Limit Theorems

Statistics 110

Summer 2006

Limit Theorems

Next is to look at the properties of a sequence of random variables $Y_1, Y_2, Y_3, \ldots$. For example, what happens to

$$Y_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

or

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

or

$$Y_n = \min(X_1, \ldots, X_n)$$

as $n \to \infty$.

Definition. *A sequence of RVs* $Y_n, n = 1, 2, 3, \ldots$ *is said to* **Converge in Probability** *to a constant* c *(denoted by* $Y_n \xrightarrow{P} c$*), if for any* $\epsilon > 0$*,*

$$P[|Y_n - c| \geq \epsilon] \to 0$$

as $n \to \infty$*.*

In other words, if I is any interval containing c, then eventually Y_n will have most of its probability concentrated in I.

Theorem. [Weak Law of Large Numbers] *Let* $X_1, X_2, \ldots$ *be independent RVs with* $E[X_i] = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2 < \infty$ *and*

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Then $\bar{X}_n \xrightarrow{P} \mu$*.*

Proof. As shown earlier, $E[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$. Thus by Chebyshev's inequality

$$P[|\bar{X}_n - \mu| \geq \epsilon] \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

□

Note that the restriction of a finite variance can be relaxed, though for most problems that isn't necessary.

Theorem. *Let $X_1, X_2, \ldots$ be independent RVs with $E[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2 < \infty$. If*

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{\overline{\sigma^2}}{n} \to 0$$

Then $\bar{X}_n \xrightarrow{P} \bar{\mu}$ (or $(\bar{X}_n - \bar{\mu}) \xrightarrow{P} 0$)

Lemma. *Suppose that $g(t)$ is a function that is continuous at $t = c$ and that $X_n \xrightarrow{P} c$. Then $g(X_n) \xrightarrow{P} g(c)$.*

Proof.

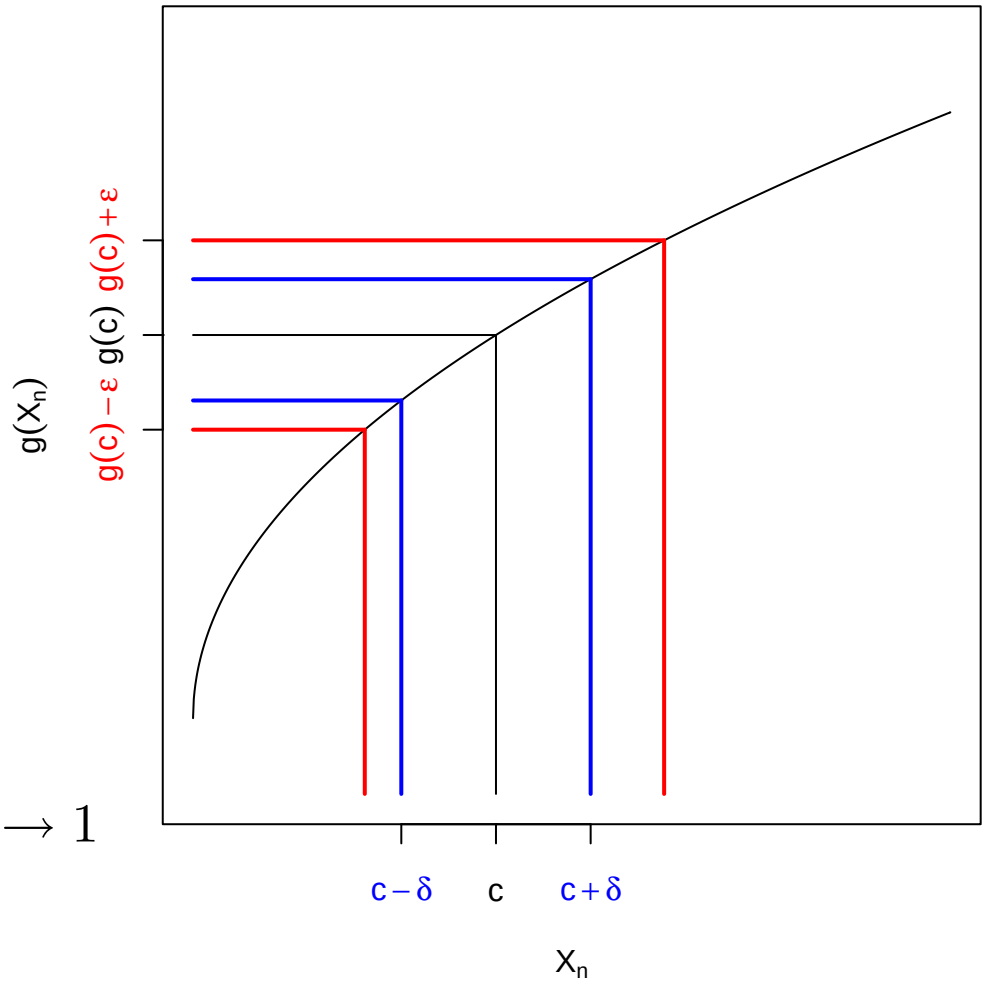
To see this, let $Z_n = g(X_n)$. Since $g(\cdot)$ is a continuous function, for any fixed $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|X_n - c| \leq \delta \implies |Z_n - g(c)| \leq \epsilon$$

Therefore,

$$P[|Z_n - g(c)| \leq \epsilon] \geq P[|X_n - c| \leq \delta] \to 1$$

□



The continuity assumption is important. If $g(\cdot)$ is not continuous at c, then we may not be able to find such a δ.

Example: Suppose

$$Y_n = \begin{cases} \frac{1}{n} & \text{with prob } 1 - \frac{1}{\sqrt{n}} \\ -n & \text{with prob } \frac{1}{\sqrt{n}} \end{cases}$$

Then $P[|Y_n - 0| > \epsilon] = \frac{1}{\sqrt{n}}$ for $n > \frac{1}{\epsilon}$ hence $Y_n \xrightarrow{P} 0$.

Let

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

Then

$$g(Y_n) = \begin{cases} 1 & \text{with prob } 1 - \frac{1}{\sqrt{n}} \\ 0 & \text{with prob } \frac{1}{\sqrt{n}} \end{cases}$$

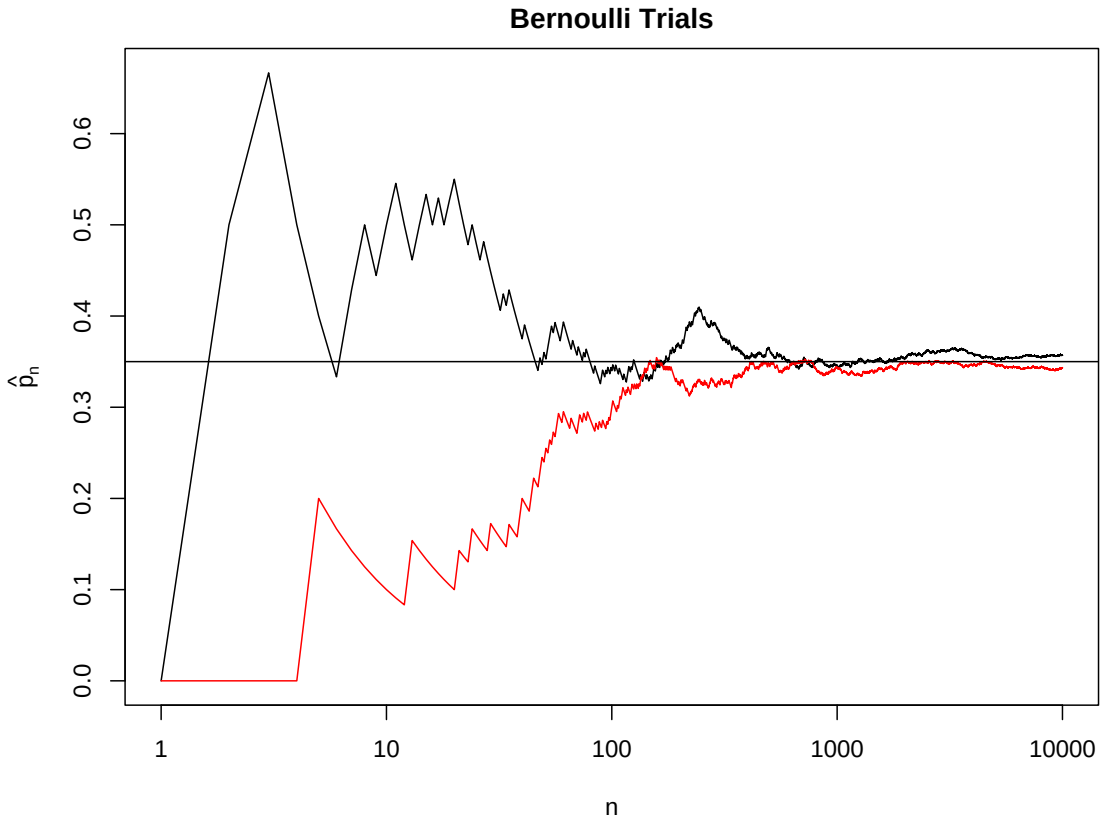
Which implies $g(Y_n) \xrightarrow{P} 1 \neq g(0)$. This is due to g being discontinuous at 0.

Example: Let $X_n \sim Bin(n, p)$ and let $\hat{p}_n = \frac{X_n}{n}$.

$$E[\hat{p}_n] = \frac{E[X_n]}{n} = \frac{np}{n} = p$$
$$\mathrm{Var}(\hat{p}_n) = \frac{\mathrm{Var}(X_n)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

As $\mathrm{Var}(\hat{p}_n) \to 0$ as $n \to \infty$, $\hat{p}_n \xrightarrow{P} p$. This result can also be seen from the law of large numbers, as p_n is the sample average of n iid $Bern(p)$.

This result supports the idea of treating probability as long run frequencies. Let A be the event of interest and Z_i be the indicator variable of whether outcome i fall in set A or not. As we have seen before $P[A] = P[Z_i = 1] = p$. $\hat{p}_n$ is the sample relative frequency after n trials which converges to $P[A]$.



We can use the lemma to show what happens to the common estimate of a Bernoulli variance. Let $g(t) = t(1-t)$. As this is a continuous function on $[0, 1]$,

$$g(\hat{p}_n) = \hat{p}_n(1 - \hat{p}_n) \xrightarrow{P} p(1-p) = \mathrm{Var}(Z_i)$$

The previous plot contains two examples of Monte Carlo integration. Let $g(x) = I\{x \leq 0.35\}$ where $X_i \sim U(0, 1)$. The last point of each line is calculated by

$$\hat{I}(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

which is an estimate of the quantity

$$I(g) = \int_0^1 g(x)dx = E[g(X)]$$

which by the law of large numbers, $\hat{I}(g)$ converges in probability to $I(g)$

Monte Carlo is usually used to calculate difficult integrals (or expected values)

For example, the book discusses calculating

$$\frac{1}{\sqrt{2\pi}} \int_0^1 e^{-x^2/2} dx = \Phi(1) - \Phi(0)$$

by Monte Carlo.

We have seen another example of Monte Carlo. The forecast SST maps were also calculated by Monte Carlo. They were based on the following setup.

Assume that X has density $f(x)$ and suppose you are interested in

$$E[g(X)] = \int_{\mathcal{X}} g(x) f(x) dx$$

This can be estimated by generating $X_1, X_2, \ldots, X_n$ from density $f(x)$ by calculating

$$\hat{I}(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i)$$

which by the law of large numbers, this converges in probability to $E[g(X)]$.

In the temperature maps, for a pixel in the map $g(x) = x$, where x is the temperature in that pixel.

Lemma. *If $X_n \xrightarrow{P} c$ and $Y_n \xrightarrow{P} d$ and $g(x, y)$ a continuous function in a neighbourhood containing the point (c, d), then $g(X_n, Y_n) \xrightarrow{P} g(c, d)$.*

For example, we can use this to prove that the sample variance S_n^2 converges in probability to $\mathrm{Var}(X_i) = \sigma^2$.

$$
\begin{aligned}
S_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) \\
&\xrightarrow{P} E[X^2] - (E[X])^2 = \mathrm{Var}(X)
\end{aligned}
$$

As shown, some of the probability bounds can be fairly loose. It is possible, if necessary, to get tighter bounds by looking at higher moments. For example, suppose X_i's are iid with $E[X_i] = 0$ and $\mathrm{Var}(X_i) = \sigma^2$.

By Chebyshev,

$$P[|\bar{X}_n| \geq a] \leq \frac{\sigma^2}{na^2} \to 0$$

at a rate of $\frac{1}{n}$. However we can do better. Assume that $E[X_i^4]$ is finite.

$$\begin{aligned} E[\bar{X}_n^4] &= \frac{1}{n^4} E[(X_1 + X_2 + \ldots + X_n)^4] \\ &= \frac{1}{n^4} \left[\sum_i E[X_i^4] + \sum_{i \neq j} E[X_i^3 X_j] + \sum_{i \neq j} E[X_i^2 X_j^2] \right. \\ &\qquad \left. + \sum_{i \neq j \neq k} E[X_i^2 X_j X_k] + \sum_{i \neq j \neq k \neq l} E[X_i X_j X_k X_l] \right] \end{aligned}$$

$$E[\bar{X}_n^4] = \frac{1}{n^4}\left[nE[X_i^4] + \frac{n(n-1)}{2}(E[X_i^2])^2\right]$$

$$\begin{aligned} P[|\bar{X}_n| \geq a] &= P[|\bar{X}_n|^4 \geq a^4] \\ &\leq \frac{E[\bar{X}_n^4]}{a^4} = \frac{n-1}{2n^3}\frac{\sigma^4}{a^4} + \frac{1}{n^3}\frac{E[X_i^4]}{a^4} \\ &\leq \frac{1}{n^2}\frac{\sigma^4}{a^4} + \frac{1}{n^3}\frac{E[X_i^4]}{a^4} \to 0 \end{aligned}$$

at a rate of $\frac{1}{n^2}$.

Being able to show convergence at higher rates can be useful for more complicated problems. Examples where this may occur include nonparametric regression, stochastic processes such as Brownian motion, and Markov Chain Monte Carlo. The following example gives a flavour of the advantages to be able to use higher rates of convergence.

Suppose you have multiple sequences of random variables where

$$X_{1n} \xrightarrow{P} c_1, X_{2n} \xrightarrow{P} c_2, \ldots, X_{in} \xrightarrow{P} c_i, \ldots$$

and you are interested in $\sum_{i=1}^{n} X_{in}$ as $n \to \infty$.

If each converges only at a rate of only $\frac{1}{n}$, the sum of the probabilities may not go to 0 (n terms of order $\frac{1}{n}$). However if each converges at a rate of $\frac{1}{n^2}$, the sum of the probabilities will go to 0 at a rate of at least $\frac{1}{n}$.

If a moment generating function exists (thus all moments exists) you can get even tighter bounds.

Theorem. [Chernoff bound] *Assume that RV X has a MGF $M_X(t)$. Then*

$$P[X \geq a] \leq e^{-ta} M_X(t) \qquad \text{for } t > 0$$
$$P[X \leq a] \leq e^{-ta} M_X(t) \qquad \text{for } t < 0$$

Proof. For $t > 0$,

$$P[X \geq a] = P[e^{tX} \geq e^{ta}] \leq \frac{E[e^{tX}]}{e^{ta}} = e^{-ta} M_X(t)$$

For $t < 0$

$$P[X \leq a] = P[e^{tX} \geq e^{ta}] \leq \frac{E[e^{tX}]}{e^{ta}} = e^{-ta} M_X(t)$$

□

Note that we get a different bound for each different t, so, if possible, we want to find which t will minimize the probability bound.

Example (back to the Widgets). Lets assume that $X \sim N(500, 100)$. What is a bound on $P[X > 550]$. The MGF for this normal is

$$M_X(t) = e^{500t + 50t^2}$$

$$P[X > 550] \leq \frac{e^{500t+50t^2}}{e^{550t}} = e^{50(t^2-t)}$$

This will be minimized by the t which minimizes $t^2 - t$, which happens to be $t = 0.5$. Thus

$$P[X > 550] \leq e^{-12.5} = 0.00000372$$

As seen before $P[X > 550] = 0.000000287$, so this is much closer than the one-sided Chebyshev bound of 0.0384 (a factor of 10 not 10,000).

Types of Convergence

There are different types of convergence of a random variable to a constant c. The most important are

1. Convergence almost surely (sometimes called almost everywhere or with probability 1)

$$X_n \xrightarrow{a.s.} c \Longleftrightarrow \text{For any } \epsilon > 0, P[|X_n - c| > \epsilon \text{ only finitely often}] = 1$$

   So eventually X_n gets within ϵ of c and never leaves, but when this happens is random.

2. Convergence in probability

$$X_n \xrightarrow{P} c \Longleftrightarrow \text{For any } \epsilon > 0, P[|X_n - c| < \epsilon] \to 1$$

Example: Let $\omega \sim U(-1,1)$ and

$$Y_n(\omega) = \begin{cases} -n & \frac{-1}{\sqrt{n}} < \omega < \frac{1}{\sqrt{n}} \\ \frac{1}{n} & \text{Otherwise} \end{cases}$$

Then

- $P[|Y_n - 0| > \epsilon] \to 0$. Hence $Y_n \xrightarrow{P} 0$.
- For any $\omega \neq 0$, $Y_n(\omega) \to 0$. Hence $Y_n \xrightarrow{a.s.} 0$.

Note that $E[Y_n] = -n\frac{1}{\sqrt{n}} + \frac{1}{n}(1 - \frac{1}{\sqrt{n}}) \approx -\sqrt{n} \to -\infty \neq 0$

Hence in general $Y_n \xrightarrow{a.s.} c$ does not imply $E[Y_n] \to c$.

Stronger results are needed for $Y_n \xrightarrow{a.s.} c$ to imply $E[Y_n] \to c$. One such result is the dominated convergence theorem (in appendix).

In general, almost sure convergence is much stronger than convergence in probability. The following theorem supports this

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

Proof. Omitted $\square$

The other direction doesn't hold. It is possible to have $X_n \xrightarrow{P} c$ but it will not converge almost everywhere.

The following example shows that you can have convergence in probability, but not convergence almost surely.

Let $\Omega = [0, 1)$ and let $\omega \sim U(0, 1)$

$$X_1 = 1$$

$$X_2 = I\{0 \leq \omega < 0.5\}$$

$$X_3 = I\{0.5 \leq \omega < 1\}$$

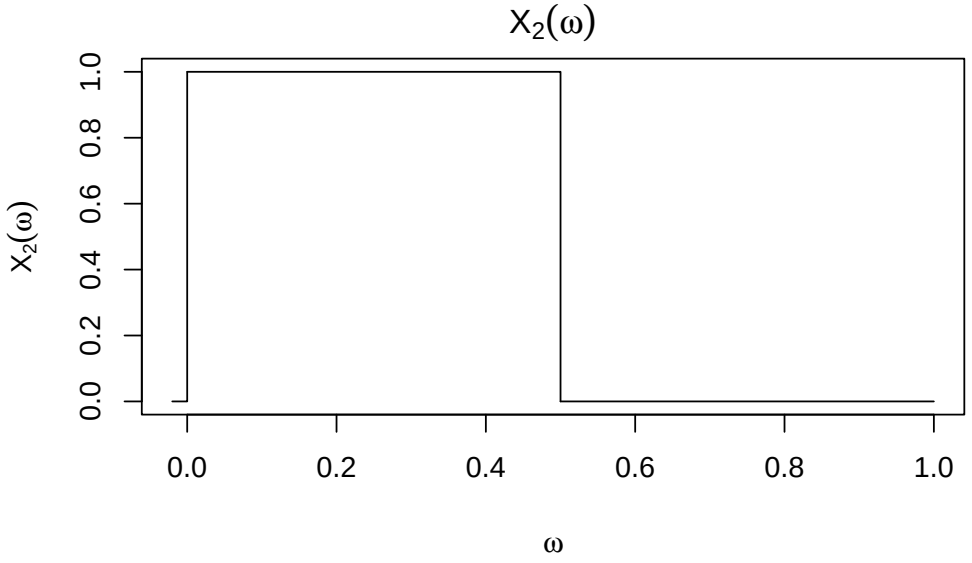
$$X_4 = I\{0 \leq \omega < 0.25\}$$

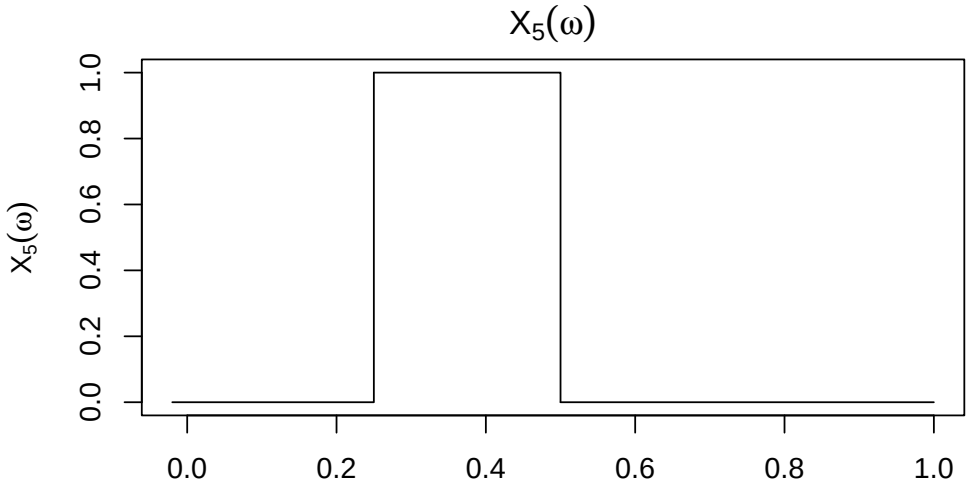
$$X_5 = I\{0.25 \leq \omega < 0.5\}$$

$$X_6 = I\{0.5 \leq \omega < 0.75\}$$

$$X_7 = I\{0.75 \leq \omega < 1\}$$

$$X_8 = I\{0 \leq \omega < 0.125\}$$





• • •

$$X_n = I\left\{\frac{m}{2^k} \leq \omega < \frac{m+1}{2^k}\right\} \quad \text{if } n = 2^k + m$$

- $P[X_n = 0] = 1 - \frac{1}{2^k} \to 1$, hence $X_n \xrightarrow{P} 0$

- For any ω, $X_n(\omega) = 1$ for infinitely many n (its 1 for X_1, then for one of the next 2, then for one of the next 4, and so on). So for no ω does $X_n(\omega)$ converge to 0 (it always comes back to 1 every so often). Hence $X_n \overset{a.s.}{\not\longrightarrow} 0$.

For many problems, both forms of convergence hold. In addition, for most problems that you would be interested in, convergence in probability is probably adequate.

One useful result involving almost sure convergence is

Theorem. [Strong Law of Large Numbers] *Let $X_1, X_2, \ldots$ be independent identically distribution RVs with $E[X_i] = \mu$ and $E[|X_i|] < \infty$ and define*

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Then $\bar{X}_n \xrightarrow{a.s.} \mu$.

Proof. Omitted $\square$

Remarks:

1. An important part of this theorem is that finite expectation of any RV implies that its sample mean converges almost surely.

2. This result implies that the condition $\mathrm{Var}(X_i) < \infty$ is not needed for the weak law of large numbers to hold.

3. These laws of large numbers give us the third motivation of the definition of expectation: $\mu = E[X]$ is the "fair bet" for playing a game with payoff X. This version strengthens the interpretation of probabilities as long run relative frequencies.

Types of Convergence - Appendix

Let $X_1, X_2, X_3, \ldots$ be a sequence of RVs defined on a sample space Ω, and X be another RV defined on the same sample space Ω. What is the meaning of "the sequence of random variables X_n converging to a random variable X"?

We've seen one type of convergence (in probability). There are others that are used.

First recall that a random variable is a function from Ω to $\mathbb{R}^1$, i.e. $X_n = X_n(\omega), X = X(\omega), \omega \in \Omega$.

When we are talking about convergence of random variables, we are actually talking about events on the sample space Ω.

1. Convergence everywhere

$$X_n \overset{everywhere}{\longrightarrow} X \iff X_n(\omega) \to X(\omega) \text{ for all } \omega \in \Omega$$

2. Convergence almost surely (sometimes called almost everywhere or with probability 1)

$$X_n \xrightarrow{a.s.} X \iff P[\{\omega : X_n(\omega) \to X(\omega)\}] = 1$$

3. Convergence in probability

$$X_n \xrightarrow{P} X \iff \text{For any } \epsilon > 0, P[\{\omega : |X_n(\omega) - X(\omega)| < \epsilon\}] \to 1$$

4. Convergence in mean of order p (L_p convergence)

$$E[|X_n - X|^p] \to 0 \text{ for } p > 0$$

5. Convergence in distribution

   To come later

Notice that the first one, convergence everywhere does not involve probability, so it isn't useful in the course.

Also while these are stated in terms of convergence to random variables, this includes convergence to a constant c as discussed earlier. Just let X be a random variable such that $P[X = c] = 1, P[X \neq c] = 0$.

Similarly to convergence to a constant, in general $Y_n \xrightarrow{a.s.} Y$ does not imply $E[Y_n] \to E[Y]$.

For the means to converge, we need the Y_n's to be suitably bounded. One such way is with the following theorem.

Theorem. [Dominated Convergence] *If* $Y_n \xrightarrow{a.s.} Y$ *and* $|Y_n(\omega)| < X(\omega)$ *for some RV* X *with* $E[X] < \infty$*, then*

$$E[Y] < \infty \text{ and } E[Y_n] \to E[Y]$$

Proof. Omitted $\square$

In general, almost everywhere convergence is much stronger than convergence in probability. The following theorem supports this

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

Proof. Omitted $\square$

The other direction doesn't hold. It is possible to have $X_n \xrightarrow{P} X$ but it will not converge almost everywhere.