

1. Introduction: A Simulation Experiment

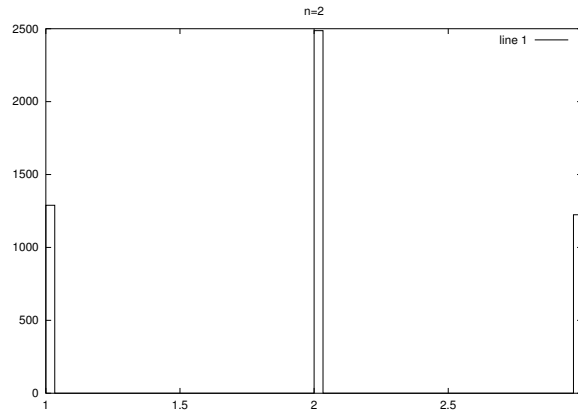
This note focuses on limiting behavior of sequences of random variables, and particularly on the properties of sample averages.

As an informal introduction to some of the key ideas, suppose that for each $i = 1, 2, \dots, n$, the random variables X_i are independent, identically distributed random variables, equal to 1 with probability .5 and equal to 3 with probability .5. (So each X_i has expected value 2.)

Consider the sample average

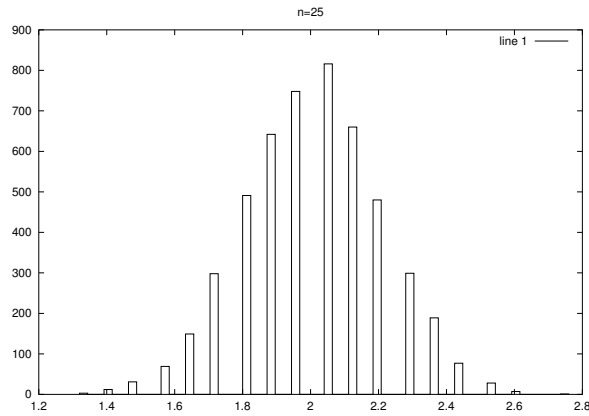
$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

We want to know what the distribution of \bar{X}_n looks like, and how it changes with n . To do this, we can carry out a simulation experiment in Matlab. We draw a sample of size $n = 2$ from the distribution of X_i . We then form the sample mean $\bar{X}_n = \frac{1}{n} \sum_i X_i$. We repeat this 5000 times to approximate the distribution of \bar{X}_n , which looks like this:



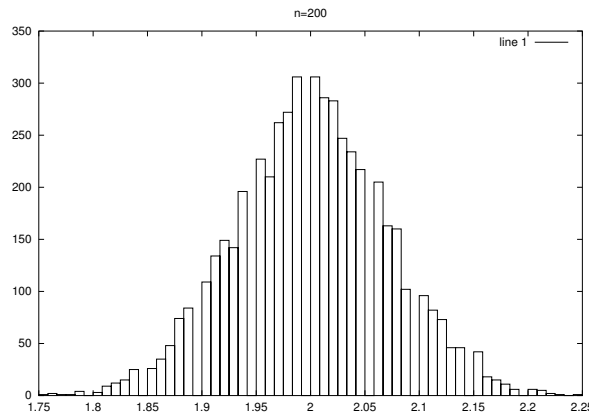
The sample average has a discrete distribution which takes on only 3 possible values: 1,2, and 3. Notice that it is centered at 2, which is not surprising since each X_i has expected value 2.

What happens if we repeat the experiment, but with a larger sample size? Suppose we set $n = 25$. This results in the following distribution of \bar{X}_n :



The distribution is still discrete, but it can take on more possible values and has something of a bell-curve shape. Notice also that the values it takes on tend to be closer to 2.

Going further, we set $n = 200$. Then we get:



As we increase n , we notice two things. First, the distribution becomes tighter around the mean of 2. Second, its distribution looks more and more like a normal distribution!

This illustrates two fundamental results of probability theory:

1. The **law of large numbers** says loosely that as the sample size increases, the sample average gets close, in probability, to the true mean of the observations.
2. The **central limit theorem** says loosely that as the sample size increases, the distribution of the sample average becomes more and more like a normal distribution.

These results are quite remarkable. Depending on the distribution of the individual X_i 's, the "exact" distribution of \bar{X}_n might be very complicated. But if the sample size is reasonably large, the distribution of \bar{X}_n is close to the true mean, and approximately normal, for a very wide range of distributions of X_i . This turns out to be incredibly useful for the later parts of the course, where we try to learn about unknown parameters of a distribution using random samples.

2. Convergence Concepts and the Law of Large Numbers

Approximations that hold as the sample size n becomes large (that is, as $n \rightarrow \infty$), are called *asymptotic approximations*. To develop these approximations formally, let us first go back to a setting with two random variables. Earlier, we looked at the mean and variance of two independent random variables X and Y :

$$E[X + Y] = E[X] + E[Y],$$

and

$$V(X + Y) = V(X) + V(Y).$$

More generally, if X_1, \dots, X_n are n independent random variables, we have mean

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i],$$

and variance

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

Note that we only use the independence for the variance. Without independence the mean is unchanged, but the variance becomes

$$\begin{aligned} V\left(\sum_{i=1}^n X_i\right) &= E\left[\left(\sum_{i=1}^n X_i - E\left[\sum_{i=1}^n X_i\right]\right)^2\right] \\ &= E\left[\left(\sum_{i=1}^n (X_i - E[X_i])\right)^2\right] = E\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - E[X_i]) \cdot (X_j - E[X_j])\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n E\left[(X_i - E[X_i]) \cdot (X_j - E[X_j])\right] \\ &= \sum_{i=1}^n V(X_i) + \sum_{i=1}^n \sum_{j \neq i} C(X_i, X_j). \end{aligned}$$

Independence ensures that all the covariance terms in the double sum are equal to zero.

Now suppose that all random variables are independent and have the same mean μ and variance σ^2 . Then the sum has mean

$$E\left[\sum_{i=1}^n X_i\right] = n \cdot \mu,$$

and variance

$$V\left[\sum_{i=1}^n X_i\right] = n \cdot \sigma^2.$$

The mean and variance of the average are

$$E\left[\sum_{i=1}^n X_i/n\right] = \mu,$$

and

$$V\left[\sum_{i=1}^n X_i/n\right] = n \cdot \sigma^2/n^2 = \sigma^2/n.$$

Now consider the behavior of the average of the first n random variables, $\bar{X}_n = \sum_{i=1}^n X_i/n$, as n gets large. In that case the variance of the sample average gets smaller and smaller. Using Chebyshev's inequality, this implies that the probability that the sample average is more than ε away from μ can be made arbitrarily small by taking n large enough: Fix ε . Then

$$Pr(|\bar{X}_n - \mu| > \varepsilon) < \sigma^2/(n \cdot \varepsilon^2),$$

which can be made arbitrarily small for fixed ε by choosing n large enough. It would seem uncontroversial to say that in this case \bar{X}_n converges to μ .

In other cases this is not so clear. Consider the following sequence of random variables X_1, X_2, \dots with the pdf of X_n equal to

$$f_n(x) = \begin{cases} (n-1)/2 & -1/n < x < 1/n, \\ 1/n & n < x < n+1, \\ 0 & \text{elsewhere} \end{cases}.$$

The mean of X_n is $1 + 1/2n$. The variance increases with n and approaches infinity as n goes to infinity. However, the probability that X_n is more than ε away from zero is at most $1/n$. Does X_n converge to 0? Does X_n converge to its asymptotic mean of 1?

This example demonstrates the need for different concepts of convergence. We consider three such concepts.

Definition 1 A sequence of random variables X_n converges to μ in probability if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} Pr(|X_n - \mu| > \varepsilon) = 0.$$

Definition 2 A sequence of random variables X_n converges to μ in quadratic mean if

$$\lim_{n \rightarrow \infty} E[(X_n - \mu)^2] = 0.$$

Definition 3 A sequence of random variables X_n converges to μ almost surely if for all $\varepsilon > 0$,

$$P\left(\lim_{n \rightarrow \infty} |X_n - \mu| > \varepsilon\right) = 0.$$

The last definition might initially seem a bit mysterious. Recall that a random variable is a function from some sample space Ω into the real line. So we could be more precise and write each X_n as $X_n(\omega)$, to emphasize this fact. So we could write

$$P\left(\lim_{n \rightarrow \infty} |X_n - \mu| > \varepsilon\right) = P\left(\left\{\omega \in \Omega : \lim_{n \rightarrow \infty} |X_n(\omega) - \mu| > \varepsilon\right\}\right).$$

In words: each $\omega \in \Omega$ implies a certain sequence $X_1(\omega), X_2(\omega), X_3(\omega), \dots$. For almost sure convergence to hold, we must have that essentially none of these sequences satisfies $\lim_{n \rightarrow \infty} |X_n(\omega) - \mu| > \varepsilon$.

Note that in the first example the convergence is clearly in quadratic mean and probability. The independence implies it is also convergence almost surely. The following relations hold between the different convergence concepts:

- (i) Convergence in quadratic mean implies convergence in probability.
- (ii) Convergence almost surely implies convergence in probability.
- (iii) Convergence in quadratic mean does not imply, and is not implied by, convergence almost surely.

Let us consider an example of the difference between convergence in quadratic mean and convergence almost surely. Consider the following sequence of random variables, defined as $X_n(\omega)$, for $\omega \in \Omega = [0, 1]$, and with the probability of ω in some interval (a, b) equal to $b - a$ for $0 \leq a \leq b \leq 1$:

$$X_1(\omega) = 1 \text{ for } \omega \in [0, 1] \text{ and zero elsewhere,}$$

$$X_2(\omega) = 1 \text{ for } \omega \in [0, 1/2] \text{ and zero elsewhere,}$$

$$X_3(\omega) = 1 \text{ for } \omega \in [1/2, 1/2 + 1/3] \text{ and zero elsewhere,}$$

$$X_4(\omega) = 1 \text{ for } \omega \in [1/2 + 1/3, 1] \cup [0, 1/12] \text{ and zero elsewhere,}$$

$$X_4(\omega) = 1 \text{ for } \omega \in [1/2 + 1/3, 1] \cup [1/12, 1/12 + 1/5] \text{ and zero elsewhere.}$$

For $X_n(\omega)$ the intervals where $X_n(\omega)$ is equal to one have total length equal to $1/n$, and therefore probability $1/n$. They shift to the right till they hit 1, and then start over again at 0. Clearly the mean of X_n is $p_n = 1/n$, and the variance is $p_n(1 - p_n) = 1/n - 1/n^2$, both of which go to zero, so we have convergence to zero in quadratic mean and probability. Now consider for a particular value of ω the sequence of values $X_1(\omega)$, $X_2(\omega)$. Does this sequence converge? No—no matter how large n , the remaining sum $\sum_{i=n}^{\infty} 1/i$ always diverges, implying that the sequence is always going to return to 1. Hence the probability of an ω such that the limit $X_n(\omega)$ even exists, let alone that it equals zero, is zero, and not one as required by almost sure convergence.

With these convergence concepts we can formulate laws of large numbers.

Result 1 *Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables with common mean μ and variance σ^2 . Let $\bar{X}_n = \sum_{i=1}^n X_i/n$ be the average up to the n th random variable. Then*

$$\bar{X}_n \xrightarrow{qm} \mu.$$

The proof is straightforward: the variance of \bar{X}_n is σ^2/n which goes to zero. The result implies convergence in probability as well, as we already showed using Chebyshev's inequality.

A second result gives an even stronger for almost sure convergence. In this case convergence in quadratic mean does not necessarily hold because the variance does not necessarily exist.

Result 2 *Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables with common mean μ . Let $\bar{X}_n = \sum_{i=1}^n X_i/n$ be the average up to the n th random variable. Then*

$$\bar{X}_n \xrightarrow{as} \mu.$$

Most advanced probability textbooks contain detailed proofs of this type of result. See, for example, Dudley, Theorem 8.3.5.

3. Convergence in Distribution and the Central Limit Theorem

For the next set of results we need an additional mode of convergence:

Definition 4 *A sequence of random variables X_1, X_2, \dots converges in distribution to a random variable Y if at all continuity points of $F_Y(y)$,*

$$\lim_{n \rightarrow \infty} F_{X_n}(y) = F_Y(y).$$

Convergence in distribution is also called weak convergence.

The restriction to continuity points is for the following reason: suppose

$$f_{X_n}(x) = n, \text{ for } 0 < x < 1/n,$$

and zero elsewhere. The value of $F_{X_n}(0)$ is zero for all n , but X_n converges in distribution to a random variable Y with $F_Y(0) = 1 \neq \lim_{n \rightarrow \infty} F_{X_n}(0)$.

If a random variable converges in distribution to a (degenerate) random variable with all mass in a single point μ , the random variable converges to μ in probability, and vice versa.

The following result is useful for proving results about convergence in distribution. This is Theorem 2.3.12 in Casella and Berger:

Result 3 *(CB 2.3.12) Suppose $\{X_i, i = 1, 2, \dots\}$ is a sequence of random variables, each with MGF $M_{X_i}(t)$. Furthermore, suppose that*

$$\lim_{i \rightarrow \infty} M_{X_i}(t) = M_X(t),$$

for all t in a neighborhood of 0, and $M_X(t)$ is an MGF. Then there is a unique CDF F_X whose moments are determined by $M_X(t)$, and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i \rightarrow \infty} F_{X_i}(x) = F_X(x).$$

This says that convergence of MGFs for $|t| < h$, implies convergence of CDFs and hence convergence in distribution.

This result is useful for proving one of the most important theorems in probability theory, the central limit theorem:

Result 4 *Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables with twice differentiable moment generating function $M_X(t)$ in a neighbourhood of zero, and with mean μ and variance σ^2 . Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) / \sigma = \sqrt{n} \cdot (\bar{X}_n - \mu) / \sigma \xrightarrow{d} \mathcal{N}(0, 1).$$

That is, the normalized sum converges to a standard normal distribution. The proof below exploits the existence of the moment generating function, although this is not necessary: the result also holds with just the existence of the mean and variance (using characteristic functions). Also we do not need identically distributed random variables, or even independent ones. These assumptions can all be weakened at some expense.

proof:

Define $Y_i = (X_i - \mu)/\sigma$. Then $M'_Y(0) = E[Y] = 0$, and $M''_Y(0) = V(Y) + E[Y]^2 = 1$. Also,

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(X_i - \mu \right) / \sigma = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i.$$

The moment generating function of Z_n is

$$M_{Z_n}(t) = E \exp \left(t \sum_{i=1}^n Y_i / \sqrt{n} \right) = E \left[\prod_{i=1}^n \exp(t Y_i / \sqrt{n}) \right].$$

By independence this is equal to

$$\prod_{i=1}^n E[\exp(t Y_i / \sqrt{n})] = \prod_{i=1}^n M_{Y_i}(t / \sqrt{n}) = M_Y(t / \sqrt{n})^n.$$

Using a Taylor series expansion for $M_Y(t / \sqrt{n})$ around zero we get

$$M_Y(t / \sqrt{n}) = M_Y(0) + M'_Y(0)t / \sqrt{n} + M''_Y(\tilde{t})t^2 / (2n),$$

for some \tilde{t} between t / \sqrt{n} and zero. Because $M_Y(0) = 1$ and $M'_Y(0) = 0$, we have

$$M_Y(t / \sqrt{n})^n = (1 + M''_Y(\tilde{t})t^2 / (2n))^n.$$

Taking the limit as $n \rightarrow \infty$ we get $\tilde{t} \rightarrow 0$, and therefore $M''_Y(\tilde{t}) \rightarrow M''_Y(0) = 1$, and

$$\lim_{n \rightarrow \infty} M_Y(t / \sqrt{n})^n = \exp(t^2 / 2).$$

This is the moment generating function for a standard normal random variable, which completes the proof.

The next set of results are known as Slutsky's theorem:

Result 5 *If X_n converges in distribution to X , and Y_n converges in probability to a constant c , then*

$$X_n \cdot Y_n \xrightarrow{d} c \cdot X,$$

and

$$X_n + Y_n \xrightarrow{d} c + X.$$

If in addition $c \neq 0$,

$$X_n / Y_n \xrightarrow{d} X / c.$$

The final result is known as the delta method:

Result 6 *If a sequence of random variables X_n satisfies*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

then for any function $g(\cdot)$ continuously differentiable in a neighbourhood of μ with derivative $g'(\mu)$,

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2 \cdot \sigma^2).$$

Proof:

First, it is clear that $X_n \rightarrow \mu$ in probability. If the probability that $|X_n - \mu| > \varepsilon$ can be made arbitrarily small, it must be the case that the probability that $|g(X_n) - g(\mu)| > \varepsilon$ can be made arbitrarily small, and hence $g(X_n) \rightarrow g(\mu)$ in probability. Now linearize $g(X_n)$ around μ to get

$$g(X_n) = g(\mu) + (X_n - \mu) \cdot g'(\tilde{X}_n).$$

Clearly as $n \rightarrow \infty$, $g'(\tilde{X}_n) \rightarrow g'(\mu)$, and $g(X_n) - g(\mu) \approx (X_n - \mu) \cdot g'(\mu)$. \square

Consider the following example. Using a central limit theorem we might find that

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Then we can use the delta method to conclude that

$$\sqrt{n}((\bar{X})^2 - \mu^2) \xrightarrow{d} \mathcal{N}(0, 4\mu^2\sigma^2).$$