} N(0, Var((X-EX)^2)).$$

But how can we make it rigorously? Without loss of generality, we can assume that $EX = 0$ and $Var(X) = 1$. Write

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n-1}{n}\sum_{i=1}^{n}X_i^2 - \frac{1}{n}\sum_{i\neq j}X_iX_j.$$

Observe that

$$\begin{aligned}
E\left(\frac{1}{n}\sum_{i\neq j}X_iX_j\right) &= 0 \\
var(X_1, X_2) &= E(X_1X_2)^2 \\
Cov(X_1X_2, X_1X_3) &= E(X_1^2X_2X_3) - [E(X_1X_2)][E(X_1X_3)] = 0 \\
Cov(X_1X_2, X_3X_4) &= E(X_1X_2X_3X_4) - [E(X_1X_2)][E(X_3X_4)] = 0 \\
Var\left(\frac{1}{n}\sum_{i\neq j}X_iX_j\right) &= \frac{1}{n^2}\{n(n-1)Var(X_1X_2) + n(n-1)(n-2)cov(X_1X_2, X_1X_3) \\
&\quad + n(n-1)(n-2)(n-3)(n-4)cov(X_1X_2, X_3X_4)\} = \frac{n-1}{n}E(X_1^2X_2^2).
\end{aligned}$$

Hence,

$$\frac{\sqrt{n}}{n}\sum_{i\neq j} X_iX_j \xrightarrow{P} 0.$$

By Slustky Theorem, we have

$$\sqrt{n}(s_n^2-\sigma^2) \xrightarrow{d} N(0, Var((X-EX)^2))$$

or

$$\begin{aligned} P\left(\sqrt{n}(s_n^2-\sigma^2) \le 1.96\sqrt{\mu_4-\sigma^4}\right) &\approx 0.975 \\ P\left(s_n^2 \le \sigma^2 + 1.96\frac{\sqrt{\mu_4-\sigma^4}}{\sqrt{n}}\right) &\approx 0.975. \end{aligned}$$

On the other hand, $(n-1)s_n^2/\sigma^2 \sim \chi^2_{n-1}$ when $X_i$'s are normally distributed. If we can show that the following statement holds by checking the chi-square table, it implies that the procedure for finding confidence interval of $\sigma^2$ or doing hypothesis testing on $\sigma^2$ based on normal theory is generally applicable when $n$ is large.

$$P\left(s_n^2 \le \sigma^2 + 1.96\frac{\sqrt{\mu_4-\sigma^4}}{\sqrt{n}}\right) \approx 0.975.$$

Suppose that $V \sim \chi^2_{n-1}$. We can write $V = Z_1^2+\cdots+Z_{n-1}^2$ where $Z_i \sim N(0,1)$. Hence,

$$\frac{V-(n-1)}{\sqrt{2(n-1)}} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$. It means that a Chi-square distribution of degree of freedom $n-1$ can be approximated well by a normal distribution with mean $n-1$ and variance $2(n-1)$ as n tends to infinity. Suppose we look up the chi-square distribution table to derive confidence interval for $\sigma^2$.

$$\begin{aligned} & P\left(s_n^2 \le \sigma^2 + 1.96\frac{\sqrt{\mu_4-\sigma^4}}{\sqrt{n}}\right) \\ = \;& P\left(\frac{(n-1)s_n^2}{\sigma^2} \le n-1+1.96\frac{(n-1)\sqrt{\mu_4-\sigma^4}}{\sqrt{n}\sigma^2}\right) \\ = \;& P\left(\frac{\frac{(n-1)s_n^2}{\sigma^2}-(n-1)}{\sqrt{2(n-1)}} \le \frac{1.96(n-1)\sqrt{\mu_4-\sigma^4}/\sqrt{n}\sigma^2}{\sqrt{2(n-1)}}\right) \\ \approx \;& P\left(\frac{\frac{(n-1)s_n^2}{\sigma^2}-(n-1)}{\sqrt{2(n-1)}} \le 1.96\frac{\sqrt{\mu_4-\sigma^4}}{\sqrt{2\sigma^4}}\right). \end{aligned}$$

When $\mu_4 = 3\sigma^4$,

$$P\left(s_n^2 \le \sigma^2 + 1.96\frac{\sqrt{\mu_4-\sigma^4}}{\sqrt{n}}\right) \approx 0.975.$$

Otherwise, it cannot be close to 0.975. Therefore, the distribution of $s^2$ is not robust against the assumption of *normality*.

# 7 Convergence Properties of Transformed Sequences

Given that $X_n \to X$ in some sense of convergence, and given a function $g$, a basic question is whether $g(X_n) \to g(X)$ in the same sense of convergence. The following theorem states that the answer is *yes* if the function $g$ is continuous. As a remark, commonly used $\delta$-method for approximating moments and distributions relies on the following theorem. Further discussion of $\delta$-method can be found in the following two sections.

**Theorem 17** *Let $X_1, X_2, \ldots$ and $X$ be random $k$-vectors defined on a probability space and let $g$ be a vector-valued Borel function defined on $R^k$. Suppose that $g$ is continuous with $P_X$-probability* 1. *Then*
*(i) $X_n \xrightarrow{wp1} X$ implies $g(X_n) \xrightarrow{wp1} g(X)$;*
*(ii) $X_n \xrightarrow{p} X$ implies $g(X_n) \xrightarrow{p} g(X)$;*
*(iii) $X_n \xrightarrow{d} X$ implies $g(X_n) \xrightarrow{d} g(X)$.*

Suppose that $X = k$, a constant, $X_n \xrightarrow{p} k$ and $g$ is a function continuous at $k$, then $g(X_n) \xrightarrow{p} g(k)$.

**Proof.** If $g$ is continuous at $k$, then for every $\epsilon > 0$, there exists a constant $\delta > 0$ such that

$$|x - k| < \delta \quad \Rightarrow \quad |g(x) - g(k)| < \epsilon$$

so that

$$P(|X_n - k| < \delta) \leq P(|g(X_n) - g(k)| < \epsilon).$$

But $X_n \xrightarrow{p} k$, so $P(|X_n - k| < \delta) \to 1$. This implies that

$$P(|g(X_n) - g(k)| < \epsilon) \to 1.$$

Or, $g(X_n) \xrightarrow{p} g(k)$.

Consider the following example on finding the asymptotic distribution of $\hat{p}^2$, the square of a binomial proportion. We begin by observing that

$$\hat{p}^2 = p^2 + 2p(\hat{p} - p) + (\hat{p} - p)^2,$$

or

$$\sqrt{n}(\hat{p}^2 - p^2) = 2p\sqrt{n}(\hat{p} - p)\sqrt{n}(\hat{p} - p)^2 = 2pZ_n + n^{-1/2}Z_n^2,$$

where we set $Z_n = \sqrt{n}(\hat{p} - p)$. We know that $Z_n$ converges in distribution, so from above theorem we know that $Z_n^2$ also converges in distribution. Then $Z_n^2 = O_P(1)$, and hence $n^{-1/2}Z_n^2 = o_P(1)$. These facts give us

$$\sqrt{n}(\hat{p}^2 - p^2) = 2pZ_n + o_P(1).$$

We conclude that $\hat{p}^2$ has an approximate normal distribution with mean $p^2$ and variance $n^{-1}4p^3(1-p)$.

The most commonly considered functions of vectors converging in some stochastic sense are linear transformations and quadratic forms. As an example, consider the residual of sum squares in linear regression.

**Corollary 1** *Suppose that the $k$-vectors $\mathbf{X}_n$ converge to the $k$-vector $\mathbf{X}$ wp 1, or in probability, or in distribution. Let $\mathbf{A}_{m\times k}$ and $\mathbf{B}_{k\times k}$ be matrices. Then $\mathbf{AX}_n^{'} \to \mathbf{AX}^{'}$ and $\mathbf{X}_n\mathbf{BX}_n^{'} \to \mathbf{XBX}^{'}$ in the given mode of convergence.*

**Proof.** The vector-valued function

$$\mathbf{Ax}^{'} = \left(\sum_{i=1}^{k} a_{1i}x_i, \ldots, \sum_{i=1}^{k} a_{mi}x_i\right)$$

and the real-valued function

$$\mathbf{xBx}^{'} = \sum_{i=1}^{k}\sum_{j=1}^{k} b_{ij}x_ix_j$$

are continuous functions of $\mathbf{x} = (x_1, \ldots, x_k)$.

## 7.1 The $\delta$ Method for Calculating Asymptotic Distribution

If every individual in the population under study can be classified as falling into one and only one of $k$ categories, we say that the categories are mutually exclusive and exhaustive. A randomly selected member of the population will fall into one of the $k$ categories with probability $\mathbf{p}$, where $\mathbf{p}$ is the vector of cell probabilities

$$\mathbf{p} = (p_1, p_2, \ldots, p_k)$$

and $\sum_{i=1}^{k} p_i = 1$. Here the cells are strung out into a line for purposes of indexing only; their arrangement and ordering does not reflect anything about the characteristics of individuals falling into a particular cell. The $p_i$ reflect the relative frequency of each category in the population.

As an example, we might be interested in *whether hair color is related to eye color*. We then can conduct a study by collecting a random sample and get a count of the number of people who fall in this particular cross-classification

determined by hair color and eye color. When the cells are defined in terms of the categories of two or more variables, a structure relating to the nature of the data is imposed. The natural structure for two variables is often a rectangular array with columns corresponding to the categories of one variable and rows to categories of the second variable; three variables creates layers of two-way tables, and so on. The simplest contingency table is based on four cells, and the categories depend on two variables. The four cells are arranged in a $2 \times 2$ table whose two rows correspond to the categorical variable $A$ and whose two columns correspond to the second categorical variable $B$. Double subscripts refer to the position of the cells in our arrangement. The first subscript gives the category number of variable $A$, the second of variable $B$, and the two-dimensional array is displayed as a grid with two rows and two columns. The probability $p_{ij}$ is the probability of an individual being in category $i$ of variable $A$ and category $j$ of variable $B$. Usually, we have some theory in mind which can be checked in terms of hypothesis testing such as

$$H_0 : \boldsymbol{p} = \boldsymbol{\pi} \ \ (\boldsymbol{\pi} \text{ a fixed value}).$$

Then the problem can phrased as $n$ observations from the $k$-cell multinomial distribution with cell probabilities $p_1, \ldots, p_k$. Then we encounter the problem of *proving asymptotic multivariate normality of cell frequency vectors.* To test $H_0$, it can be proceed by the Pearson chi square test, which is to reject $H_0$ if $X^2$ is too large, where

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i}.$$

This test statistic was first derived by Pearson (1900). Then we need to answer two questions. The first one is to determine what kind of the magnitude of $X^2$ is the so-called *too large.* The second one is whether the Pearson chi-square test is a reasonable testing procedure. These questions will be tackled by deriving the asymptotic distribution of the Pearson chi square statistic under $H_0$ and a local alternative of $H_0$.

Using matrix notation, $X^2$ can be written as

$$X^2 = \boldsymbol{U}_n \mathbf{D}_{\boldsymbol{\pi}}^{-1} \boldsymbol{U}_n^t,$$

where

$$\mathbf{U}_n = \sqrt{n}(\hat{\mathbf{p}} - \boldsymbol{\pi}), \quad \hat{\boldsymbol{p}} = n^{-1}(n_1, \ldots, n_k), \ \text{ and } \mathbf{D}_{\boldsymbol{\pi}} = diag(\boldsymbol{\pi}).$$

Let $g(\mathbf{x}) = \mathbf{x}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{x}^t$ for $\mathbf{x} = (x_1, \ldots, x_k)$. Evidently, $g$ is a continuous function of $\mathbf{x}$. It can be shown that $\mathbf{U}_n \xrightarrow{d} \mathbf{U}$, where $\mathbf{U}$ has the multivariate normal

distribution $\mathcal{N}(0, \mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}^t\boldsymbol{\pi})$. Then by the Corollary in previous section, we have

$$\mathbf{U}_n\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}_n^t \xrightarrow{d} \mathbf{U}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}^t.$$

Thus the asymptotic distribution of $X^2$ under $H_0$, which is the distribution of $\mathbf{U}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}^t$, where $\mathbf{U}$ has the $\mathcal{N}(0, \mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}^t\boldsymbol{\pi})$ distribution. This reduces the problem to finding the distribution of a quadratic form of a multivariate normal random vector. The above process is the so-called $\delta$ method.

Now we state without proof the following general result on the distribution of a quadratic form of a multivariate normal random variable. It can be found in Chapter 3b in Rao (1973) and Chapter 3.5 of Serfling (1980).

**Theorem 18** *If* $\mathbf{X} = (X_1, \ldots, X_d)$ *has the multivariate normal distribution* $\mathcal{N}(0, \Sigma)$ *and* $Y = \mathbf{X}\mathbf{A}\mathbf{X}^t$ *for some symmetric matrix* $\mathbf{A}$*, then* $\mathcal{L}[Y] = \mathcal{L}[\sum_{i=1}^d \lambda_i Z_i^2]$*, where* $Z_1^2, \ldots, Z_d^2$ *are independent chi square variables with one degree of freedom each and* $\lambda_1, \ldots, \lambda_d$ *are the eigenvalues of* $\mathbf{A}^{1/2}\Sigma(\mathbf{A}^{1/2})^t$*.*

Apply the above theorem to the present problem, we see that $\mathcal{L}[\mathbf{U}\mathbf{D}_{\boldsymbol{\pi}}^{-1}\mathbf{U}^t] = \mathcal{L}[\sum_{i=1}^d \lambda_i Z_i^2]$, where $\lambda_i$ are the eigenvalues of

$$\mathbf{B} = \mathbf{D}_{\boldsymbol{\pi}}^{-1/2}(\mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}^t\boldsymbol{\pi})\mathbf{D}_{\boldsymbol{\pi}}^{-1/2} = \mathbf{I} - \sqrt{\boldsymbol{\pi}^t}\sqrt{\boldsymbol{\pi}},$$

where $\sqrt{\boldsymbol{\pi}} = (\sqrt{\pi}_1, \ldots, \sqrt{\pi}_k)$. Now it remains to find the eigenvalues of $\mathbf{B}$. Since $\mathbf{B}^2 = \mathbf{B}$ and $\mathbf{B}$ is symmetric, the eigenvalues of $\mathbf{B}$ are all either 1 or 0. Moreover,

$$\sum_{i=1}^k \lambda_i = tr(\mathbf{B}) = k - 1.$$

Therefore, we establish the result that under the simple hypothesis $H_0$, Pearson's chi-square statistic $X^2$ has an asymptotic chi square distribution with $k - 1$ degrees of freedom.

We already examined the limiting distribution of the Pearson chi square statistic under $H_0$ by employing $\delta$ method. In essence, the $\delta$ method requires two ingredients: first, a random variable (which we denote here by $\hat{\theta}_n$) whose distribution depends on a real-valued parameter $\theta$ in such a way that

$$\mathcal{L}[\sqrt{n}(\hat{\theta}_n - \theta)] \to N(0, \sigma^2(\theta)); \tag{6}$$

and second, a function $f(x)$ that can be differentiated at $x = \theta$ so that it possesses the following expansion about $\theta$:

$$f(x) = f(\theta) + (x - \theta)f'(\theta) + o(|x - \theta|) \text{ as } x \to \theta. \tag{7}$$

The $\delta$ method for finding approximate means and variances (asymptotic mean and asymptotic variance) of a function of a random variable is justified by the following theorem.

**Theorem 19** *(The one-dimensional $\delta$ method.) If $\hat{\theta}_n$ is a real-valued random variable and $\theta$ is a real-valued parameter such that (6) holds, and if $f$ is a function satisfying (7), then the asymptotic distribution of $f(\hat{\theta}_n)$ is given by*

$$\mathcal{L}[\sqrt{n}(f(\hat{\theta}_n) - f(\theta))] \to N(0, \sigma^2(\theta)[f^{'}(\theta)]^2). \tag{8}$$

**Proof**. Set $\Omega_n = R$, $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n \times \cdots = \times_{n=1}^{\infty}\Omega_n$, and $P_n$ to be the probability distribution of $\hat{\theta}_n$ on $R$. Note that $\Omega$ is the set of all sequences $\{t_n\}$ such that $t_n \in \Omega_n$. We define two subsets of $\Omega$:

$$\begin{aligned} S &= \{\{t_n\} \in \Omega : t_n - \theta = O(n^{-1/2})\}, \\ T &= \{\{t_n\} \in \Omega : f(t_n) - f(\theta) - (t_n - \theta)f^{'}(\theta) = o(n^{-1/2})\}. \end{aligned}$$

Since $f$ satisfies (7), then $S \subset T$. By (6), we have

$$n^{1/2}(\hat{\theta}_n - \theta) = O_P(1) \text{ and hence } \hat{\theta}_n - \theta = O_P(n^{-1/2}). \tag{9}$$

Note that $S$ occurs in probability and hence $T$ also occur in probability since $S \subset T$. Finally,

$$f(\hat{\theta}_n) - f(\theta) - (\hat{\theta}_n - \theta)f^{'}(\theta) = o_P(n^{-1/2}) \tag{10}$$

or

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta)) = \sqrt{n}(\hat{\theta}_n - \theta)f^{'}(\theta) + o_P(1). \tag{11}$$

Now let $V_n = \sqrt{n}(f(\hat{\theta}_n) - f(\theta))$, $U_n = \sqrt{n}(\hat{\theta}_n - \theta)$, and $g(x) = xf^{'}(\theta)$ for all real numbers $x$. Then (11) may be rewritten as

$$V_n = g(U_n) + o_P(1).$$

Now we discuss the power of Pearson's chi square test when $\boldsymbol{p} = \boldsymbol{\pi} + n^{-1/2}\mu$. This case is useful in the study of goodness-of-fit tests when the model being tested is wrong but not far wrong. In this case,

$$\begin{aligned} E(\mathbf{X}_n) &= n\boldsymbol{\pi} + \sqrt{n}\boldsymbol{\mu}, \\ Cov(\mathbf{X}_n) &= n(\mathbf{D}_\pi - \boldsymbol{\pi}^{'}\boldsymbol{\pi}) + \sqrt{n}(\mathbf{D}_\mu - 2\boldsymbol{\pi}^{'}\boldsymbol{\mu}) + \boldsymbol{\mu}^{'}\boldsymbol{\mu}. \end{aligned} \tag{12}$$

The coordinates of $\boldsymbol{\pi}$ and $\mathbf{p}$ both sum to 1 so that $\boldsymbol{\mu}$ satisfies the condition $\sum_{i=1}^{k}\mu_i = 0$. Here $\boldsymbol{\mu}$ acts as a noncentrality parameter.

Set $\boldsymbol{U}_n = \sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{\pi})$, so that $E(\boldsymbol{U}_n) = \boldsymbol{\mu}$, and

$$Cov(\boldsymbol{U}_n) = \boldsymbol{D}_\pi - \boldsymbol{\pi}^t\boldsymbol{\pi} + n^{-1/2}(\boldsymbol{D}_\mu - 2\boldsymbol{\pi}^t\boldsymbol{\mu}) + n^{-1}\boldsymbol{\mu}^t\boldsymbol{\mu}. \tag{13}$$

**Theorem 20** . $\mathcal{L}[\boldsymbol{U}_n] \to \mathcal{L}[\mathbf{U}]$, *where* $\mathbf{U}$ *has the multivariate normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{D}_\pi - \boldsymbol{\pi}^t\boldsymbol{\pi}$.

**Proof.** We compute the characteristic function of $E\exp(it\sum_{i=1}^k u_i)$ and show that it converges to that of $\boldsymbol{U}$:

$$\begin{aligned}
E\left(\exp\left[it\sum_{j=1}^k \lambda_j u_j\right]\right) &= E\left(\exp\left[\sum_{j=1}^k it\lambda_j\left(\frac{n_j}{\sqrt{n}} - \sqrt{n}p_j\right)\right]\right) \\
&= \exp\left(-it\sqrt{n}\sum_{j=1}^k \lambda_j p_j\right)\cdot E\left(\exp\left[\frac{it}{\sqrt{n}}\sum_{j=1}^k \lambda_j n_j\right]\right) \\
&= \exp\left(-it\sqrt{n}\sum_{j=1}^k \lambda_j p_j\right)\cdot\left(\sum_{j=1}^k p_j \exp\left(\frac{it}{\sqrt{n}}\lambda_j\right)\right)^n \\
&= \left(\sum_{j=1}^k p_j \cdot \exp\left[\frac{it}{\sqrt{n}}\left(\lambda_j - \sum_{i=1}^k \lambda_i\right)\right]\right)^n \\
&= \left\{\sum_{j=1}^k p_j\left[1 + \frac{it}{\sqrt{n}}(\lambda_j - \sum_{i=1}^k \lambda_i p_i) - \frac{t^2}{2n}(\lambda_j - \sum_{i=1}^k \lambda_i p_i)^2 + o(n^{-1})\right]\right\}^n \\
&= \left\{1 - \frac{t^2}{2n}\sum_{j=1}^k p_j\left(\lambda_j - \sum_{i=1}^k \lambda_i p_i\right)^2 + o(n^{-1})\right\}^n \\
&\to \exp\left(it\sum_{j=1}^k \lambda_i\mu_i - \frac{t^2}{2}(\lambda_1,\ldots,\lambda_k)(\mathbf{D}_\pi - \boldsymbol{\pi}^t\boldsymbol{\pi})(\lambda_1,\ldots,\lambda_k)^t\right)
\end{aligned}$$

using the facts that $\sum_{i=1}^k \mu_i t_i = \boldsymbol{\mu}\mathbf{t}^t$ and

$$\sum_{i=1}^k \pi_i(t_i - \mathbf{t}\boldsymbol{\pi}^t)^2 = \mathbf{t}(\mathbf{D}_\pi - \boldsymbol{\pi}^t\boldsymbol{\pi})\mathbf{t}^t.$$

The limit being the ch.f. of the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{D}_\pi - \boldsymbol{\pi}^t\boldsymbol{\pi}$.

## 7.2 Variance-Stabilizing Transformations

Sometime the statistic of interest for inference about a parameter $\theta$ is conveniently asymptotically normal, but with an asymptotic variance parameter functionally dependent on $\theta$, i.e., the asymptotic variance is $\sigma^2(\theta)$. According to the $\delta$ method, a smooth transformation, $g$, of statistic also are approximately normally distributed. It turns out to be useful to know transformations $g$ called *variance stabilizing*, such that the asymptotic variance of that statistic is approximately independent of the parameter $\theta$. Usually, $g$ can be found by solving a differential equation.

**Example**. (Binomial Proportion) It is known that $\sqrt{n}(\hat{p}-p) \xrightarrow{d} N(0, p(1-p))$, where $\hat{p}$ is a binomial proportion. Here the variance of $\hat{p}$ depends on both $n$ and $p$. The problem of variance stabilization is to find a one-to-one function $g : D \to R$ such that the variance of $g(\hat{p})$ is proportional to $n^{-1}$ and does not depend on $p$. Suppose that such a $g$ exists. Applying the $\delta$ method to $g$, we have a differential equation

$$\frac{dg}{dp} = \frac{1}{\sqrt{p(1-p)}}.$$

It can be solved easily that

$$g(p) = 2 \arcsin \sqrt{p} + \ constant.$$

Further examples of the variance-stabilizing technique may be found in Rao (1973), Section 6.g. Refer to the first five sections of Chapter 3 of Serfling (1980) for additional reading on transformation of given statistics.

### 7.3 Multivariate versions of the $\delta$ method

Let $\hat{\boldsymbol{\theta}}_n$ be a $T$-dimensional random vector: $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nT})$, and let $\boldsymbol{\theta}$ be a $T$-dimensional vector parameter: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_T)$. We assume that $\hat{\boldsymbol{\theta}}_n$ has an asymptotic normal distribution in the sense that

$$\mathcal{L}[\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})] \to N(0, \Sigma(\hat{\boldsymbol{\theta}}_n)). \tag{14}$$

Here $\Sigma(\hat{\boldsymbol{\theta}}_n)$ is the $T \times T$ asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_n$.

Now suppose $\boldsymbol{f}$ is a function defined on an open subset of $T$-dimensional space and taking values in $R$-dimensional space, i.e.,

$$\mathbf{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_R(\boldsymbol{\theta}))\,.$$

We assume that $\boldsymbol{f}$ has a differential at $\boldsymbol{\theta}$, i.e., that $\mathbf{f}$ has the following expansion as $\mathbf{x} \to \boldsymbol{\theta}$:

$$f_i(\mathbf{x}) = f_i(\boldsymbol{\theta}) + \sum_{j=1}^{T} (x_j - \theta_j) \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{x}=\boldsymbol{\theta}} + o(\|\mathbf{x} - \boldsymbol{\theta}\|) \tag{15}$$

for $i = 1, \ldots, R$. If we let $(\partial \mathbf{f}/\partial \boldsymbol{\theta})$ denote the $R \times T$ matrix whose $(i, j)$ entry is the partial derivative of $f_i$ with respect to the $j$th coordinate of $\mathbf{x} = (x_1, \ldots, x_T)$ evaluated at $\mathbf{x} = \ \boldsymbol{\theta}$, i.e.,

$$\left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)_{ij} = \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{x}=\boldsymbol{\theta}},$$

then (15) can be expressed neartly in matrix notation as

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\boldsymbol{\theta}) + (\mathbf{x} - \boldsymbol{\theta}) \left( \left. \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} \right|_{\mathbf{x}} + o(\|\mathbf{x} - \boldsymbol{\theta}\|) \right) \tag{16}$$

as $\mathbf{x} \to \boldsymbol{\theta}$. Within this framework, the $\delta$ method can be stated as follows:

**Theorem 21** *(Multivariate $\delta$ method.) If $\hat{\boldsymbol{\theta}}_n$, $\boldsymbol{\theta}$, and $\mathbf{f}$ be as described above and suppose (14) and (16) hold. Then the asymptotic distribution of $f(\hat{\boldsymbol{\theta}}_n)$ is given by:*

$$\mathcal{L}[\sqrt{n}(f(\hat{\boldsymbol{\theta}}_n) - f(\boldsymbol{\theta}))] \to N\left(0, \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right) \Sigma(\boldsymbol{\theta}) \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)'\right). \tag{17}$$

Note that a broad class of statistics of interest, such as the sample coefficient of variation $s/\bar{x}$, may be expressed as a smooth function of a vector of the basic sample statistics. The sample moments are known to be asymptotically jointly normal statistics. Then the above theorem can be used to find out the asymptotic distribution of statistics of interest. Refer to Chapter 8 of Ferguson (1996) on the asymptotic behavior of the sample correlation coefficient.

## References

[1] Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden=Day, San Francisco.

[2] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, Mass.

[3] Chung, K.L. (1974). *A Course in Probability Theory.* 2nd ed., Academic Press, New York.

[4] Cramer, H. and Wold, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.* **11** 290-294.

[5] Ferguson, T.S. (1996). *A Course in PLarge Sample Theory.* 1st ed., Chapman & Hall, London.

[6] Lehmann, E.L. (1983). *Theory of Point Estimation.* John Wiley & Sons, New York.

[7] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variable is such that it can reasonably be supposed to have arisen from random sampling. *Phil. Mag. Series 5* **50** 157-175.

[8] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed., wiley, New York.

[9] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.