7

Limit Theorems

Contents

Consider a sequence $X_1, X_2, \ldots$ of independent identically distributed random variables with mean μ and variance σ^2. Let

$$S_n = X_1 + \cdots + X_n$$

be the sum of the first n of them. Limit theorems are mostly concerned with the properties of S_n and related random variables, as n becomes very large.

Because of independence, we have

$$\text{var}(S_n) = \text{var}(X_1) + \cdots + \text{var}(X_n) = n\sigma^2.$$

Thus, the distribution of S_n spreads out as n increases, and does not have a meaningful limit. The situation is different if we consider the **sample mean**

$$M_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

A quick calculation yields

$$\mathbf{E}[M_n] = \mu, \qquad \text{var}(M_n) = \frac{\sigma^2}{n}.$$

In particular, the variance of M_n decreases to zero as n increases, and the bulk of its distribution must be very close to the mean μ. This phenomenon is the subject of certain laws of large numbers, which generally assert that the sample mean M_n (a random variable) converges to the true mean μ (a number), in a precise sense. These laws provide a mathematical basis for the loose interpretation of an expectation $\mathbf{E}[X] = \mu$ as the average of a large number of independent samples drawn from the distribution of X.

We will also consider a quantity which is intermediate between S_n and M_n. We first subtract $n\mu$ from S_n, to obtain the zero-mean random variable $S_n - n\mu$ and then divide by $\sigma\sqrt{n}$, to obtain

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

It can be verified (see Section 7.4) that

$$\mathbf{E}[Z_n] = 0, \qquad \text{var}(Z_n) = 1.$$

Since the mean and the variance of Z_n remain unchanged as n increases, its distribution neither spreads, nor shrinks to a point. The **central limit theorem** is concerned with the asymptotic shape of the distribution of Z_n and asserts that it becomes the standard normal distribution.

Limit theorems are useful for several reasons:

(a) Conceptually, they provide an interpretation of expectations (as well as probabilities) in terms of a long sequence of identical independent experiments.

(b) They allow for an approximate analysis of the properties of random variables such as S_n. This is to be contrasted with an exact analysis which would require a formula for the PMF or PDF of S_n, a complicated and tedious task when n is large.

7.1 SOME USEFUL INEQUALITIES

In this section, we derive some important inequalities. These inequalities use the mean, and possibly the variance, of a random variable to draw conclusions on the probabilities of certain events. They are primarily useful in situations where the mean and variance of a random variable X are easily computable, but the distribution of X is either unavailable or hard to calculate.

We first present the **Markov inequality**. Loosely speaking it asserts that if a *nonnegative* random variable has a small mean, then the probability that it takes a large value must also be small.

Markov Inequality

If a random variable X can only take nonnegative values, then

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}[X]}{a}, \qquad \text{for all } a > 0.$$

To justify the Markov inequality, let us fix a positive number a and consider the random variable Y_a defined by

$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

It is seen that the relation

$$Y_a \leq X$$

always holds and therefore,

$$\mathbf{E}[Y_a] \leq \mathbf{E}[X].$$

On the other hand,

$$\mathbf{E}[Y_a] = a\mathbf{P}(Y_a = a) = a\mathbf{P}(X \geq a),$$

from which we obtain

$$a\mathbf{P}(X \geq a) \leq \mathbf{E}[X].$$

Example 7.1. Let X be uniformly distributed on the interval $[0, 4]$ and note that $\mathbf{E}[X] = 2$. Then, the Markov inequality asserts that

$$\mathbf{P}(X \geq 2) \leq \frac{2}{2} = 1, \qquad \mathbf{P}(X \geq 3) \leq \frac{2}{3} = 0.67, \qquad \mathbf{P}(X \geq 4) \leq \frac{2}{4} = 0.5.$$

By comparing with the exact probabilities

$$\mathbf{P}(X \geq 2) = 0.5, \qquad \mathbf{P}(X \geq 3) = 0.25, \qquad \mathbf{P}(X \geq 4) = 0,$$

we see that the bounds provided by the Markov inequality can be quite loose.

We continue with the **Chebyshev inequality**. Loosely speaking, it asserts that if the variance of a random variable is small, then the probability that it takes a value far from its mean is also small. Note that the Chebyshev inequality does not require the random variable to be nonnegative.

Chebyshev Inequality

If X is a random variable with mean μ and variance σ^2, then

$$\mathbf{P}\big(|X - \mu| \geq c\big) \leq \frac{\sigma^2}{c^2}, \qquad \text{for all } c > 0.$$

To justify the Chebyshev inequality, we consider the nonnegative random variable $(X - \mu)^2$ and apply the Markov inequality with $a = c^2$. We obtain

$$\mathbf{P}\big((X - \mu)^2 \geq c^2\big) \leq \frac{\mathbf{E}\big[(X - \mu)^2\big]}{c^2} = \frac{\sigma^2}{c^2}.$$

The derivation is completed by observing that the event $(X-\mu)^2 \geq c^2$ is identical to the event $|X - \mu| \geq c$ and

$$\mathbf{P}\big(|X - \mu| \geq c\big) = \mathbf{P}\big((X - \mu)^2 \geq c^2\big) \leq \frac{\sigma^2}{c^2}.$$

An alternative form of the Chebyshev inequality is obtained by letting $c = k\sigma$, where k is positive, which yields

$$\mathbf{P}\big(|X - \mu| \geq k\sigma\big) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Thus, the probability that a random variable takes a value more than k standard deviations away from its mean is at most $1/k^2$.

The Chebyshev inequality is generally more powerful than the Markov inequality (the bounds that it provides are more accurate), because it also makes use of information on the variance of X. Still, the mean and the variance of a random variable are only a rough summary of the properties of its distribution, and we cannot expect the bounds to be close approximations of the exact probabilities.

Example 7.2. As in Example 7.1, let X be uniformly distributed on $[0, 4]$. Let us use the Chebyshev inequality to bound the probability that $|X - 2| \geq 1$. We have $\sigma^2 = 16/12 = 4/3$, and

$$\mathbf{P}\big(|X - 2| \geq 1\big) \leq \frac{4}{3},$$

which is not particularly informative.

For another example, let X be exponentially distributed with parameter $\lambda = 1$, so that $\mathbf{E}[X] = \text{var}(X) = 1$. For $c > 1$, using Chebyshev's inequality, we obtain

$$\mathbf{P}(X \geq c) = \mathbf{P}(X - 1 \geq c - 1) \leq \mathbf{P}\big(|X - 1| \geq c - 1\big) \leq \frac{1}{(c-1)^2}.$$

This is again conservative compared to the exact answer $\mathbf{P}(X \geq c) = e^{-c}$.

7.2 THE WEAK LAW OF LARGE NUMBERS

The weak law of large numbers asserts that the sample mean of a large number of independent identically distributed random variables is very close to the true mean, with high probability.

As in the introduction to this chapter, we consider a sequence $X_1, X_2, \ldots$ of independent identically distributed random variables with mean μ and variance σ^2, and define the sample mean by

$$M_n = \frac{X_1 + \cdots + X_n}{n}.$$

We have

$$\mathbf{E}[M_n] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \frac{n\mu}{n} = \mu,$$

and, using independence,

$$\text{var}(M_n) = \frac{\text{var}(X_1 + \cdots + X_n)}{n^2} = \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

We apply Chebyshev's inequality and obtain

$$\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) \leq \frac{\sigma^2}{n\epsilon^2}, \qquad \text{for any } \epsilon > 0.$$

We observe that for any fixed $\epsilon > 0$, the right-hand side of this inequality goes to zero as n increases. As a consequence, we obtain the weak law of large numbers, which is stated below. It turns out that this law remains true even if the X_i

have infinite variance, but a much more elaborate argument is needed, which we omit. The only assumption needed is that $\mathbf{E}[X_i]$ is well-defined and finite.

The Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ. For every $\epsilon > 0$, we have

$$\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big) = \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0, \qquad \text{as } n \to \infty.$$

The WLLN states that for large n, the "bulk" of the distribution of M_n is concentrated near μ. That is, if we consider a positive length interval $[\mu-\epsilon, \mu+\epsilon]$ around μ, then there is high probability that M_n will fall in that interval; as $n \to \infty$, this probability converges to 1. Of course, if ϵ is very small, we may have to wait longer (i.e., need a larger value of n) before we can assert that M_n is highly likely to fall in that interval.

Example 7.3. Probabilities and Frequencies. Consider an event A defined in the context of some probabilistic experiment. Let $p = \mathbf{P}(A)$ be the probability of that event. We consider n independent repetitions of the experiment, and let M_n be the fraction of time that event A occurred; in this context, M_n is often called the **empirical frequency** of A. Note that

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where X_i is 1 whenever A occurs, and 0 otherwise; in particular, $\mathbf{E}[X_i] = p$. The weak law applies and shows that when n is large, the empirical frequency is most likely to be within ϵ of p. Loosely speaking, this allows us to say that empirical frequencies are faithful estimates of p. Alternatively, this is a step towards interpreting the probability p as the frequency of occurrence of A.

Example 7.4. Polling. Let p be the fraction of voters who support a particular candidate for office. We interview n "randomly selected" voters and record the fraction M_n of them that support the candidate. We view M_n as our estimate of p and would like to investigate its properties.

We interpret "randomly selected" to mean that the n voters are chosen independently and uniformly from the given population. Thus, the reply of each person interviewed can be viewed as an independent Bernoulli trial X_i with success probability p and variance $\sigma^2 = p(1-p)$. The Chebyshev inequality yields

$$\mathbf{P}\big(|M_n - p| \geq \epsilon\big) \leq \frac{p(1-p)}{n\epsilon^2}.$$

The true value of the parameter p is assumed to be unknown. On the other hand, it is easily verified that $p(1-p) \leq 1/4$, which yields

$$\mathbf{P}\big(|M_n - p| \geq \epsilon\big) \leq \frac{1}{4n\epsilon^2}.$$

For example, if $\epsilon = 0.1$ and $n = 100$, we obtain

$$\mathbf{P}\big(|M_{100} - p| \geq 0.1\big) \leq \frac{1}{4 \cdot 100 \cdot (0.1)^2} = 0.25.$$

In words, with a sample size of $n = 100$, the probability that our estimate is wrong by more than 0.1 is no larger than 0.25.

Suppose now that we impose some tight specifications on our poll. We would like to have high confidence (probability at least 95%) that our estimate will be very accurate (within .01 of p). How many voters should be sampled?

The only guarantee that we have at this point is the inequality

$$\mathbf{P}\big(|M_n - p| \geq 0.01\big) \leq \frac{1}{4n(0.01)^2}.$$

We will be sure to satisfy the above specifications if we choose n large enough so that

$$\frac{1}{4n(0.01)^2} \leq 1 - 0.95 = 0.05,$$

which yields $n \geq 50,000$. This choice of n has the specified properties but is actually fairly conservative, because it is based on the rather loose Chebyshev inequality. A refinement will be considered in Section 7.4.

7.3 CONVERGENCE IN PROBABILITY

We can interpret the WLLN as stating that "M_n converges to μ." However, since $M_1, M_2, \ldots$ is a sequence of random variables, not a sequence of numbers, the meaning of convergence has to be made precise. A particular definition is provided below. To facilitate the comparison with the ordinary notion of convergence, we also include the definition of the latter.

Convergence of a Deterministic Sequence

Let $a_1, a_2, \ldots$ be a sequence of real numbers, and let a be another real number. We say that the sequence a_n converges to a, or $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon, \qquad \text{for all } n \geq n_0.$$

Intuitively, for any given accuracy level ϵ, a_n must be within ϵ of a, when n is large enough.

Convergence in Probability

Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to a in probability**, if for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n - a| \geq \epsilon\big) = 0.$$

Given this definition, the WLLN simply says that the sample mean converges in probability to the true mean μ.

If the random variables $Y_1, Y_2, \ldots$ have a PMF or a PDF and converge in probability to a, then according to the above definition, "almost all" of the PMF or PDF of Y_n is concentrated to within a an ϵ-interval around a for large values of n. It is also instructive to rephrase the above definition as follows: for every $\epsilon > 0$, and for every $\delta > 0$, there exists some n_0 such that

$$\mathbf{P}\big(|Y_n - a| \geq \epsilon\big) \leq \delta, \qquad \text{for all } n \geq n_0.$$

If we refer to ϵ as the *accuracy* level, and δ as the *confidence* level, the definition takes the following intuitive form: for any given level of accuracy and confidence, Y_n will be equal to a, within these levels of accuracy and confidence, provided that n is large enough.

Example 7.5. Consider a sequence of independent random variables X_n that are uniformly distributed over the interval $[0, 1]$, and let

$$Y_n = \min\{X_1, \ldots, X_n\}.$$

The sequence of values of Y_n cannot increase as n increases, and it will occasionally decrease (when a value of X_n that is smaller than the preceding values is obtained). Thus, we intuitively expect that Y_n converges to zero. Indeed, for $\epsilon > 0$, we have using the independence of the X_n,

$$\begin{aligned}\mathbf{P}\big(|Y_n - 0| \geq \epsilon\big) &= \mathbf{P}(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) \\ &= \mathbf{P}(X_1 \geq \epsilon) \cdots \mathbf{P}(X_n \geq \epsilon) \\ &= (1 - \epsilon)^n.\end{aligned}$$

Since this is true for every $\epsilon > 0$, we conclude that Y_n converges to zero, in probability.

Example 7.6. Let Y be an exponentially distributed random variable with parameter $\lambda = 1$. For any positive integer n, let $Y_n = Y/n$. (Note that these random variables are dependent.) We wish to investigate whether the sequence Y_n converges to zero.

For $\epsilon > 0$, we have

$$\mathbf{P}\big(|Y_n - 0| \geq \epsilon\big) = \mathbf{P}(Y_n \geq \epsilon) = \mathbf{P}(Y \geq n\epsilon) = e^{-n\epsilon}.$$

In particular,

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n - 0| \geq \epsilon\big) = \lim_{n\to\infty} e^{-n\epsilon} = 0.$$

Since this is the case for every $\epsilon > 0$, Y_n converges to zero, in probability.

One might be tempted to believe that if a sequence Y_n converges to a number a, then $\mathbf{E}[Y_n]$ must also converge to a. The following example shows that this need not be the case.

Example 7.7. Consider a sequence of discrete random variables Y_n with the following distribution:

$$\mathbf{P}(Y_n = y) = \begin{cases} 1 - \dfrac{1}{n}, & \text{for } y = 0, \\ \dfrac{1}{n}, & \text{for } y = n^2, \\ 0, & \text{elsewhere.} \end{cases}$$

For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbf{P}\big(|Y_n| \geq \epsilon\big) = \lim_{n\to\infty} \frac{1}{n} = 0,$$

and Y_n converges to zero in probability. On the other hand, $\mathbf{E}[Y_n] = n^2/n = n$, which goes to infinity as n increases.

7.4 THE CENTRAL LIMIT THEOREM

According to the weak law of large numbers, the distribution of the sample mean M_n is increasingly concentrated in the near vicinity of the true mean μ. In particular, its variance tends to zero. On the other hand, the variance of the sum $S_n = X_1 + \cdots + X_n = nM_n$ increases to infinity, and the distribution of S_n cannot be said to converge to anything meaningful. An intermediate view is obtained by considering the deviation $S_n - n\mu$ of S_n from its mean $n\mu$, and scaling it by a factor proportional to $1/\sqrt{n}$. What is special about this particular scaling is that it keeps the variance at a constant level. The central limit theorem

asserts that the distribution of this scaled random variable approaches a normal distribution.

More specifically, let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2. We define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

An easy calculation yields

$$\mathbf{E}[Z_n] = \frac{\mathbf{E}[X_1 + \cdots + X_n] - n\mu}{\sigma\sqrt{n}} = 0,$$

and

$$\mathrm{var}(Z_n) = \frac{\mathrm{var}(X_1 + \cdots + X_n)}{\sigma^2 n} = \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1.$$

The Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with common mean μ and variance σ^2, and define

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Then, the CDF of Z_n converges to the standard normal CDF

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

in the sense that

$$\lim_{n\to\infty} \mathbf{P}(Z_n \leq z) = \Phi(z), \qquad \text{for every } z.$$

The central limit theorem is surprisingly general. Besides independence, and the implicit assumption that the mean and variance are well-defined and finite, it places no other requirement on the distribution of the X_i, which could be discrete, continuous, or mixed random variables. It is of tremendous importance for several reasons, both conceptual, as well as practical. On the conceptual side, it indicates that the sum of a large number of independent random variables is approximately normal. As such, it applies to many situations in which a random effect is the sum of a large number of small but independent random

factors. Noise in many natural or engineered systems has this property. In a wide array of contexts, it has been found empirically that the statistics of noise are well-described by normal distributions, and the central limit theorem provides a convincing explanation for this phenomenon.

On the practical side, the central limit theorem eliminates the need for detailed probabilistic models and for tedious manipulations of PMFs and PDFs. Rather, it allows the calculation of certain probabilities by simply referring to the normal CDF table. Furthermore, these calculations only require the knowledge of means and variances.

Approximations Based on the Central Limit Theorem

The central limit theorem allows us to calculate probabilities related to Z_n as if Z_n were normal. Since normality is preserved under linear transformations, this is equivalent to treating S_n as a normal random variable with mean $n\mu$ and variance $n\sigma^2$.

Normal Approximation Based on the Central Limit Theorem

Let $S_n = X_1 + \cdots + X_n$, where the X_i are independent identically distributed random variables with mean μ and variance σ^2. If n is large, the probability $\mathbf{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure.

1. Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n.
2. Calculate the normalized value $z = (c - n\mu)/\sigma\sqrt{n}$.
3. Use the approximation

$$\mathbf{P}(S_n \leq c) \approx \Phi(z),$$

where $\Phi(z)$ is available from standard normal CDF tables.

Example 7.8. We load on a plane 100 packages whose weights are independent random variables that are uniformly distributed between 5 and 50 pounds. What is the probability that the total weight will exceed 3000 pounds? It is not easy to calculate the CDF of the total weight and the desired probability, but an approximate answer can be quickly obtained using the central limit theorem.

We want to calculate $\mathbf{P}(S_{100} > 3000)$, where S_{100} is the sum of the 100 packages. The mean and the variance of the weight of a single package are

$$\mu = \frac{5+50}{2} = 27.5, \qquad \sigma^2 = \frac{(50-5)^2}{12} = 168.75,$$

based on the formulas for the mean and variance of the uniform PDF. We thus calculate the normalized value

$$z = \frac{3000 - 100 \cdot 27.5}{\sqrt{168.75 \cdot 100}} = \frac{250}{129.9} = 1.92,$$

and use the standard normal tables to obtain the approximation

$$\mathbf{P}(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726.$$

Thus the desired probability is

$$\mathbf{P}(S_{100} > 3000) = 1 - \mathbf{P}(S_{100} \leq 3000) \approx 1 - 0.9726 = 0.0274.$$

Example 7.9. A machine processes parts, one at a time. The processing times of different parts are independent random variables, uniformly distributed on $[1, 5]$. We wish to approximate the probability that the number of parts processed within 320 time units is at least 100.

Let us call N_{320} this number. We want to calculate $\mathbf{P}(N_{320} \geq 100)$. There is no obvious way of expressing the random variable N_{320} as the sum of independent random variables, but we can proceed differently. Let X_i be the processing time of the ith part, and let $S_{100} = X_1 + \cdots + X_{100}$ be the total processing time of the first 100 parts. The event $\{N_{320} \geq 100\}$ is the same as the event $\{S_{100} \leq 320\}$, and we can now use a normal approximation to the distribution of S_{100}. Note that $\mu = \mathbf{E}[X_i] = 3$ and $\sigma^2 = \text{var}(X_i) = 16/12 = 4/3$. We calculate the normalized value

$$z = \frac{320 - n\mu}{\sigma\sqrt{n}} = \frac{320 - 300}{\sqrt{100 \cdot 4/3}} = 1.73,$$

and use the approximation

$$\mathbf{P}(S_{100} \leq 320) \approx \Phi(1.73) = 0.9582.$$

If the variance of the X_i is unknown, but an upper bound is available, the normal approximation can be used to obtain bounds on the probabilities of interest.

Example 7.10. Let us revisit the polling problem in Example 7.4. We poll n voters and record the fraction M_n of those polled who are in favor of a particular candidate. If p is the fraction of the entire voter population that supports this candidate, then

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where the X_i are independent Bernoulli random variables with parameter p. In particular, M_n has mean p and variance $p(1-p)/n$. By the normal approximation,

$X_1 + \cdots + X_n$ is approximately normal, and therefore M_n is also approximately normal.

We are interested in the probability $\mathbf{P}\big(|M_n - p| \geq \epsilon\big)$ that the polling error is larger than some desired accuracy ϵ. Because of the symmetry of the normal PDF around the mean, we have

$$\mathbf{P}\big(|M_n - p| \geq \epsilon\big) \approx 2\mathbf{P}(M_n - p \geq \epsilon).$$

The variance $p(1-p)/n$ of $M_n - p$ depends on p and is therefore unknown. We note that the probability of a large deviation from the mean increases with the variance. Thus, we can obtain an upper bound on $\mathbf{P}\big(M_n - p \geq \epsilon\big)$ by assuming that $M_n - p$ has the largest possible variance, namely, $1/4n$. To calculate this upper bound, we evaluate the standardized value

$$z = \frac{\epsilon}{1/(2\sqrt{n})},$$

and use the normal approximation

$$\mathbf{P}\big(M_n - p \geq \epsilon\big) \leq 1 - \Phi(z) = 1 - \Phi\big(2\epsilon\sqrt{n}\big).$$

For instance, consider the case where $n = 100$ and $\epsilon = 0.1$. Assuming the worst-case variance, we obtain

$$\begin{aligned}\mathbf{P}\big(|M_{100} - p| \geq 0.1\big) &\approx 2\mathbf{P}(M_n - p \geq 0.1)\\ &\leq 2 - 2\Phi\big(2 \cdot 0.1 \cdot \sqrt{100}\big) = 2 - 2\Phi(2) = 2 - 2 \cdot 0.977 = 0.046.\end{aligned}$$

This is much smaller (more accurate) than the estimate that was obtained in Example 7.4 using the Chebyshev inequality.

We now consider a reverse problem. How large a sample size n is needed if we wish our estimate M_n to be within 0.01 of p with probability at least 0.95? Assuming again the worst possible variance, we are led to the condition

$$2 - 2\Phi\big(2 \cdot 0.01 \cdot \sqrt{n}\big) \leq 0.05,$$

or

$$\Phi\big(2 \cdot 0.01 \cdot \sqrt{n}\big) \geq 0.975.$$

From the normal tables, we see that $\Phi(1.96) = 0.975$, which leads to

$$2 \cdot 0.01 \cdot \sqrt{n} \geq 1.96,$$

or

$$n \geq \frac{(1.96)^2}{4 \cdot (0.01)^2} = 9604.$$

This is significantly better than the sample size of 50,000 that we found using Chebyshev's inequality.

The normal approximation is increasingly accurate as n tends to infinity, but in practice we are generally faced with specific and finite values of n. It

would be useful to know how large an n is needed before the approximation can be trusted, but there are no simple and general guidelines. Much depends on whether the distribution of the X_i is close to normal to start with and, in particular, whether it is symmetric. For example, if the X_i are uniform, then S_8 is already very close to normal. But if the X_i are, say, exponential, a significantly larger n will be needed before the distribution of S_n is close to a normal one. Furthermore, the normal approximation to $\mathbf{P}(S_n \leq c)$ is generally more faithful when c is in the vicinity of the mean of S_n.

The De Moivre – Laplace Approximation to the Binomial

A binomial random variable S_n with parameters n and p can be viewed as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$, with common parameter p:

$$S_n = X_1 + \cdots + X_n.$$

Recall that

$$\mu = \mathbf{E}[X_i] = p, \qquad \sigma = \sqrt{\mathrm{var}(X_i)} = \sqrt{p(1-p)},$$

We will now use the approximation suggested by the central limit theorem to provide an approximation for the probability of the event $\{k \leq S_n \leq \ell\}$, where k and ℓ are given integers. We express the event of interest in terms of a standardized random variable, using the equivalence

$$k \leq S_n \leq \ell \qquad \Longleftrightarrow \qquad \frac{k-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}}.$$

By the central limit theorem, $(S_n - np)/\sqrt{np(1-p)}$ has approximately a standard normal distribution, and we obtain

$$\begin{aligned} \mathbf{P}(k \leq S_n \leq \ell) =& \mathbf{P}\left(\frac{k-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}} \right) \\ \approx& \Phi\left(\frac{\ell - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}} \right). \end{aligned}$$

An approximation of this form is equivalent to treating S_n as a normal random variable with mean np and variance $np(1-p)$. Figure 7.1 provides an illustration and indicates that a more accurate approximation may be possible if we replace k and ℓ by $k - \frac{1}{2}$ and $\ell + \frac{1}{2}$, respectively. The corresponding formula is given below.

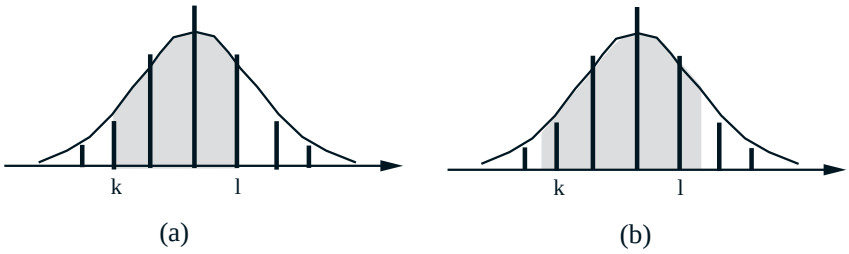

Figure 7.1: The central limit approximation treats a binomial random variable S_n as if it were normal with mean np and variance $np(1-p)$. This figure shows a binomial PMF together with the approximating normal PDF. (a) A first approximation of a binomial probability $\mathbf{P}(k \leq S_n \leq \ell)$ is obtained by integrating the area under the normal PDF from k to ℓ, which is the shaded area in the figure. (b) With the approach in (a), if we have $k = \ell$, the probability $\mathbf{P}(S_n = k)$ would be approximated by zero. A potential remedy would be to use the normal probability between $k - \frac{1}{2}$ and $k + \frac{1}{2}$ to approximate $\mathbf{P}(S_n = k)$. By extending this idea, $\mathbf{P}(k \leq S_n \leq \ell)$ can be approximated by using the area under the normal PDF from $k - \frac{1}{2}$ to $\ell + \frac{1}{2}$, which corresponds to the shaded area.

De Moivre – Laplace Approximation to the Binomial

If S_n is a binomial random variable with parameters n and p, n is large, and k, ℓ are nonnegative integers, then

$$\mathbf{P}(k \leq S_n \leq \ell) \approx \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

Example 7.11. Let S_n be a binomial random variable with parameters $n = 36$ and $p = 0.5$. An exact calculation yields

$$\mathbf{P}(S_n \leq 21) = \sum_{k=0}^{21} \binom{36}{k} (0.5)^{36} = 0.8785.$$

The central limit approximation, without the above discussed refinement, yields

$$\mathbf{P}(S_n \leq 21) \approx \Phi\left(\frac{21 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21 - 18}{3}\right) = \Phi(1) = 0.8413.$$

Using the proposed refinement, we have

$$\mathbf{P}(S_n \leq 21) \approx \Phi\left(\frac{21.5 - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{21.5 - 18}{3}\right) = \Phi(1.17) = 0.879,$$

which is much closer to the exact value.

The de Moivre – Laplace formula also allows us to approximate the probability of a single value. For example,

$$\mathbf{P}(S_n = 19) \approx \Phi\left(\frac{19.5 - 18}{3}\right) - \Phi\left(\frac{18.5 - 18}{3}\right) = 0.6915 - 05675 = 0.124.$$

This is very close to the exact value which is

$$\binom{36}{19}(0.5)^{36} = 0.1251.$$

7.5 THE STRONG LAW OF LARGE NUMBERS

The strong law of large numbers is similar to the weak law in that it also deals with the convergence of the sample mean to the true mean. It is different, however, because it refers to another type of convergence.

The Strong Law of Large Numbers (SLLN)

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with mean μ. Then, the sequence of sample means $M_n = (X_1 + \cdots + X_n)/n$ converges to μ, *with probability 1*, in the sense that

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

In order to interpret the SSLN, we need to go back to our original description of probabilistic models in terms of sample spaces. The contemplated experiment is infinitely long and generates experimental values for each one of the random variables in the sequence $X_1, X_2, \ldots$. Thus, it is best to think of the sample space Ω as a set of infinite sequences $\omega = (x_1, x_2, \ldots)$ of real numbers: any such sequence is a possible outcome of the experiment. Let us now define the subset A of Ω consisting of those sequences $(x_1, x_2, \ldots)$ whose long-term average is μ, i.e.,

$$(x_1, x_2, \ldots) \in A \qquad \Longleftrightarrow \qquad \lim_{n\to\infty} \frac{x_1 + \cdots + x_n}{n} = \mu.$$

The SLLN states that all of the probability is concentrated on this particular subset of Ω. Equivalently, the collection of outcomes that do not belong to A (infinite sequences whose long-term average is not μ) has probability zero.

The difference between the weak and the strong law is subtle and deserves close scrutiny. The weak law states that the probability $\mathbf{P}\big(|M_n - \mu| \geq \epsilon\big)$ of a significant deviation of M_n from μ goes to zero as $n \to \infty$. Still, for any finite n, this probability can be positive and it is conceivable that once in a while, even if infrequently, M_n deviates significantly from μ. The weak law provides no conclusive information on the number of such deviations, but the strong law does. According to the strong law, and with probability 1, M_n converges to μ. This implies that for any given $\epsilon > 0$, the difference $|M_n - \mu|$ will exceed ϵ only a finite number of times.

Example 7.12. Probabilities and Frequencies. As in Example 7.3, consider an event A defined in terms of some probabilistic experiment. We consider a sequence of independent repetitions of the same experiment, and let M_n be the fraction of the first n trials in which A occurs. The strong law of large numbers asserts that M_n converges to $\mathbf{P}(A)$, with probability 1.

We have often talked intuitively about the probability of an event A as the frequency with which it occurs in an infinitely long sequence of independent trials. The strong law backs this intuition and establishes that the long-term frequency of occurrence of A is indeed equal to $\mathbf{P}(A)$, with certainty (the probability of this happening is 1).

Convergence with Probability 1

The convergence concept behind the strong law is different than the notion employed in the weak law. We provide here a definition and some discussion of this new convergence concept.

Convergence with Probability 1

Let $Y_1, Y_2, \ldots$ be a sequence of random variables (not necessarily independent) associated with the same probability model. Let c be a real number. We say that Y_n converges to c **with probability 1** (or **almost surely**) if

$$\mathbf{P}\left(\lim_{n\to\infty} Y_n = c\right) = 1.$$

Similar to our earlier discussion, the right way of interpreting this type of convergence is in terms of a sample space consisting of infinite sequences: all of the probability is concentrated on those sequences that converge to c. This does not mean that other sequences are impossible, only that they are extremely unlikely, in the sense that their total probability is zero.

The example below illustrates the difference between convergence in probability and convergence with probability 1.

Example 7.13. Consider a discrete-time arrival process. The set of times is partitioned into consecutive intervals of the form $I_k = \{2^k, 2^k + 1, \ldots, 2^{k+1} - 1\}$. Note that the length of I_k is 2^k, which increases with k. During each interval I_k, there is exactly one arrival, and all times within an interval are equally likely. The arrival times within different intervals are assumed to be independent. Let us define $Y_n = 1$ if there is an arrival at time n, and $Y_n = 0$ if there is no arrival.

We have $\mathbf{P}(Y_n \neq 0) = 1/2^k$, if $n \in I_k$. Note that as n increases, it belongs to intervals I_k with increasingly large indices k. Consequently,

$$\lim_{n\to\infty} \mathbf{P}(Y_n \neq 0) = \lim_{k\to\infty} \frac{1}{2^k} = 0,$$

and we conclude that Y_n converges to 0 in probability. However, when we carry out the experiment, the total number of arrivals is infinite (one arrival during each interval I_k). Therefore, Y_n is unity for infinitely many values of n, the event $\{\lim_{n\to\infty} Y_n = 0\}$ has zero probability, and we do not have convergence with probability 1.

Intuitively, the following is happening. At any given time, there is a small (and diminishing with n) probability of a substantial deviation from 0 (convergence in probability). On the other hand, given enough time, a substantial deviation from 0 is certain to occur, and for this reason, we do not have convergence with probability 1.

Example 7.14. Let $X_1, X_2, \ldots$ be a sequence of independent random variables that are uniformly distributed on $[0, 1]$, and let $Y_n = \min\{X_1, \ldots, X_n\}$. We wish to show that Y_n converges to 0, with probability 1.

In any execution of the experiment, the sequence Y_n is nonincreasing, i.e., $Y_{n+1} \leq Y_n$ for all n. Since this sequence is bounded below by zero, it must have a limit, which we denote by Y. Let us fix some $\epsilon > 0$. If $Y \geq \epsilon$, then $X_i \geq \epsilon$ for all i, which implies that

$$\mathbf{P}(Y \geq \epsilon) \leq \mathbf{P}(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) = (1 - \epsilon)^n.$$

Since this is true for all n, we must have

$$\mathbf{P}(Y \geq \epsilon) \leq \lim_{n\to\infty} (1 - \epsilon)^n = 0.$$

This shows that $\mathbf{P}(Y \geq \epsilon) = 0$, for any positive ϵ. We conclude that $\mathbf{P}(Y > 0) = 0$, which implies that $\mathbf{P}(Y = 0) = 1$. Since Y is the limit of Y_n, we see that Y_n converges to zero with probability 1.