ECE 302: Probability and Applications[1]

Week 11 Topics

- Markov & Chebyshev Inequalities
- Sample Mean of IID RVs
- Law of Large Numbers
- Central Limit Theorem

1 The Markov and Chebyshev Inequalities

In general, the mean and variance of a random variable do not provide enough information to determine the cdf/pdf. However, in this section we will show that the mean and variance of a random variable X do allow us to obtain bounds for probabilities of the form $P[|X| \geq t]$.

First consider a random variable X that is a nonnegative and that has mean $E[X]$. The **Markov inequality** then states that

$$P[X \geq a] \leq \frac{E[X]}{a} \quad \text{for } X \text{ nonnegative.}$$

We obtain Eq. (4.75) as follows:

$$\begin{aligned} E[X] &= \int_0^a t f_X(t)dt + \int_a^\infty t f_X(t)dt \geq \int_a^\infty t f_X(t)dt \\ &\geq \int_a^\infty a f_X(t)dt = aP[X \geq a] \end{aligned}$$

The first inequality results from discarding the integral from zero to a; the second inequality results from replacing t with the smaller number a.

Example 1

The mean height of children in a kindergarten class is 3 feet, 6 inches. Find the bound on the probability that a kid in the class is taller than 9 feet. The Markov inequality gives $P[H \geq 9] \leq 42/108 = .389$.

The bound in the above example appears to be ridiculous. However, a bound, by its very nature, must take the worst case into consideration. One can easily construct a random variable for which the bound given by the Markov inequality is exact. The reason we know that the bound in the above example is ridiculous is that we have knowledge about the variability of the children's height about their mean.

Now suppose that the mean $E[X] = m$ and the variance $\text{VAR}[X] = \sigma^2$ of a random variable are known, and that we are interested in bounding $P[|X - m| \geq a]$. The **Chebyshev inequality** states that

$$P[|X - m| \geq a] \leq \frac{\sigma^2}{a^2}$$

The Chebyshev inequality is a consequence of the Markov inequality. Let $D^2 = (X - m)^2$ be the squared deviation from the mean. Then the Markov inequality applied to D^2 gives

$$P\left[D^2 \geq a^2\right] \leq \frac{E\left[(X-m)^2\right]}{a^2} = \frac{\sigma^2}{a^2}$$

Equation (4.76) follows when we note that $\{D^2 \geq a^2\}$ and $\{|X - m| \geq a\}$ are equivalent events.

Suppose that a random variable X has zero variance; then the Chebyshev inequality implies that

$$P[X = m] = 1,$$

that is, the random variable is equal to its mean with probability one. In other words, X is equal to the constant m in almost all experiments.

Example 2

The mean response time and the standard deviation in a multi-user computer system are known to be 15 seconds and 3 seconds, respectively. Estimate the probability that the response time is more than 5 seconds from the mean.

The Chebyshev inequality with $m = 15$ seconds, $\sigma = 3$ seconds, and $a = 5$ seconds gives

$$P[|X - 15| \geq 5] \leq \frac{9}{25} = .36.$$

Example 3

If X has mean m and variance σ^2, then the Chebyshev inequality for $a = k\sigma$ gives

$$P[|X - m| \geq k\sigma] \leq \frac{1}{k^2}$$

Now suppose that we know that X is a Gaussian random variable, then for $k = 2, P[|X - m| \geq 2\sigma] = .0456$, whereas the Chebyshev inequality gives the upper bound .25.

Example 4: Chebyshev Bound Is Tight

Let the random variable X have $P[X = -v] = P[X = v] = 0.5$. The mean is zero and the variance is $\text{VAR}[X] = E[X^2] = (-v)^2 0.5 + v^2 0.5 = v^2$.

Note that $P[|X| \geq v] = 1$. The Chebyshev inequality states:

$$P[|X| \geq v] \leq 1 - \frac{\text{VAR}[X]}{v^2} = 1$$

We see that the bound and the exact value are in agreement, so the bound is tight.

We see from the example above that for certain random variables, the Chebyshev inequality can give rather loose bounds. Nevertheless, the inequality is useful in situations in which we have no knowledge about the distribution of a given random variable other than its mean and variance. In an upcoming section, we will use the Chebyshev inequality to prove that the arithmetic average of independent measurements of the same random variable is highly likely to be close to the expected value of the random variable when the number of measurements is large.

If more information is available than just the mean and variance, then it is possible to obtain bounds that are tighter than the Markov and Chebyshev inequalities. Consider the Markov inequality again. The region of interest is $A = \{t \geq a\}$, so let $I_A(t)$ be the indicator function, that is, $I_A(t) = 1$ if $t \in A$ and $I_A(t) = 0$ otherwise. The key step in the derivation is to note that $t/a \geq 1$ in the region of interest. In effect we bounded $I_A(t)$ by t/a as shown in Fig. 1.

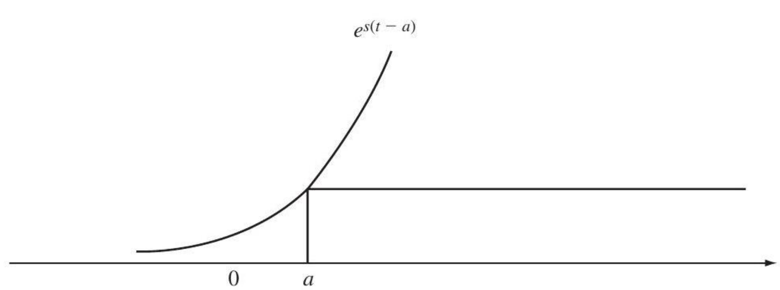

Figure 1: Bounds on indicator function for $A = \{t \geq a\}$.

We then have:

$$P[X \geq a] = \int_0^\infty I_A(t) f_X(t) dt \leq \int_0^\infty \frac{t}{a} f_X(t) dt = \frac{E[X]}{a}$$

By changing the upper bound on $I_A(t)$, we can obtain different bounds on $P[X \geq a]$. Consider the bound $I_A(t) \leq e^{s(t-a)}$, also shown in Fig. 1, where $s > 0$. The resulting bound is:

$$\begin{aligned} P[X \geq a] &= \int_0^\infty I_A(t) f_X(t) dt \leq \int_0^\infty e^{s(t-a)} f_X(t) dt \\ &= e^{-sa} \int_0^\infty e^{st} f_X(t) dt = e^{-sa} E\left[e^{sX}\right] \end{aligned}$$

This bound is called the **Chernoff bound**, which can be seen to depend on the expected value of an exponential function of X. This function is called the moment generating function and is related to the transforms that are introduced in the next section. We develop the Chernoff bound further in the next section.

Sums of Random Variables and Long-Term Averages

Many problems involve the counting of the number of occurrences of events, the measurement of cumulative effects, or the computation of arithmetic averages in a series of measurements. Usually these problems can be reduced to the problem of finding, exactly or approximately, the distribution of a random variable that consists of the sum of n independent, identically distributed random variables. In this chapter, we investigate sums of random variables and their properties as n becomes large.

In Section 7.1, we show how the characteristic function is used to compute the pdf of the sum of independent random variables. In Section 7.2, we discuss the sample mean estimator for the expected value of a random variable and the relative frequency estimator for the probability of an event. We introduce measures for assessing the goodness of these estimators. We then discuss the laws of large numbers, which are theorems that state that the sample mean and relative frequency estimators converge to the corresponding expected values and probabilities as the number of samples is increased. These theoretical results demonstrate the remarkable consistency between probability theory and observed behavior, and they reinforce the relative frequency interpretation of probability.

In Section 7.3, we present the central limit theorem, which states that, under very general conditions, the cdf of a sum of random variables approaches that of a Gaussian random variable even though the cdf of the individual random variables may be far from Gaussian. This result enables us to approximate the pdf of sums of random variables by the pdf of a Gaussian random variable. The result also explains why the Gaussian random variable appears in so many diverse applications.

In Section 7.4 we consider sequences of random variables and their convergence properties. In Section 7.5 we discuss random experiments in which events occur at random times. In these experiments we are interested in the average rate at which events occur as well as the rate at which quantities associated with the events grow. Finally, Section 7.6 introduces computer methods based on the discrete Fourier transform that prove very useful in the numerical calculation of pmf's and pdf's from their transforms.

2 Sums of Random Variables

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables, and let S_n be their sum:

$$S_n = X_1 + X_2 + \cdots + X_n$$

In this section, we find the mean and variance of S_n, as well as the pdf of S_n in the important special case where the X_j's are independent random variables.

2.1 Mean and Variance of Sums of Random Variables

In an earlier section we found that *regardless of statistical dependence, the expected value of a sum of* n *random variables is equal to the sum of the expected values*:

$$E\left[X_1 + X_2 + \cdots + X_n\right] = E\left[X_1\right] + \cdots + E\left[X_n\right] \tag{1}$$

Thus knowledge of the means of the X_j's suffices to find the mean of S_n. The following example shows that in order to compute the variance of a sum of random variables, we need to know the variances and covariances of the X_j's.

Example 5

Find the variance of $Z = X + Y$.
From Eq. (1), $E[Z] = E[X + Y] = E[X] + E[Y]$. The variance of Z is therefore

$$\begin{aligned}\mathrm{VAR}(Z) =& E\left[(Z - E[Z])^2\right] = E\left[(X + Y - E[X] - E[Y])^2\right]\\ =& E\left[\{(X - E[X]) + (Y - E[Y])\}^2\right]\\ =& E\left[(X - E[X])^2 + (Y - E[Y])^2 + (X - E[X])(Y - E[Y])\right.\\ &\left. + (Y - E[Y])(X - E[X])\right]\\ =& \mathrm{VAR}[X] + \mathrm{VAR}[Y] + \mathrm{COV}(X, Y) + \mathrm{COV}(Y, X)\\ =& \mathrm{VAR}[X] + \mathrm{VAR}[Y] + 2\,\mathrm{COV}(X, Y)\end{aligned}$$

In general, the covariance $\mathrm{COV}(X, Y)$ is not equal to zero, so the variance of a sum is not necessarily equal to the sum of the individual variances.

The result in Example 5 can be generalized to the case of n random variables:

$$\begin{aligned}\mathrm{VAR}\left(X_1 + X_2 + \cdots + X_n\right) &= E\left\{\sum_{j=1}^{n}\left(X_j - E\left[X_j\right]\right)\sum_{k=1}^{n}\left(X_k - E\left[X_k\right]\right)\right\}\\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} E\left[\left(X_j - E\left[X_j\right]\right)\left(X_k - E\left[X_k\right]\right)\right]\\ &= \sum_{k=1}^{n}\mathrm{VAR}\left(X_k\right) + \sum_{\substack{j=1\\ j\neq k}}^{n}\sum_{k=1}^{n}\mathrm{COV}\left(X_j, X_k\right)\end{aligned}$$

Thus *in general, the variance of a sum of random variables is not equal to the sum of the individual variances.*

An important special case is when the X_j's are independent random variables. If $X_1, X_2, \ldots, X_n$ *are independent random variables*, then $\mathrm{COV}(X_j, X_k) = 0$ for $j \neq k$ and

$$\mathrm{VAR}\left(X_1 + X_2 + \cdots + X_n\right) = \mathrm{VAR}\left(X_1\right) + \cdots + \mathrm{VAR}\left(X_n\right) \tag{2}$$

Example 6: Sum of iid Random Variables

Find the mean and variance of the sum of n independent, identically distributed (iid) random variables, each with mean μ and variance σ^2.

The mean of S_n is obtained from Eq. (1):

$$E[S_n] = E[X_1] + \cdots + E[X_n] = n\mu$$

The covariance of pairs of independent random variables is zero, so by Eq. (2),

$$\text{VAR}[S_n] = n\,\text{VAR}[X_j] = n\sigma^2$$

since $\text{VAR}[X_j] = \sigma^2$ for $j = 1, \ldots, n$.

2.2 pdf of Sums of Independent Random Variables

Let $X_1, X_2, \ldots, X_n$ be n *independent* random variables. In this section we use transform methods to find the pdf of $S_n = X_1 + X_2 + \cdots + X_n$.

First, consider the $n = 2$ case, $Z = X + Y$, where X and Y are independent random variables. The characteristic function of Z is given by

$$\begin{aligned} \Phi_Z(\omega) &= E\left[e^{j\omega Z}\right] \\ &= E\left[e^{j\omega(X+Y)}\right] \\ &= E\left[e^{j\omega X} e^{j\omega Y}\right] \\ &= E\left[e^{j\omega X}\right] E\left[e^{j\omega Y}\right] \\ &= \Phi_X(\omega)\Phi_Y(\omega) \end{aligned} \tag{3}$$

where the fourth equality follows from the fact that functions of independent random variables (i.e., $e^{j\omega X}$ and $e^{j\omega Y}$) are also independent random variables. Thus the characteristic function of Z is the product of the individual characteristic functions of X and Y.

When we considered $Z = X + Y$ we found that its pdf is given by the convolution of the pdf's of X and Y:

$$f_Z(z) = f_X(x) * f_Y(y) \tag{4}$$

Since $\Phi_Z(\omega)$ can be viewed as the Fourier transform of the pdf of Z:

$$\Phi_Z(\omega) = \mathscr{F}\{f_Z(z)\}$$

By equating the transform of Eq. (4) to Eq. (3) we obtain

$$\Phi_Z(\omega) = \mathscr{F}\left\{f_Z(z)\right\} = \mathscr{F}\left\{f_X(x) * f_Y(y)\right\} = \Phi_X(\omega)\Phi_Y(\omega) \tag{5}$$

Equation (5) states the well-known result that the Fourier transform of a convolution of two functions is equal to the product of the individual Fourier transforms.

Now consider the sum of n independent random variables:

$$S_n = X_1 + X_2 + \cdots + X_n$$

The characteristic function of S_n is

$$\begin{aligned}\Phi_{S_n}(\omega) &= E\left[e^{j\omega S_n}\right] = E\left[e^{j\omega(X_1+X_2+\cdots+X_n)}\right]\\ &= E\left[e^{j\omega X_1}\right]\ldots E\left[e^{j\omega X_n}\right]\\ &= \Phi_{X_1}(\omega)\ldots\Phi_{X_n}(\omega)\end{aligned} \tag{6}$$

Thus the pdf of S_n can then be found by finding the inverse Fourier transform of the product of the individual characteristic functions of the X_j's.

$$f_{S_n}(X) = \mathscr{F}^{-1}\left\{\Phi_{X_1}(\omega)\ldots\Phi_{X_n}(\omega)\right\}$$

Example 7: Sum of Independent Gaussian Random Variables

Let S_n be the sum of n independent Gaussian random variables with respective means and variances, $m_1, \ldots, m_n$ and $\sigma_1^2, \ldots, \sigma_n^2$. Find the pdf of S_n.

The characteristic function of X_k is

$$\Phi_{X_k}(\omega) = e^{+j\omega m_k - \omega^2\sigma_k^2 n}$$

so by Eq. (6),

$$\begin{aligned}\Phi_{S_n}(\omega) &= \prod_{k=1}^{n} e^{+j\omega m_k - \omega^2\sigma_k^2/2}\\ &= \exp\left\{+j\omega\left(m_1+\cdots+m_n\right) - \omega^2\left(\sigma_1^2+\cdots+\sigma_n^2\right)/2\right\}\end{aligned}$$

This is the characteristic function of a Gaussian random variable. Thus S_n is a Gaussian random variable with mean $m_1 + \cdots + m_n$ and variance $\sigma_1^2 + \cdots + \sigma_n^2$.

Example 8: Sum of iid Random Variables

Find the pdf of a sum of n independent, identically distributed random variables with characteristic functions

$$\Phi_{X_k}(\omega) = \Phi_X(\omega) \quad \text{for } k = 1, \ldots, n$$

Equation (6) immediately implies that the characteristic function of S_n is

$$\Phi_{S_n}(\omega) = \{\Phi_X(\omega)\}^n \tag{7}$$

The pdf of S_n is found by taking the inverse transform of this expression.

Example 9: Sum of iid Exponential Random Variables

Find the pdf of a sum of n independent exponentially distributed random variables, all with parameter α.

The characteristic function of a single exponential random variable is

$$\Phi_X(\omega) = \frac{\alpha}{\alpha - j\omega}$$

From the previous example we then have that

$$\Phi_{S_n}(\omega) = \left\{\frac{\alpha}{\alpha - j\omega}\right\}^n$$

From **Table 4.1**, we see that S_n is an m-Erlang random variable.

When dealing with integer-valued random variables it is usually preferable to work with the probability generating function

$$G_N(z) = E\left[z^N\right]$$

The generating function for a sum of independent discrete random variables, $N = X_1 + \cdots + X_n$, is

$$\begin{aligned} G_N(z) &= E\left[z^{X_1+\cdots+X_n}\right] = E\left[z^{X_1}\right] \ldots E\left[z^{X_n}\right] \\ &= G_{X_1}(z) \ldots G_{X_n}(z) \end{aligned}$$

Example 10

Find the generating function for a sum of n independent, identically geometrically distributed random variables.

The generating function for a single geometric random variable is given by

$$G_X(z) = \frac{pz}{1 - qz}$$

Therefore the generating function for a sum of n such independent random variables is

$$G_N(z) = \left\{ \frac{pz}{1 - qz} \right\}^n$$

From **Table 3.1**, we see that this is the generating function of a negative binomial random variable with parameters p and n.

3 The Sample Mean and the Laws of Large Numbers

Let X be a random variable for which the mean, $E[X] = \mu$, is unknown. Let $X_1, \ldots, X_n$ denote n independent, repeated measurements of X; that is, the X_j's are **independent, identically distributed** (iid) random variables with the same pdf as X. The sample mean of the sequence is used to estimate $E[X]$:

$$M_n = \frac{1}{n} \sum_{j=1}^{n} X_j$$

In this section, we compute the expected value and variance of M_n in order to assess the effectiveness of M_n as an estimator for $E[X]$. We also investigate the behavior of M_n as n becomes large.

The following example shows that the relative frequency estimator for the probability of an event is a special case of a sample mean. Thus the results derived below for the sample mean are also applicable to the relative frequency estimator.

Example 11: Relative Frequency

Consider a sequence of independent repetitions of some random experiment, and let the random variable I_j be the indicator function for the occurrence of event A in the j th trial. The total number of occurrences of A in the first n trials is then

$$N_n = I_1 + I_2 + \cdots + I_n$$

The **relative frequency** of event A in the first n repetitions of the experiment is then

$$f_A(n) = \frac{1}{n} \sum_{j=1}^{n} I_j$$

Thus the relative frequency $f_A(n)$ is simply the sample mean of the random variables I_j.

The sample mean is itself a random variable, so it will exhibit random variation. A good estimator should have the following two properties: (1) On the average, it should give the correct value of the parameter being estimated, that is, $E[M_n] = \mu$; and (2) It should not vary too much about the correct value of this parameter, that is, $E\left[(M_n - \mu)^2\right]$ is small.

The expected value of the sample mean is given by

$$E[M_n] = E\left[\frac{1}{n} \sum_{j=1}^{n} X_j\right] = \frac{1}{n} \sum_{j=1}^{n} E[X_j] = \mu \tag{8}$$

since $E[X_j] = E[X] = \mu$ for all j. Thus the sample mean is equal to $E[X] = \mu$, on the average. For this reason, we say that the sample mean is an **unbiased estimator** for μ.

Equation (8) implies that the mean square error of the sample mean about μ is equal to the variance of M_n, that is,

$$E\left[(M_n - \mu)^2\right] = E\left[(M_n - E[M_n])^2\right]$$

Note that $M_n = S_n/n$, where $S_n = X_1 + X_2 + \cdots + X_n$. From Eq. (2), $\mathrm{VAR}[S_n] = n\ \mathrm{VAR}[\mathrm{X_j}] = n\sigma^2$, since the X_j's are iid random variables. Thus

$$\mathrm{VAR}[M_n] = \frac{1}{n^2} \mathrm{VAR}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \tag{9}$$

Equation (9) states that the variance of the sample mean approaches zero as the number of samples is increased. This implies that the probability that the sample mean is close to the true mean approaches one as n becomes very large. We can formalize this statement by using the Chebyshev inequality:

$$P\left[|M_n - E\left[M_n\right]| \geq \varepsilon\right] \leq \frac{\text{VAR}\left[M_n\right]}{\varepsilon^2}$$

Substituting for $E\left[M_n\right]$ and $\text{VAR}\left[M_n\right]$, we obtain

$$P\left[|M_n - \mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2} \tag{10}$$

If we consider the complement of the event considered in Eq. (10), we obtain

$$P\left[|M_n - \mu| < \varepsilon\right] \geq 1 - \frac{\sigma^2}{n\varepsilon^2} \tag{11}$$

Thus for any choice of error ε and probability $1-\delta$, we can select the number of samples n so that M_n is within ε of the true mean with probability $1-\delta$ or greater. The following example illustrates this.

Example 12

A voltage of constant, but unknown, value is to be measured. Each measurement X_j is actually the sum of the desired voltage v and a noise voltage N_j of zero mean and standard deviation of 1 microvolt (μV):

$$X_j = v + N_j$$

Assume that the noise voltages are independent random variables. How many measurements are required so that the probability that M_n is within $\varepsilon = 1\mu V$ of the true mean is at least. 99 ?

Each measurement X_j has mean v and variance 1, so from Eq. (11) we require that n satisfy

$$1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - \frac{1}{n} = .99$$

This implies that $n = 100$.

Thus if we were to repeat the measurement 100 times and compute the sample mean, on the average, at least 99 times out of 100, the resulting sample mean will be within $1\mu V$ of the true mean.

Note that if we let n approach infinity in **Eq. (5.20)** we obtain

$$\lim_{n\to\infty} P\left[|M_n - \mu| < \varepsilon\right] = 1$$

Equation (11) requires that the X_j's have finite variance. It can be shown that this limit holds even if the variance of the X_j's does not exist [Gnedenko, p. 203]. We state this more general result:

> **Weak Law of Large Numbers:** Let $X_1, X_2, \ldots$ be a sequence of iid random variables with finite mean $E[X] = \mu$, then for $\varepsilon > 0$,
>
> $$\lim_{n\to\infty} P\left[|M_n - \mu| < \varepsilon\right] = 1 \tag{12}$$

The weak law of large numbers states that for a large enough fixed value of n, the sample mean using n samples will be close to the true mean with high probability. The weak law of large numbers does not address the question about what happens to the sample mean as a function of n as we make additional measurements. This question is taken up by the strong law of large numbers, which we discuss next.

Suppose we make a series of independent measurements of the same random variable. Let $X_1, X_2, \ldots$ be the resulting sequence of iid random variables with mean μ. Now consider the sequence of sample means that results from the above measurements: $M_1, M_2, \ldots$, where M_j is the sample mean computed using X_1 through X_j. The notion of statistical regularity discussed in Chapter 1 leads us to expect that this sequence of sample means converges to μ, that is, we expect that with high probability, *each particular sequence of sample means approaches μ and stays there*, as shown in Fig. 7.1. In terms of probabilities, we expect the following:

$$P\left[\lim_{n\to\infty} M_n = \mu\right] = 1$$

that is, with virtual certainty, every sequence of sample mean calculations converges to the true mean of the quantity. The proof of this result is well beyond the level of this course (see [Gnedenko, p. 216]), but we will have the opportunity in later sections to apply the result in various situations.

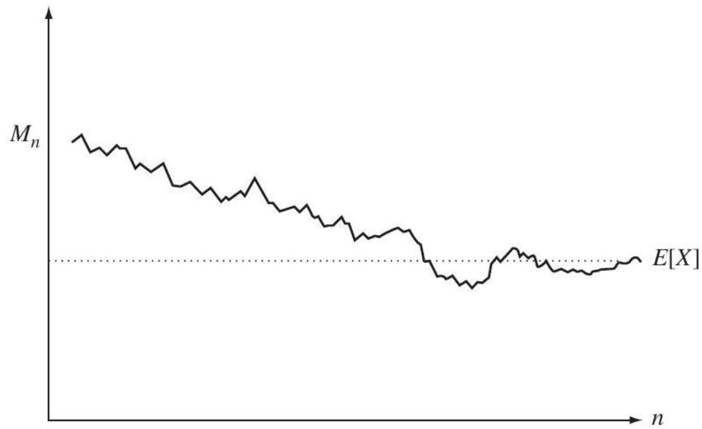

Figure 2: Convergence of sequence of sample means to $E[X]$.

Strong Law of Large Numbers: Let $X_1, X_2, \ldots$ be a sequence of iid random variables with finite mean $E[X] = \mu$ and finite variance, then

$$P\left[\lim_{n\to\infty} M_n = \mu\right] = 1 \tag{13}$$

Equation (13) appears similar to Eq. (12), but in fact it makes a dramatically different statement. It states that with probability 1, *every sequence of sample mean calculations will eventually approach and stay close to* $E[X] = \mu$. This is the type of convergence we expect in physical situations where statistical regularity holds.

With the strong law of large numbers we come full circle in the modeling process. We began in Chapter 1 by noting that statistical regularity is observed in many physical phenomena, and from this we deduced a number of properties of relative frequency. These properties were used to formulate a set of axioms from which we developed a mathematical theory of probability. We have now come full circle and shown that, under certain conditions, the *theory* predicts the convergence of sample means to expected values. There are still gaps between the mathematical theory and the real world (i.e., we can never actually carry out an infinite number of measurements and compute an infinite number of sample means). Nevertheless, the strong law of large numbers demonstrates the remarkable consistency between the theory and the observed physical behavior.

We already indicated that relative frequencies are special cases of sample averages. If we apply the weak law of large numbers to the relative frequency of an event A, $f_A(n)$, in a sequence of independent repetitions of a random experiment, we obtain

$$\lim_{n\to\infty} P\left[|f_A(n) - P[A]| < \varepsilon\right] = 1$$

If we apply the strong law of large numbers, we obtain

$$P\left[\lim_{n\to\infty} f_A(n) = P[A]\right] = 1$$

Example 13

In order to estimate the probability of an event A, a sequence of Bernoulli trials is carried out and the relative frequency of A is observed. How large should n be in order to have a .95 probability that the relative frequency is within 0.01 of $p = P[A]$?

Let $X = I_A$ be the indicator function of A. From Table 3.1 we have that the mean of I_A is $\mu = p$ and the variance is $\sigma^2 = p(1-p)$. Since p is unknown, σ^2 is also unknown. However, it is easy to show that $p(1-p)$ is at most 1/4 for $0 \leq p \leq 1$. Therefore, by Eq. (10),

$$P\left[|f_A(n) - p| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

The desired accuracy is $\varepsilon = 0.01$ and the desired probability is

$$1 - .95 = \frac{1}{4n\varepsilon^2}$$

We then solve for n and obtain $n = 50,000$. It has already been pointed out that the Chebyshev inequality gives very loose bounds, so we expect that this value for n is probably overly conservative. In the next section, we present a better estimate for the required value of n.

4 The Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence of iid random variables with finite mean μ and finite variance σ^2, and let S_n be the sum of the first n random variables in the sequence:

$$S_n = X_1 + X_2 + \cdots + X_n \tag{14}$$

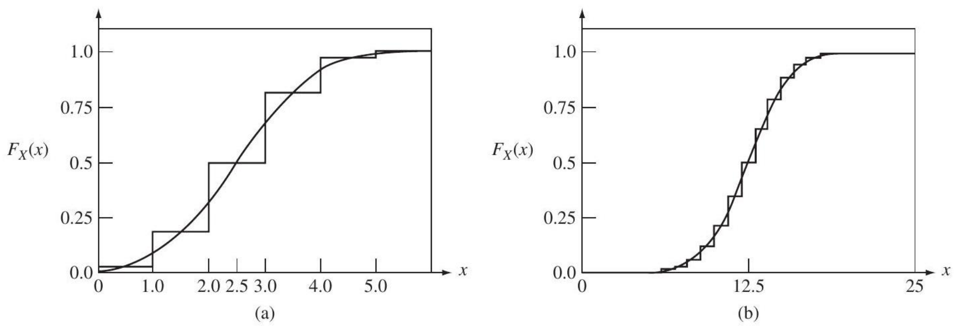

Figure 3: (a) The cdf of the sum of five independent Bernoulli random variables with $p = 1/2$ and the cdf of a Gaussian random variable of the same mean and variance. (b) The cdf of the sum of 25 independent Bernoulli random variables with $p = 1/2$ and the cdf of a Gaussian random variable of the same mean and variance.

In Section 7.1, we developed methods for determining the exact pdf of S_n. We now present the central limit theorem, which states that, as n becomes large, the cdf of a properly normalized S_n approaches that of a Gaussian random variable. This enables us to approximate the cdf of S_n with that of a Gaussian random variable.

The central limit theorem explains why the Gaussian random variable appears in so many diverse applications. In nature, many macroscopic phenomena result from the addition of numerous independent, microscopic processes; this gives rise to the Gaussian random variable. In many man-made problems, we are interested in averages that often consist of the sum of independent random variables. This again gives rise to the Gaussian random variable.

From Example 6, we know that if the X_j's are iid, then S_n has mean $n\mu$ and variance $n\sigma^2$. The central limit theorem states that the cdf of a suitably normalized version of S_n approaches that of a Gaussian random variable.

Central Limit Theorem: Let S_n be the sum of n iid random variables with finite mean $E[X] = \mu$ and finite variance σ^2, and let Z_n be the zero-mean, unitvariance random variable defined by

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

then

$$\lim_{n\to\infty} P\left[Z_n \leq z\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

Note that Z_n is sometimes written in terms of the sample mean:

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

The amazing part about the central limit theorem is that the summands X_j can have any distribution as long as they have a finite mean and finite variance. This gives the result its wide applicability.

Figures 3 through 5 compare the exact cdf and the Gaussian approximation for the sums of Bernoulli, uniform, and exponential random variables, respectively. In all three cases, it can be seen that the approximation improves as number of terms in the sum increases. The proof of the central limit theorem is discussed in the last part of this section.

Example 14

Suppose that orders at a restaurant are iid random variables with mean $\mu = \$8$ and standard deviation $\sigma = \$2$. Estimate the probability that the first 100 customers spend a total of more than \$840. Estimate the probability that the first 100 customers spend a total of between \$780 and \$820.

Let X_k denote the expenditure of the k th customer, then the total spent by the first 100 customers is

$$S_{100} = X_1 + X_2 + \cdots + X_{100}$$

The mean of S_{100} is $n\mu = 800$ and the variance is $n\sigma^2 = 400$. Figure 6 shows the pdf of S_{100} where it can be seen that the pdf is highly concen-

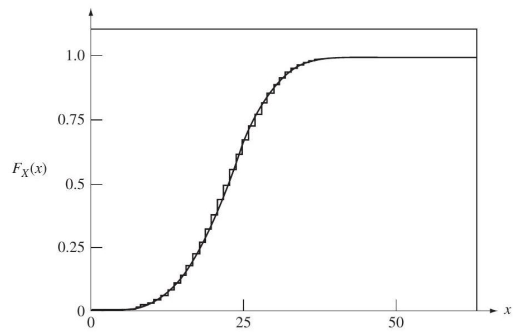

Figure 4: The cdf of the sum of five independent discrete, uniform random variables from the set $\{0, 1, \ldots, 9\}$ and the cdf of a Gaussian random variable of the same mean and variance.

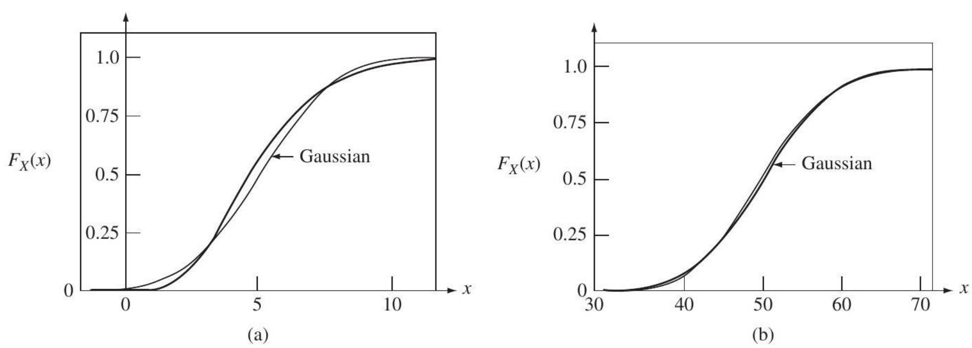

Figure 5: (a) The *cdf* of the sum of five independent exponential random variables of mean 1 and the *cdf* of a Gaussian random variable of the same mean and variance. (b) The cdf of the sum of 50 independent exponential random variables of mean 1 and the cdf of a Gaussian random variable of the same mean and variance.

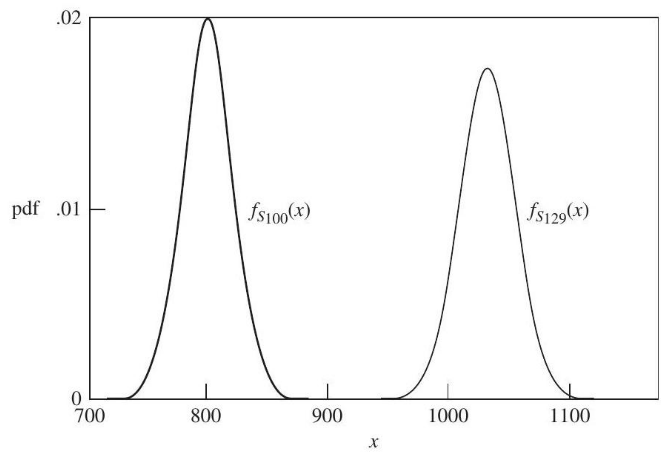

Figure 6: Gaussian pdf approximations S_{100} and S_{129} in Examples 14 and 15.

trated about the mean. The normalized form of S_{100} is

$$Z_{100} = \frac{S_{100} - 800}{20}$$

Thus

$$\begin{aligned} P\left[S_{100} > 840\right] &= P\left[Z_{100} > \frac{840 - 800}{20}\right] \\ &\simeq Q(2) = 2.28\left(10^{-2}\right) \end{aligned}$$

where we used Table 4.2 to evaluate $Q(2)$. Similarly,

$$\begin{aligned} P\left[780 \leq S_{100} \leq 820\right] &= P\left[-1 \leq Z_{100} \leq 1\right] \\ &\simeq 1 - 2Q(1) \\ &= .682 \end{aligned}$$

Example 15

In Example 14, after how many orders can we be 90% sure that the total spent by all customers is more than \$1000 ?

The problem here is to find the value of n for which

$$P\left[S_n > 1000\right] = .90$$

S_n has mean $8n$ and variance $4n$. Proceeding as in the previous example, we have

$$P[S_n > 1000] = P\left[Z_n > \frac{1000 - 8n}{2\sqrt{n}}\right] = .90$$

Using the fact that $Q(-x) = 1 - Q(x)$, Table 4.3 implies that n must satisfy

$$\frac{1000 - 8n}{2\sqrt{n}} = -1.2815$$

which yields the following quadratic equation for $\sqrt{n}$:

$$8n - 1.2815(2)\sqrt{n} - 1000 = 0$$

The positive root of the equation yields $\sqrt{n} = 11.34$, or $n = 128.6$. Figure 6 shows the pdf for S_{129}.

Example 16

The time between events in a certain random experiment is iid exponential random variables with mean m seconds. Find the probability that the 1000th event occurs in the time interval $(1000 \pm 50)m$.

Let X_j be the time between events and let S_n be the time of the nth event, then S_n is given by Eq. (14). From Table 4.1, the mean and variance of X_j is given by $E[X_j] = m$ and $\text{VAR}[X_j] = m^2$. The mean and variance of S_n are then $E[S_n] = nE[X_j] = nm$ and $\text{VAR}[S_n] = n\,\text{VAR}[X_j] = nm^2$. The central limit theorem then gives

$$\begin{aligned} P[950m \le S_{1000} \le 1050m] &= P\left[\frac{950m - 1000m}{m\sqrt{1000}} \le Z_n \le \frac{1050m - 1000m}{m\sqrt{1000}}\right] \\ &\simeq Q(1.58) - Q(-1.58) \\ &= 1 - 2Q(1.58) \\ &= 1 - 2(0.0567) = .8866 \end{aligned}$$

Thus as n becomes large, S_n is very likely to be close to its mean nm. We can therefore conjecture that the long-term average rate at which events occur is

$$\frac{n \text{ events}}{S_n \text{ seconds}} = \frac{n}{nm} = \frac{1}{m} \text{ events / second}$$

The calculation of event occurrence rates and related averages is discussed in Section 7.5.

4.1 Gaussian Approximation for Binomial Probabilities

We found in Chapter 2 that the binomial random variable becomes difficult to compute directly for large n because of the need to calculate factorial terms. A particularly important application of the central limit theorem is in the approximation of binomial probabilities. Since the binomial random variable is a sum of iid Bernoulli random variables (which have finite mean and variance), its cdf approaches that of a Gaussian random variable. Let X be a binomial random variable with mean np and variance $np(1-p)$, and let Y be a Gaussian random variable with the same mean and variance, then by the central limit theorem for n large the probability that $X = k$ is approximately equal to the integral of the Gaussian pdf in an interval of unit length about k, as shown in Fig. 7.6:

$$\begin{aligned} P[X=k] &\simeq P\left[k-\frac{1}{2} < Y < k+\frac{1}{2}\right] \\ &= \frac{1}{\sqrt{2\pi np(1-p)}} \int_{k-1/2}^{k+1/2} e^{-(x-np)^2/2np(1-p)} dx \end{aligned}$$

The above approximation can be simplified by approximating the integral by the product of the integrand at the center of the interval of integration (that is, $x = k$) and the length of the interval of integration (one):

$$P[X=k] \simeq \frac{1}{\sqrt{2\pi np(1-p)}} e^{-(k-np)^2/2np(1-p)} \tag{15}$$

Figures 7(a) and 7(b) compare the binomial probabilities and the Gaussian approximation using Eq. (15).

Example 17

In Example 13 in Section 7.2, we used the Chebyshev inequality to estimate the number of samples required for there to be a .95 probability that the relative frequency estimate for the probability of an event A would be within 0.01 of $P[A]$. We now estimate the required number of samples using the Gaussian approximation for the binomial distribution.

Let $f_A(n)$ be the relative frequency of A in n Bernoulli trials. Since

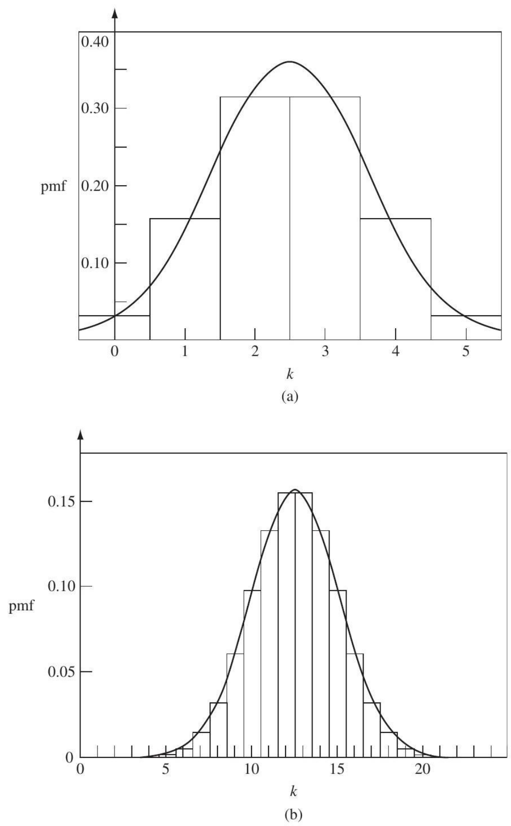

Figure 7: (a) Gaussian approximation for binomial probabilities with $n = 5$ and $p = 1/2$. (b) Gaussian approximation for binomial with $n = 25$ and $p = 1/2$.

$f_A(n)$ has mean p and variance $p(1-p)/n$, then

$$Z_n = \frac{f_A(n) - p}{\sqrt{p(1-p)/n}}$$

has zero mean and unit variance, and is approximately Gaussian for n sufficiently large. The probability of interest is

$$P\left[|f_A(n) - p| < \varepsilon\right] \simeq P\left[|Z_n| < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right] = 1 - 2Q\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

The above probability cannot be computed because p is unknown. However, it can be easily shown that $p(1-p) \leq 1/4$ for p in the unit interval.

It then follows that for such p, $\sqrt{p(1-p)} \leq 1/2$, and since $Q(x)$ decreases with increasing argument

$$P\left[|f_A(n) - p| < \varepsilon\right] > 1 - 2Q(2\varepsilon\sqrt{n})$$

We want the above probability to equal 95. This implies that $Q(2\varepsilon\sqrt{n}) = (1 - .95)/2 = .025$. From Table 4.2, we see that the argument of $Q(x)$ should be approximately 1.95, thus

$$2\varepsilon\sqrt{n} = 1.95$$

Solving for n, we obtain

$$n = (.98)^2/\varepsilon^2 = 9506$$

4.2 Chernoff Bound for Binomial Random Variable

The Gaussian pdf extends over the entire real line. When taking the sum of random variables that have a finite range, such as the binomial random variable, the central limit theorem can be inaccurate at the extreme values of the sum. The Chernoff bound introduced in Chapter 3 gives better estimates.

The Chernoff bound for the binomial is given by:

$$P[X \geq a] \leq e^{-sa} E\left[e^{sX}\right] = e^{-sa} E\left[(e^s)^X\right] = e^{-sa} G_N(e^s) = e^{-sa}\left(q + pe^s\right)^n$$

where $s > 0$, and $G_N(z)$ is the pgf for the binomial random variable. To minimize the bound we take the derivative with respect to s and set it to zero:

$$0 = \frac{d}{ds} e^{-sa} G_N(e^s) = -ae^{-sa}\left(q + pe^s\right)^n + e^{-sa} e^s np \left(q + pe^s\right)^{n-1}$$
$$a\left(q + pe^s\right) = e^s np$$

where the second line results after canceling common terms. The optimum s and the associated bound are:

$$e^s = \frac{aq}{p(n-a)}$$
$$P[X \geq a] \leq \left(\frac{p(n-a)}{aq}\right)^a \left(q + p\frac{aq}{p(n-a)}\right)^n = \left(\frac{p(n-a)}{aq}\right)^a \left(\frac{qn}{(n-a)}\right)^n$$

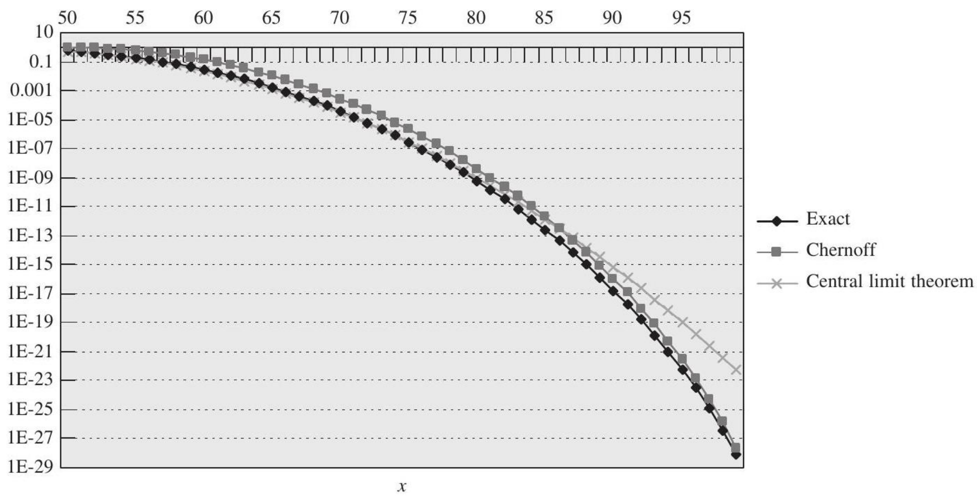

Figure 8: Comparison of Chernoff bound and central limit theorem.

$$= \left(\frac{p(1-a/n)}{(a/n)q}\right)^a \left(\frac{q}{1-a/n}\right)^n = \left(\frac{p^{a/n}q^{1-a/n}}{(a/n)^{\frac{a}{n}}(1-a/n)^{1-a/n}}\right)^n$$

Example 18

Compare the central limit estimate for $P[X > x]$ with the Chernoff bound for the binomial random variable with $n = 100$ and $p = 0.5$.

The central limit gives the estimate:

$$P[X \geq a] \approx Q\left(\frac{x - np}{\sqrt{npq}}\right) = Q\left(\frac{x-50}{5}\right)$$

The Chernoff bound is:

$$P[X \geq a] \leq \left(\frac{1/2}{(x/100)^{\frac{x}{100}}(1-x/100)^{1-x/100}}\right)^{100}$$

Figure 8 shows a comparison of the exact values of the tail distribution with the Chernoff bound and the estimate from the central limit theorem. The central limit theorem estimate is more accurate than the Chernoff bounds up to about $x = 86$. At the extreme values of x, the Chernoff bound remains accurate while the central limit estimate loses its accuracy.

4.3 *Proof of the Central Limit Theorem

We now sketch a proof of the central limit theorem. First note that

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)\,.$$

The characteristic function of Z_n is given by

$$\begin{aligned}
\Phi_{Z_n}(\omega) &= E\left[e^{j\omega Z_n}\right] \\
&= E\left[\exp\left\{\frac{j\omega}{\sigma\sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)\right\}\right] \\
&= E\left[\prod_{k=1}^{n} e^{j\omega(X_k-\mu)/\sigma\sqrt{n}}\right] \\
&= \prod_{k=1}^{n} E\left[e^{j\omega(X_k-\mu)/\sigma\sqrt{n}}\right] \\
&= \left\{E\left[e^{j\omega(X-\mu)/\sigma\sqrt{n}}\right]\right\}^n \qquad (16)
\end{aligned}$$

The third equality follows from the independence of the X_k's and the last equality follows from the fact that the X_k's are identically distributed.

By expanding the exponential in the expression, we obtain an expression in terms of n and the central moments of X:

$$\begin{aligned}
&E\left[e^{j\omega(X-\mu)/\sigma\sqrt{n}}\right] \\
&\quad = E\left[1 + \frac{j\omega}{\sigma\sqrt{n}}(X-\mu) + \frac{(j\omega)^2}{2!n\sigma^2}(X-\mu)^2 + R(\omega)\right] \\
&\quad = 1 + \frac{j\omega}{\sigma\sqrt{n}} E[(X-\mu)] + \frac{(j\omega)^2}{2!n\sigma^2} E\left[(X-\mu)^2\right] + E[R(\omega)]
\end{aligned}$$

Noting that $E[(X-\mu)] = 0$ and $E\left[(X-\mu)^2\right] = \sigma^2$, we have

$$E\left[e^{j\omega(X-\mu)/\sigma\sqrt{n}}\right] = 1 - \frac{\omega^2}{2n} + E[R(\omega)] \qquad (17)$$

The term $E[R(\omega)]$ can be neglected relative to $\omega^2/2n$ as n becomes large. If we substitute Eq. (17) into Eq. (16), we obtain

$$\Phi_{Z_n}(\omega) = \left\{1 - \frac{\omega^2}{2n}\right\}^n$$

$$\to e^{-\omega^2/2} \text{ as } n \to \infty$$

The latter expression is the characteristic function of a zero-mean, unit-variance Gaussian random variable. Thus the cdf of Z_n approaches the cdf of a zero-mean, unit-variance Gaussian random variable.