CHAPTER 5.
Convergence of Random Variables

5.1. Introduction

One of the most important parts of probability theory concerns the behavior of sequences of random variables. This part of probability is often called "large sample theory" or "limit theory" or "asymptotic theory." This material is extremely important for statistical inference. The basic question is this: what can we say about the limiting behavior of a sequence of random variables $X_1, X_2, X_3, \ldots$? Since statistics is all about gathering data, we will naturally be interested in what happens as we gather more and more data, hence our interest in this question.

Recall that in calculus, we say that a sequence of real numbers x_n converges to a limit x if, for every $\epsilon > 0$, $|x_n - x| < \epsilon$ for all large n. In probability, convergence is more subtle. Going back to calculus for a moment, suppose that $x_n = x$ for all n. Then, trivially, $\lim_n x_n = x$. Consider a probabilistic version of this example. Suppose that $X_1, X_2, \ldots$ are a sequence of random variables which are independent and suppose each has a $N(0,1)$ distribution. Since these all have the same distribution, we are tempted to say that X_n "converges" to $Z \sim N(0,1)$. But this can't quite be right since $P(X_n = Z) = 0$ for all n.

Here is another example. Consider $X_1, X_2, \ldots$ where $X_i \sim N(0, 1/n)$. Intuitively, X_n is very concentrated around 0 for large n. But $P(X_n = 0) = 0$ for all n. The next section develops appropriate methods of discussing convergence of random variables.

5.2. Types of Convergence

Let us start by giving some definitions of different types of convergence. It is easy to get overwhelmed. Just hang on and remember this: the two key ideas in what follows are "convergence in probability" and "convergence in distribution."

Suppose that $X_1, X_2, \ldots$ have finite second moments. X_n converges to X *in quadratic mean* (also called convergence in L_2), written $X_n \stackrel{q.m.}{\to} X$, if,

$$E(X_n - X)^2 \to 0$$

as $n \to \infty$.

X_n converges to X *in probability*, written $X_n \xrightarrow{p} X$, if, for every $\epsilon > 0$,

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

as $n \to \infty$.

Let F_n denote the cdf of X_n and let F denote the cdf of X. X_n converges to X *in distribution*, written $X_n \xrightarrow{d} X$, if,

$$\lim_n F_n(t) = F(t)$$

at all t for which F is continuous.

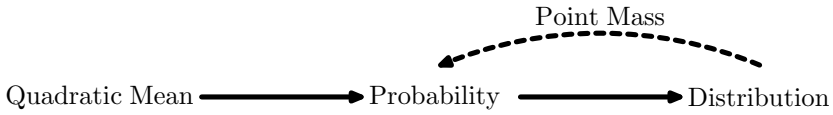
Here is a summary:

Quadratic Mean	$E(X_n - X)^2 \to 0$
In probability	$P(\|X_n - X\| > \epsilon) \to 0$ for all $\epsilon > 0$
In distribution	$F_n(t) \to F(t)$ at continuity points t

Recall that X is a point mass at c if $P(X = c) = 1$. The distribution function for X is $F(x) = 0$ if $x < c$ and $F(x) = 1$ if $x \geq c$. In this case, we write the convergence of X_n to X as $X \xrightarrow{q.m.} c$, $X \xrightarrow{p} c$, or $X \xrightarrow{d} c$, depending on the type of convergence. Notice that $X \xrightarrow{d} c$ means that $F_n(t) \to 0$ for $t < c$ and $F_n(t) \to 1$ for $t > c$. We do not require that $F_n(c)$ converge to 1, since c is not a point of continuity in the limiting distribution function.

EXAMPLE 5.2.1. Let $X_n \sim N(0, 1/n)$. Intuitively, X_n is concentrating at 0 so we would like to say that $X_n \xrightarrow{d} 0$. Let's see if this is true. Let F be the distribution function for a point mass at 0. Note that $\sqrt{n}X_n \sim N(0, 1)$. Let Z denote a standard normal random variable. For $t < 0$, $F_n(t) = P(X_n < t) = P(\sqrt{n}X_n < \sqrt{n}t) = P(Z < \sqrt{n}t) \to 0$ since $\sqrt{n}t \to -\infty$. For $t > 0$, $F_n(t) = P(X_n < t) = P(\sqrt{n}X_n < \sqrt{n}t) = P(Z < \sqrt{n}t) \to 1$ since $\sqrt{n}t \to \infty$. Hence, $F_n(t) \to F(t)$ for all $t \neq 0$ and so $X_n \xrightarrow{d} 0$. But notice that $F_n(0) = 1/2 \neq F(1/2) = 1$ so convergence fails at $t = 0$. But that doesn't matter because $t = 0$ is not a continuity point of F and the definition of convergence in distribution only requires convergence at continuity points.

The following diagram summarized the relationship between the types of convergence.



Here is the theorem that corresponds to the diagram.

THEOREM 5.2.1. *The following relationships hold:*
(a) $X_n \stackrel{q.m.}{\to} X$ *implies that* $X_n \stackrel{p}{\to} X$.
(b) $X_n \stackrel{p}{\to} X$ *implies that* $X_n \stackrel{d}{\to} X$.
(c) If $X_n \stackrel{d}{\to} X$ *and if* $P(X = c) = 1$ *for some real number* c, *then* $X_n \stackrel{p}{\to} X$.
In general, none of the reverse implications hold except the special case in (c).

PROOF. We start by proving (a). Suppose that $X_n \stackrel{q.m.}{\to} X$. Fix $\epsilon > 0$. Then, using Chebyshev's inequality,

$$P(|X_n - X| > \epsilon) = P(|X_n - X|^2 > \epsilon^2) \leq \frac{E|X_n - X|^2}{\epsilon^2} \to 0.$$

Proof of (b). This proof is a little more complicated. You may skip if it you wish. Fix $\epsilon > 0$. Then

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

Also,

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

Hence,

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

Take the limit as $n \to \infty$ to conclude that

$$F(x - \epsilon) \leq \text{liminf}_n F_n(x) \leq \text{limsup}_n F_n(x) \leq F(x + \epsilon).$$

This holds for all $\epsilon > 0$. Take the limit as $\epsilon \to 0$ and use the fact that F is continuous at x and conclude that $\lim_n F_n(x) = F(x)$.

Proof of (c). Fix $\epsilon > 0$. Then,

$$\begin{aligned} P(|X_n - c| > \epsilon) &= P(X_n < c - \epsilon) + P(X_n > c + \epsilon) \\ &\leq P(X_n \leq c - \epsilon) + P(X_n > c + \epsilon) \\ &= F_n(c - \epsilon) + 1 - F_n(c + \epsilon) \\ &\to F(c - \epsilon) + 1 - F(c + \epsilon) \\ &= 0 + 1 - 0 = 0. \end{aligned}$$

Let us now show that the reverse implications do not hold.

CONVERGENCE IN PROBABILITY DOES NOT IMPLY CONVERGENCE IN QUADRATIC MEAN. Let $U \sim \text{Unif}(0, 1)$ and let $X_n = \sqrt{n} I_{(0,1/n)}(U)$. Then $P(|X_n| > \epsilon) = P(\sqrt{n} I_{(0,1/n)}(U) > \epsilon) = P(0 \leq U < 1/n) = 1/n \to 0$. Hence, Then $X_n \xrightarrow{p} 0$. But $E(X_n^2) = n \int_0^{1/n} du = 1$ for all n so X_n does not converge in quadratic mean.

CONVERGENCE IN DISTRIBUTION DOES NOT IMPLY CONVERGENCE IN PROBABILITY. Let $X \sim N(0, 1)$. Let $X_n = -X$ for $n = 1, 2, 3, \ldots$; hence $X_n \sim N(0, 1)$. X_n has the same distribution function as X for all n so, trivially, $\lim_n F_n(x) = F(x)$ for all x. Therefore, $X_n \xrightarrow{d} X$. But $P(|X_n - X| > \epsilon) = P(|2X| > \epsilon) = P(|X| > \epsilon/2) \neq 0$. So X_n does not tend to X in probability.

Warning! One might conjecture that if $X_n \xrightarrow{p} b$ then $E(X_n) \to b$. This is not true. Let X_n be a random variable defined by $P(X_n = n^2) = 1/n$ and $P(X_n = 0) = 1 - (1/n)$. Now, $P(|X_n| < \epsilon) = P(X_n = 0) = 1 - (1/n) \to 1$. Hence, $Z \xrightarrow{p} 0$. However, $E(X_n) = [n^2 \times (1/n)] + [0 \times (1 - (1/n))] = n$. Thus, $E(X_n) \to \infty$.

Summary. Stare at the diagram.

5.3 The Law of Large Numbers

Now we come to a crowning achievement in probability, the law of large numbers. This theorem says that, in some sense, the mean of a large sample

is close to the mean of the distribution. For example, the proportion of heads of a large number of tosses is expected to be close to 1/2. We now make this more precise.

Let $X_1, X_2, \ldots,$ be an iid sample and let $\mu = E(X_1)$ and $\sigma^2 = Var(X_1)$.[1] The sample mean is defined as $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Recall these two important facts: $E(\overline{X}_n) = \mu$ and $Var(\overline{X}_n) = \sigma^2/n$.

THEOREM. 5.3.1. *(The Weak Law of Large Numbers.) If $X_1, \ldots, X_n$ are iid, then $\overline{X}_n \xrightarrow{p} \mu$.*

PROOF. Assume that $\sigma < \infty$. This is not necessary but it simplifies the proof. Using Chebyshev's inequality,

$$P\left(|\overline{X}_n - \mu| > \epsilon\right) \leq \frac{Var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$.

There is a stronger theorem in the appendix called the strong law of large numbers.

EXAMPLE 5.3.2. Consider flipping a coin for which the probability of heads is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = P(X_i = 1) = E(X_i)$. The fraction of heads after n tosses is $\overline{X}_n$. According to the law of large numbers, $\overline{X}_n$ converges to p in probability. This does not mean that $\overline{X}_n$ will numerically equal p. It means that, when n is large, the distribution of $\overline{X}_n$ is tightly concentrated around p. Let us try to quantify this more. Suppose the coin is fair, i.e $p = 1/2$. How large should n be so that $P(.4 \leq \overline{X}_n \leq .6) \geq .7$? First, $E(\overline{X}_n) = p = 1/2$ and $Var(\overline{X}_n) = \sigma^2/n = p(1-p)/n = 1/(4n)$. Now we use Chebyshev's inequality:

$$\begin{aligned} P(.4 \leq \overline{X}_n \leq .6) &= P(|\overline{X}_n - \mu| \leq .1) \\ &= 1 - P(|\overline{X}_n - \mu| > .1) \\ &\geq 1 - \frac{1}{4n(.1)^2} = 1 - \frac{25}{n}. \end{aligned}$$

The last expression will be larger than .7 of $n = 84$. Later we shall see that this calculation is unnecessarily conservative.

[1]Note that $\mu = E(X_i)$ is the same for all i so we can define μ in terms of X_1 or any other X_i.

5.4. The Central Limit Theorem

In this section we shall show that the sum (or average) of random variables has a distribution which is approximately Normal. Suppose that $X_1, \ldots, X_n$ are iid with mean μ and variance σ. The central limit theorem (CLT) says that $\overline{X}_n = n^{-1} \sum_i X_i$ has a distribution which is approximately Normal with mean μ and variance σ^2/n. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

THEOREM 5.4.1. (Central Limit Theorem). Let $X_1, \ldots, X_n$ be i.i.d with mean μ and variance σ^2. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n$. Then

$$Z_n \equiv \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$. In other words,

$$\lim_n P(Z_n \leq z) = \Phi(z)$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

is the cdf of a standard normal.

The proof is in the appendix. The central limit theorem says that the distribution of Z_n can be approximated by a $N(0,1)$ distribution. In other words:

> probability statements about Z_n can be approximated using a Normal distribution. It's the probability statements that we are approximating, not the random variable itself.

There are several ways to denote the fact that the distribution of Z_n can be approximated be a normal. They all mean the same thing. Here they are:

$$\begin{aligned} Z_n &\approx N(0,1) \\ \overline{X}_n &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\ \overline{X}_n - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \end{aligned}$$

$$\begin{aligned}\sqrt{n}(\overline{X}_n - \mu) &\approx N\left(0,\, \sigma^2\right) \\ \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} &\approx N(0,1).\end{aligned}$$

EXAMPLE. 5.4.2. Suppose that the number of errors per computer program has a Poisson distribution with mean 5. We get 125 programs. Let $X_1, \ldots, X_{125}$ be the number of errors in the programs. Let $\overline{X}$ be the average number of errors. We want to approximate $P(\overline{X} < 5.5)$. Let $\mu = E(X_1) = \lambda = 5$ and $\sigma^2 = Var(X_1) = \lambda = 5$. So

$$Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma = \sqrt{125}(\overline{X}_n - 5)/\sqrt{5} = 5(\overline{X}_n - 5) \approx N(0,1).$$

Hence,

$$P(\overline{X} < 5.5) = P(5(\overline{X} - 5) < 2.5) \approx P(Z < 2.5) = .9938.$$

EXAMPLE 5.4.3. We will compare Chebychev to the CLT. Suppose that $n = 25$ and suppose we wish to bound

$$P\left(\frac{|\overline{X}_n - \mu|}{\sigma} > \frac{1}{4}\right).$$

First, using Chebychev,

$$\begin{aligned}P\left(\frac{|\overline{X}_n - \mu|}{\sigma} > \frac{1}{4}\right) &= P\left(|\overline{X}_n - \mu| > \frac{\sigma}{4}\right) \\ &\leq \frac{Var(\overline{X})}{\frac{\sigma^2}{16}} = \frac{16}{25} = .64\end{aligned}$$

Using the CLT,

$$\begin{aligned}P\left(\frac{|\overline{X}_n - \mu|}{\sigma} > \frac{1}{4}\right) &= P\left(\frac{5|\overline{X}_n - \mu|}{\sigma} > \frac{5}{4}\right) \\ &\approx P\left(|Z| > \frac{5}{4}\right) = .21.\end{aligned}$$

The CLT gives a sharper bound, albeit with some error.

The central limit theorem tells us that $Z_n = \sqrt{n}(\overline{X} - \mu)/\sigma$ is approximately N(0,1). This is interesting but there is a practical problem: we don't always know σ. We can estimate σ^2 from $X_1, \ldots, X_n$ by

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

This raises the following question: if we replace σ with S_n is the central limit theorem still true? The answer is yes.

THEOREM 5.4.4. *Assume the same conditions as the CLT. Then,*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} Z$$

where $Z \sim N(0,1)$. Hence we may apply the central limit theorem with S_n in place of σ.

You might wonder, how accurate is the normal approximation. The answer is given in the Berry-Essèen theorem which we state next. You may skip this theorem if you are not interested.

THEOREM 5.4.5. *Suppose that $E|X_1|^3 < \infty$. Then*

$$\sup_z |P(Z_n \leq z) - \Phi(z)| \leq \frac{33}{4}\frac{E|X_1 - \mu|^3}{\sqrt{n}\sigma^3}.$$

5.5. The Effect of Transformations

Often, but not always, convergence properties are preserved under transformations.

THEOREM 5.5.1. *Let X_n, X, Y_n, Y be random variables.*
(a) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$.
(b) If $X_n \xrightarrow{q.m.} X$ and $Y_n \xrightarrow{q.m.} Y$, then $X_n + Y_n \xrightarrow{q.m.} X + Y$.

Generally, it is **not** the case that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ implies that $X_n + Y_n \xrightarrow{d} X + Y$. However, it does hold if one of the limits is constant.

THEOREM 5.5.2 (Slutzky's Theorem.) *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{d} c$, *then* $X_n + Y_n \xrightarrow{d} X + c$.

Products also preserve some forms of convergence.

THEOREM 5.5.3.
(a) If $X_n \xrightarrow{p} X$ *and* $Y_n \xrightarrow{p} Y$, *then* $X_n Y_n \xrightarrow{p} XY$.
(b) If $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{d} c$, *then* $X_n Y_n \xrightarrow{p} cX$.

Finally, convergence is also preserved under continuous mappings.

THEOREM 5.5.4. *Let* g *be a continuous mapping.*
(a) If $X_n \xrightarrow{p} X$ *then* $g(X_n) \xrightarrow{p} g(X)$.
(b) If $X_n \xrightarrow{d} X$ *then* $g(X_n) \xrightarrow{d} g(X)$.

Appendix A5.1. L_1 convergence and Almost Sure Convergence

We say that X_n converges almost surely to X, written $X_n \xrightarrow{a.s.} X$, if

$$P(\{s : \ X_n(s) \to X(s)\}) = 1.$$

When $P(X = c) = 1$ we can write this as

$$P(\lim_n X_n = c) = 1.$$

We say that X_n converges in L_1 to X, written $X_n \xrightarrow{L_1} X$, if

$$E|X_n - X| \to 0$$

as $n \to \infty$.

The following relationships hold in addition to those in 4.2.1.

THEOREM A5.1.1. *The following relationships hold:*
(a) $X_n \xrightarrow{a.s.} X$ *implies that* $X_n \xrightarrow{p} X$.
(b) $X_n \xrightarrow{q.m.} X$ *implies that* $X_n \xrightarrow{L_1} X$.
(c) $X_n \xrightarrow{L_1} X$ *implies that* $X_n \xrightarrow{p} X$.

Appendix A5.2. The Strong Law of Large Numbers

The weak law of large numbers says that $\overline{X}_n$ converges to EX_1 in probability. The strong law asserts that this is also true almost surely.

THEOREM A5.2.1. (The strong law of large numbers.) *Let $X_1, X_2, \ldots$ be iid. If $\mu = E|X_1| < \infty$ then $\overline{X}_n \stackrel{a.s.}{\to} \mu$.*

Appendix A5.3. Proof of the Central Limit Theorem

If X is a random variable, define its moment generating function (mgf) by $\psi_X(t) = Ee^{tX}$. Assume in what follows that the mgf is finite in a neighborhood around $t = 0$.

LEMMA A5.3.1. (Convergence using mgf's). *Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let ψ_n the mgf of Z_n. Let Z be another random variable and denote its mgf by ψ. If $\psi_n(t) \to \psi(t)$ for all t in some open interval around 0, then $Z_n \stackrel{d}{\to} Z$.*

PROOF OF THE CENTRAL LIMIT THEOREM. Let $Y_i = (X_i - \mu)/\sigma$. Then, $Z_n = n^{-1/2} \sum_i Y_i$. Let $\psi(t)$ be the mgf of Y_i. The mgf of $\sum_i Y_i$ is $(\psi(t))^n$ and mgf of Z_n is $[\psi(t/\sqrt{n})]^n \equiv \xi_n(t)$. Now $\psi'(0) = E(Y_1) = 0$, $\psi''(0) = E(Y_1^2) = Var(Y_1) = 1$. So,

$$\begin{aligned}
\psi(t) &= \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi''(0) + \cdots \\
&= 1 + 0 + \frac{t^2}{2} + \frac{t^3}{3!}\psi''(0) + \cdots \\
&= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\psi''(0) + \cdots
\end{aligned}$$

Now,

$$\begin{aligned}
\xi_n(t) &= \left[\psi\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}\psi''(0) + \cdots\right]^n \\
&= \left[1 + \frac{\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\psi''(0) + \cdots}{n}\right]^n \\
&\to e^{t^2/2}
\end{aligned}$$

which is the mgf of a N(0,1). The result follows from the previous Theorem. In the last step we used the following fact from calculus:

FACT: If $a_n \to a$ then

$$\left(1 + \frac{a_n}{n}\right)^n \to e^a.$$