6

Sums of Random Variables

Random variables of the form

$$W_n = X_1 + \cdots + X_n \tag{6.1}$$

appear repeatedly in probability theory and applications. We could in principle derive the probability model of W_n from the PMF or PDF of $X_1, \ldots, X_n$. However, in many practical applications, the nature of the analysis or the properties of the random variables allow us to apply techniques that are simpler than analyzing a general n-dimensional probability model. In Section 6.1 we consider applications in which our interest is confined to expected values related to W_n, rather than a complete model of W_n. Subsequent sections emphasize techniques that apply when $X_1, \ldots, X_n$ are mutually independent. A useful way to analyze the sum of independent random variables is to transform the PDF or PMF of each random variable to a *moment generating function*.

The central limit theorem reveals a fascinating property of the sum of independent random variables. It states that the CDF of the sum converges to a Gaussian CDF as the number of terms grows without limit. This theorem allows us to use the properties of Gaussian random variables to obtain accurate estimates of probabilities associated with sums of other random variables. In many cases exact calculation of these probabilities is extremely difficult.

6.1 Expected Values of Sums

The theorems of Section 4.7 can be generalized in a straightforward manner to describe expected values and variances of sums of more than two random variables.

Theorem 6.1 *For any set of random variables $X_1, \ldots, X_n$, the expected value of $W_n = X_1 + \cdots + X_n$ is*

$$E[W_n] = E[X_1] + E[X_2] + \cdots + E[X_n].$$

Proof We prove this theorem by induction on n. In Theorem 4.14, we proved $E[W_2] = E[X_1] +$

$E[X_2]$. Now we assume $E[W_{n-1}] = E[X_1]+\cdots+E[X_{n-1}]$. Notice that $W_n = W_{n-1}+X_n$. Since W_n is a sum of the two random variables W_{n-1} and X_n, we know that $E[W_n] = E[W_{n-1}]+E[X_n] = E[X_1] + \cdots + E[X_{n-1}] + E[X_n]$.

Keep in mind that the expected value of the sum equals the sum of the expected values whether or not $X_1, \ldots, X_n$ are independent. For the variance of W_n, we have the generalization of Theorem 4.15:

Theorem 6.2 *The variance of $W_n = X_1 + \cdots + X_n$ is*

$$\operatorname{Var}[W_n] = \sum_{i=1}^{n} \operatorname{Var}[X_i] + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{Cov}\left[X_i, X_j\right].$$

Proof From the definition of the variance, we can write $\operatorname{Var}[W_n] = E[(W_n - E[W_n])^2]$. For convenience, let μ_i denote $E[X_i]$. Since $W_n = \sum_{i=1}^{n} X_n$ and $E[W_n] = \sum_{i=1}^{n} \mu_i$, we can write

$$\operatorname{Var}[W_n] = E\left[\left(\sum_{i=1}^{n}(X_i - \mu_i)\right)^2\right] = E\left[\sum_{i=1}^{n}(X_i - \mu_i)\sum_{j=1}^{n}(X_j - \mu_j)\right] \tag{6.2}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}\left[X_i, X_j\right]. \tag{6.3}$$

In terms of the random vector $\mathbf{X} = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}'$, we see that $\operatorname{Var}[W_n]$ is the sum of all the elements of the covariance matrix $\mathbf{C_X}$. Recognizing that $\operatorname{Cov}[X_i, X_i] = \operatorname{Var}[X]$ and $\operatorname{Cov}[X_i, X_j] = \operatorname{Cov}[X_j, X_i]$, we place the diagonal terms of $\mathbf{C_X}$ in one sum and the off-diagonal terms (which occur in pairs) in another sum to arrive at the formula in the theorem.

When $X_1, \ldots, X_n$ are uncorrelated, $\operatorname{Cov}[X_i, X_j] = 0$ for $i \neq j$ and the variance of the sum is the sum of the variances:

Theorem 6.3 *When $X_1, \ldots, X_n$ are uncorrelated,*

$$\operatorname{Var}[W_n] = \operatorname{Var}[X_1] + \cdots + \operatorname{Var}[X_n].$$

Example 6.1 $X_0, X_1, X_2, \ldots$ is a sequence of random variables with expected values $E[X_i] = 0$ and covariances, $\operatorname{Cov}[X_i, X_j] = 0.8^{|i-j|}$. Find the expected value and variance of a random variable Y_i defined as the sum of three consecutive values of the random sequence

$$Y_i = X_i + X_{i-1} + X_{i-2}. \tag{6.4}$$

.....

Theorem 6.1 implies that

$$E\left[Y_i\right] = E\left[X_i\right] + E\left[X_{i-1}\right] + E\left[X_{i-2}\right] = 0. \tag{6.5}$$

Applying Theorem 6.2, we obtain for each i,

$$\begin{aligned} \text{Var}[Y_i] &= \text{Var}[X_i] + \text{Var}[X_{i-1}] + \text{Var}[X_{i-2}] \\ &\quad + 2\,\text{Cov}\left[X_i, X_{i-1}\right] + 2\,\text{Cov}\left[X_i, X_{i-2}\right] + 2\,\text{Cov}\left[X_{i-1}, X_{i-2}\right]. \end{aligned} \tag{6.6}$$

We next note that $\text{Var}[X_i] = \text{Cov}[X_i, X_i] = 0.8^{i-i} = 1$ and that

$$\text{Cov}\left[X_i, X_{i-1}\right] = \text{Cov}\left[X_{i-1}, X_{i-2}\right] = 0.8^1, \qquad \text{Cov}\left[X_i, X_{i-2}\right] = 0.8^2. \tag{6.7}$$

Therefore

$$\text{Var}[Y_i] = 3 \times 0.8^0 + 4 \times 0.8^1 + 2 \times 0.8^2 = 7.48. \tag{6.8}$$

The following example shows how a puzzling problem can be formulated as a question about the sum of a set of dependent random variables.

Example 6.2 At a party of $n \geq 2$ people, each person throws a hat in a common box. The box is shaken and each person blindly draws a hat from the box without replacement. We say a match occurs if a person draws his own hat. What are the expected value and variance of V_n, the number of matches?

Let X_i denote an indicator random variable such that

$$X_i = \begin{cases} 1 & \text{person } i \text{ draws his hat,} \\ 0 & \text{otherwise.} \end{cases} \tag{6.9}$$

The number of matches is $V_n = X_1 + \cdots + X_n$. Note that the X_i are generally not independent. For example, with $n = 2$ people, if the first person draws his own hat, then the second person must also draw her own hat. Note that the ith person is equally likely to draw any of the n hats, thus $P_{X_i}(1) = 1/n$ and $E[X_i] = P_{X_i}(1) = 1/n$. Since the expected value of the sum always equals the sum of the expected values,

$$E\left[V_n\right] = E\left[X_1\right] + \cdots + E\left[X_n\right] = n(1/n) = 1. \tag{6.10}$$

To find the variance of V_n, we will use Theorem 6.2. The variance of X_i is

$$\text{Var}[X_i] = E\left[X_i^2\right] - \left(E\left[X_i\right]\right)^2 = \frac{1}{n} - \frac{1}{n^2}. \tag{6.11}$$

To find $\text{Cov}[X_i, X_j]$, we observe that

$$\text{Cov}\left[X_i, X_j\right] = E\left[X_i X_j\right] - E\left[X_i\right] E\left[X_j\right]. \tag{6.12}$$

Note that $X_i X_j = 1$ if and only if $X_i = 1$ and $X_j = 1$, and that $X_i X_j = 0$ otherwise. Thus

$$E\left[X_i X_j\right] = P_{X_i,X_j}(1,1) = P_{X_i|X_j}(1|1)\, P_{X_j}(1). \tag{6.13}$$

Given $X_j = 1$, that is, the jth person drew his own hat, then $X_i = 1$ if and only if the ith person draws his own hat from the $n-1$ other hats. Hence $P_{X_i|X_j}(1|1) = 1/(n-1)$ and

$$E\left[X_i X_j\right] = \frac{1}{n(n-1)}, \qquad \text{Cov}\left[X_i, X_j\right] = \frac{1}{n(n-1)} - \frac{1}{n^2}. \tag{6.14}$$

Finally, we can use Theorem 6.2 to calculate

$$\mathrm{Var}[V_n] = n\,\mathrm{Var}[X_i] + n(n-1)\,\mathrm{Cov}\left[X_i, X_j\right] = 1. \tag{6.15}$$

That is, both the expected value and variance of V_n are 1, no matter how large n is!

Example 6.3 Continuing Example 6.2, suppose each person immediately returns to the box the hat that he or she drew. What is the expected value and variance of V_n, the number of matches?

....

In this case the indicator random variables X_i are iid because each person draws from the same bin containing all n hats. The number of matches $V_n = X_1 + \cdots + X_n$ is the sum of n iid random variables. As before, the expected value of V_n is

$$E[V_n] = nE\left[X_i\right] = 1. \tag{6.16}$$

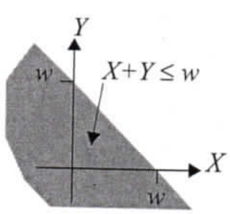
In this case, the variance of V_n equals the sum of the variances,

$$\mathrm{Var}[V_n] = n\,\mathrm{Var}[X_i] = n\left(\frac{1}{n} - \frac{1}{n^2}\right) = 1 - \frac{1}{n}. \tag{6.17}$$

The remainder of this chapter examines tools for analyzing complete probability models of sums of random variables, with the emphasis on sums of independent random variables.

Quiz 6.1 *Let W_n denote the sum of n independent throws of a fair four-sided die. Find the expected value and variance of W_n.*

6.2 PDF of the Sum of Two Random Variables



Before analyzing the probability model of the sum of n random variables, it is instructive to examine the sum $W = X + Y$ of two continuous random variables. As we see in Theorem 6.4, the PDF of W depends on the joint PDF $f_{X,Y}(x, y)$. In particular, in the proof of the theorem, we find the PDF of W using the two-step procedure in which we first find the CDF $F_W(w)$ by integrating the joint PDF $f_{X,Y}(x, y)$ over the region $X + Y \le w$ as shown.

Theorem 6.4 *The PDF of $W = X + Y$ is*

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,Y}(x, w-x)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(w-y, y)\,dy.$$

Proof

$$F_W(w) = P[X + Y \le w] = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{w-x} f_{X,Y}(x, y)\,dy\right)dx. \tag{6.18}$$

Taking the derivative of the CDF to find the PDF, we have

$$f_W(w) = \frac{dF_W(w)}{dw} = \int_{-\infty}^{\infty} \left(\frac{d}{dw} \left(\int_{-\infty}^{w-x} f_{X,Y}(x, y)\, dy \right) \right) dx \tag{6.19}$$

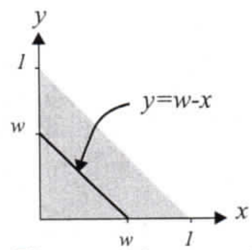
$$= \int_{-\infty}^{\infty} f_{X,Y}(x, w - x)\, dx. \tag{6.20}$$

By making the substitution $y = w - x$, we obtain

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,Y}(w - y, y)\, dy. \tag{6.21}$$

Example 6.4 Find the PDF of $W = X + Y$ when X and Y have the joint PDF

$$f_{X,Y}(x, y) = \begin{cases} 2 & 0 \le y \le 1, 0 \le x \le 1, x + y \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.22}$$



The PDF of $W = X + Y$ can be found using Theorem 6.4. The possible values of X, Y are in the shaded triangular region where $0 \le X + Y = W \le 1$. Thus $f_W(w) = 0$ for $w < 0$ or $w > 1$. For $0 \le w \le 1$, applying Theorem 6.4 yields

$$f_W(w) = \int_0^w 2\, dx = 2w, \qquad 0 \le w \le 1. \tag{6.23}$$

The complete expression for the PDF of W is

$$f_W(w) = \begin{cases} 2w & 0 \le w \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.24}$$

When X and Y are independent, the joint PDF of X and Y can be written as the product of the marginal PDFs $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. In this special case, Theorem 6.4 can be restated.

Theorem 6.5 *When X and Y are independent random variables, the PDF of $W = X + Y$ is*

$$f_W(w) = \int_{-\infty}^{\infty} f_X(w - y)\, f_Y(y)\, dy = \int_{-\infty}^{\infty} f_X(x)\, f_Y(w - x)\, dx.$$

In Theorem 6.5, we combine two univariate functions, $f_X(\cdot)$ and $f_Y(\cdot)$, in order to produce a third function, $f_W(\cdot)$. The combination in Theorem 6.5, referred to as a *convolution*, arises in many branches of applied mathematics.

When X and Y are independent integer-valued discrete random variables, the PMF of $W = X + Y$ is a convolution (Problem 4.10.9).

$$P_W(w) = \sum_{k=-\infty}^{\infty} P_X(k) P_Y(w-k). \tag{6.25}$$

You may have encountered convolutions already in studying linear systems. Sometimes, we use the notation $f_W(w) = f_X(x) * f_Y(y)$ to denote convolution.

Quiz 6.2 *Let X and Y be independent exponential random variables with expected values $E[X] = 1/3$ and $E[Y] = 1/2$. Find the PDF of $W = X + Y$.*

6.3 Moment Generating Functions

The PDF of the sum of independent random variables $X_1, \ldots, X_n$ is a sequence of convolutions involving PDFs $f_{X_1}(x)$, $f_{X_2}(x)$, and so on. In linear system theory, convolution in the time domain corresponds to multiplication in the frequency domain with time functions and frequency functions related by the Fourier transform. In probability theory, we can, in a similar way, use transform methods to replace the convolution of PDFs by multiplication of transforms. In the language of probability theory, the transform of a PDF or a PMF is a *moment generating function.*

Definition 6.1 ***Moment Generating Function (MGF)***
For a random variable X, the ***moment generating function (MGF)*** *of X is*

$$\phi_X(s) = E\left[e^{sX}\right].$$

Definition 6.1 applies to both discrete and continuous random variables X. What changes in going from discrete X to continuous X is the method of calculating the expected value. When X is a continuous random variable,

$$\phi_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx. \tag{6.26}$$

For a discrete random variable Y, the MGF is

$$\phi_Y(s) = \sum_{y_i \in S_Y} e^{sy_i} P_Y(y_i). \tag{6.27}$$

Equation (6.26) indicates that the MGF of a continuous random variable is similar to the Laplace transform of a time function. The primary difference is that the MGF is defined for real values of s. For a given random variable X, there is a range of possible values of s for which $\phi_X(s)$ exists. The set of values of s for which $\phi_X(s)$ exists is called the *region of convergence*. For example, if X is a nonnegative random variable, the region of convergence

Random Variable	PMF or PDF	MGF $\phi_X(s)$
Bernoulli (p)	$P_X(x) = \begin{cases} 1-p & x=0 \\ p & x=1 \\ 0 & \text{otherwise} \end{cases}$	$1-p+pe^s$
Binomial (n, p)	$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$	$(1-p+pe^s)^n$
Geometric (p)	$P_X(x) = \begin{cases} p(1-p)^{x-1} & x=1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{pe^s}{1-(1-p)e^s}$
Pascal (k, p)	$P_X(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$	$(\dfrac{pe^s}{1-(1-p)e^s})^k$
Poisson (α)	$P_X(x) = \begin{cases} \alpha^x e^{-\alpha}/x! & x=0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$	$e^{\alpha(e^s-1)}$
Disc. Uniform (k, l)	$P_X(x) = \begin{cases} \frac{1}{l-k+1} & x=k,k+1,\ldots,l \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{e^{sk}-e^{s(l+1)}}{1-e^s}$
Constant (a)	$f_X(x) = \delta(x-a)$	e^{sa}
Uniform (a, b)	$f_X(x) = \begin{cases} \frac{1}{b-a} & a<x<b \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{e^{bs}-e^{as}}{s(b-a)}$
Exponential (λ)	$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{\lambda}{\lambda-s}$
Erlang (n, λ)	$f_X(x) = \begin{cases} \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$(\dfrac{\lambda}{\lambda-s})^n$
Gaussian (μ, σ)	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$	$e^{s\mu+s^2\sigma^2/2}$

Table 6.1 Moment generating function for families of random variables.

includes all $s \leq 0$. Because the MGF and PMF or PDF form a transform pair, the MGF is also a complete probability model of a random variable. Given the MGF, it is possible to compute the PDF or PMF. The definition of the MGF implies that $\phi_X(0) = E[e^0] = 1$. Moreover, the derivatives of $\phi_X(s)$ evaluated at $s = 0$ are the moments of X.

Theorem 6.6 *A random variable X with MGF $\phi_X(s)$ has nth moment*

$$E\left[X^n\right] = \left.\frac{d^n \phi_X(s)}{ds^n}\right|_{s=0}.$$

Proof The first derivative of $\phi_X(s)$ is

$$\frac{d\phi_X(s)}{ds} = \frac{d}{ds}\left(\int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx\right) = \int_{-\infty}^{\infty} x e^{sx} f_X(x)\,dx. \tag{6.28}$$

Evaluating this derivative at $s = 0$ proves the theorem for $n = 1$.

$$\left.\frac{d\phi_X(s)}{ds}\right|_{s=0} = \int_{-\infty}^{\infty} x f_X(x)\,dx = E[X]. \tag{6.29}$$

Similarly, the nth derivative of $\phi_X(s)$ is

$$\frac{d^n\phi_X(s)}{ds^n} = \int_{-\infty}^{\infty} x^n e^{sx} f_X(x)\,dx. \tag{6.30}$$

The integral evaluated at $s = 0$ is the formula in the theorem statement.

Typically it is easier to calculate the moments of X by finding the MGF and differentiating than by integrating $x^n f_X(x)$.

Example 6.5 X is an exponential random variable with MGF $\phi_X(s) = \lambda/(\lambda - s)$. What are the first and second moments of X? Write a general expression for the nth moment.

The first moment is the expected value:

$$E[X] = \left.\frac{d\phi_X(s)}{ds}\right|_{s=0} = \left.\frac{\lambda}{(\lambda - s)^2}\right|_{s=0} = \frac{1}{\lambda}. \tag{6.31}$$

The second moment of X is the mean square value:

$$E\left[X^2\right] = \left.\frac{d^2\phi_X(s)}{ds^2}\right|_{s=0} = \left.\frac{2\lambda}{(\lambda - s)^3}\right|_{s=0} = \frac{2}{\lambda^2}. \tag{6.32}$$

Proceeding in this way, it should become apparent that the nth moment of X is

$$E\left[X^n\right] = \left.\frac{d^n\phi_X(s)}{ds^n}\right|_{s=0} = \left.\frac{n!\lambda}{(\lambda - s)^{n+1}}\right|_{s=0} = \frac{n!}{\lambda^n}. \tag{6.33}$$

Table 6.1 presents the MGF for the families of random variables defined in Chapters 2 and 3. The following theorem derives the MGF of a linear transformation of a random variable X in terms of $\phi_X(s)$.

Theorem 6.7 *The MGF of $Y = aX + b$ is $\phi_Y(s) = e^{sb}\phi_X(as)$.*

Proof From the definition of the MGF,

Quiz 6.3

6.4 M

Theorem

$$\phi_Y(s) = E\left[e^{s(aX+b)}\right] = e^{sb} E\left[e^{(as)X}\right] = e^{sb}\phi_X(as). \tag{6.34}$$

Quiz 6.3 *Random variable K has PMF*

$$P_K(k) = \begin{cases} 0.2 & k = 0, \ldots, 4, \\ 0 & \text{otherwise}. \end{cases} \tag{6.35}$$

Use the MGF $\phi_K(s)$ to find the first, second, third, and fourth moments of K.

6.4 MGF of the Sum of Independent Random Variables

Moment generating functions are particularly useful for analyzing sums of independent random variables, because if X and Y are independent, the MGF of $W = X + Y$ is the product:

$$\phi_W(s) = E\left[e^{sX}e^{sY}\right] = E\left[e^{sX}\right]E\left[e^{sY}\right] = \phi_X(s)\phi_Y(s). \tag{6.36}$$

Theorem 6.8 generalizes this result to a sum of n independent random variables.

Theorem 6.8 *For a set of independent random variables $X_1, \ldots, X_n$, the moment generating function of $W = X_1 + \cdots + X_n$ is*

$$\phi_W(s) = \phi_{X_1}(s)\phi_{X_2}(s)\cdots\phi_{X_n}(s).$$

When $X_1, \ldots, X_n$ are iid, each with MGF $\phi_{X_i}(s) = \phi_X(s)$,

$$\phi_W(s) = [\phi_X(s)]^n.$$

Proof From the definition of the MGF,

$$\phi_W(s) = E\left[e^{s(X_1+\cdots+X_n)}\right] = E\left[e^{sX_1}e^{sX_2}\cdots e^{sX_n}\right]. \tag{6.37}$$

Here, we have the expected value of a product of functions of independent random variables. Theorem 5.9 states that this expected value is the product of the individual expected values:

$$E\left[g_1(X_1)g_2(X_2)\cdots g_n(X_n)\right] = E\left[g_1(X_1)\right]E\left[g_2(X_2)\right]\cdots E\left[g_n(X_n)\right]. \tag{6.38}$$

By Equation (6.38) with $g_i(X_i) = e^{sX_i}$, the expected value of the product is

$$\phi_W(s) = E\left[e^{sX_1}\right]E\left[e^{sX_2}\right]\cdots E\left[e^{sX_n}\right] = \phi_{X_1}(s)\phi_{X_2}(s)\cdots\phi_{X_n}(s). \tag{6.39}$$

When $X_1, \ldots, X_n$ are iid, $\phi_{X_i}(s) = \phi_X(s)$ and thus $\phi_W(s) = (\phi_W(s))^n$.

Moment generating functions provide a convenient way to study the properties of sums of independent finite discrete random variables.

Example 6.6 J and K are independent random variables with probability mass functions

$$P_J(j) = \begin{cases} 0.2 & j = 1, \\ 0.6 & j = 2, \\ 0.2 & j = 3, \\ 0 & \text{otherwise,} \end{cases} \qquad P_K(k) = \begin{cases} 0.5 & k = -1, \\ 0.5 & k = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.40}$$

Find the MGF of $M = J + K$? What are $E[M^3]$ and $P_M(m)$?

J and K have have moment generating functions

$$\phi_J(s) = 0.2e^s + 0.6e^{2s} + 0.2e^{3s}, \qquad \phi_K(s) = 0.5e^{-s} + 0.5e^s. \tag{6.41}$$

Therefore, by Theorem 6.8, $M = J + K$ has MGF

$$\phi_M(s) = \phi_J(s)\phi_K(s) = 0.1 + 0.3e^s + 0.2e^{2s} + 0.3e^{3s} + 0.1e^{4s}. \tag{6.42}$$

To find the third moment of M, we differentiate $\phi_M(s)$ three times:

$$E\left[M^3\right] = \left.\frac{d^3\phi_M(s)}{ds^3}\right|_{s=0} \tag{6.43}$$

$$= 0.3e^s + 0.2(2^3)e^{2s} + 0.3(3^3)e^{3s} + 0.1(4^3)e^{4s}\Big|_{s=0} = 16.4. \tag{6.44}$$

The value of $P_M(m)$ at any value of m is the coefficient of e^{ms} in $\phi_M(s)$:

$$\phi_M(s) = E\left[e^{sM}\right] = \underbrace{0.1}_{P_M(0)} + \underbrace{0.3}_{P_M(1)} e^s + \underbrace{0.2}_{P_M(2)} e^{2s} + \underbrace{0.3}_{P_M(3)} e^{3s} + \underbrace{0.1}_{P_M(4)} e^{4s}. \tag{6.45}$$

The complete expression for the PMF of M is

$$P_M(m) = \begin{cases} 0.1 & m = 0, 4, \\ 0.3 & m = 1, 3, \\ 0.2 & m = 2, \\ 0 & \text{otherwise.} \end{cases} \tag{6.46}$$

Besides enabling us to calculate probabilities and moments for sums of discrete random variables, we can also use Theorem 6.8 to derive the PMF or PDF of certain sums of iid random variables. In particular, we use Theorem 6.8 to prove that the sum of independent Poisson random variables is a Poisson random variable, and the sum of independent Gaussian random variables is a Gaussian random variable.

Theorem 6.9 *If $K_1, \ldots, K_n$ are independent Poisson random variables, $W = K_1 + \cdots + K_n$ is a Poisson random variable.*

Proof We adopt the notation $E[K_i] = \alpha_i$ and note in Table 6.1 that K_i has MGF $\phi_{K_i}(s) = e^{\alpha_i(e^s-1)}$.

By Theorem 6.8,

$$\phi_W(s) = e^{\alpha_1(e^s-1)}e^{\alpha_2(e^s-1)}\cdots e^{\alpha_n(e^s-1)} = e^{(\alpha_1+\cdots+\alpha_n)(e^s-1)} = e^{(\alpha_T)(e^s-1)} \tag{6.47}$$

where $\alpha_T = \alpha_1 + \cdots + \alpha_n$. Examining Table 6.1, we observe that $\phi_W(s)$ is the moment generating function of the Poisson (α_T) random variable. Therefore,

$$P_W(w) = \begin{cases} \alpha_T^w e^{-\alpha}/w! & w = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases} \tag{6.48}$$

Theorem 6.10 *The sum of n independent Gaussian random variables $W = X_1 + \cdots + X_n$ is a Gaussian random variable.*

Proof For convenience, let $\mu_i = E[X_i]$ and $\sigma_i^2 = \text{Var}[X_i]$. Since the X_i are independent, we know that

$$\phi_W(s) = \phi_{X_1}(s)\phi_{X_2}(s)\cdots\phi_{X_n}(s) \tag{6.49}$$

$$= e^{s\mu_1+\sigma_1^2 s^2/2}e^{s\mu_2+\sigma_2^2 s^2/2}\cdots e^{s\mu_n+\sigma_n^2 s^2/2} \tag{6.50}$$

$$= e^{s(\mu_1+\cdots+\mu_n)+(\sigma_1^2+\cdots+\sigma_n^2)s^2/2}. \tag{6.51}$$

From Equation (6.51), we observe that $\phi_W(s)$ is the moment generating function of a Gaussian random variable with expected value $\mu_1 + \cdots + \mu_n$ and variance $\sigma_1^2 + \cdots + \sigma_n^2$.

In general, the sum of independent random variables in one family is a different kind of random variable. The following theorem shows that the Erlang (n, λ) random variable is the sum of n independent exponential (λ) random variables.

Theorem 6.11 *If $X_1, \ldots, X_n$ are iid exponential (λ) random variables, then $W = X_1 + \cdots + X_n$ has the Erlang PDF*

$$f_W(w) = \begin{cases} \frac{\lambda^n w^{n-1} e^{-\lambda w}}{(n-1)!} & w \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Proof In Table 6.1 we observe that each X_i has MGF $\phi_X(s) = \lambda/(\lambda - s)$. By Theorem 6.8, W has MGF

$$\phi_W(s) = \left(\frac{\lambda}{\lambda - s}\right)^n. \tag{6.52}$$

Returning to Table 6.1, we see that W has the MGF of an Erlang (n, λ) random variable.

Similar reasoning demonstrates that the sum of n Bernoulli (p) random variables is the binomial (n, p) random variable, and that the sum of k geometric (p) random variables is a Pascal (k, p) random variable.

Quiz 6.4

(A) Let $K_1, K_2, \ldots, K_m$ be iid discrete uniform random variables with PMF

$$P_K(k) = \begin{cases} 1/n & k = 1, 2, \ldots, n, \\ 0 & \text{otherwise}. \end{cases} \tag{6.53}$$

Find the MGF of $J = K_1 + \cdots + K_m$.

(B) Let $X_1, \ldots, X_n$ be independent Gaussian random variables with $E[X_i = 0]$ and $\text{Var}[X_i] = i$. Find the PDF of

$$W = \alpha X_1 + \alpha^2 X_2 + \cdots + \alpha^n X_n. \tag{6.54}$$

6.5 Random Sums of Independent Random Variables

Many practical problems can be analyzed by reference to a sum of iid random variables in which the number of terms in the sum is also a random variable. We refer to the resultant random variable, R, as a *random sum* of iid random variables. Thus, given a random variable N and a sequence of iid random variables $X_1, X_2, \ldots$, let

$$R = X_1 + \cdots + X_N. \tag{6.55}$$

The following two examples describe experiments in which the observations are random sums of random variables.

Example 6.7 At a bus terminal, count the number of people arriving on buses during one minute. If the number of people on the ith bus is K_i and the number of arriving buses is N, then the number of people arriving during the minute is

$$R = K_1 + \cdots + K_N. \tag{6.56}$$

In general, the number N of buses that arrive is a random variable. Therefore, R is a random sum of random variables.

Example 6.8 Count the number N of data packets transmitted over a communications link in one minute. Suppose each packet is successfully decoded with probability p, independent of the decoding of any other packet. The number of successfully decoded packets in the one-minute span is

$$R = X_1 + \cdots + X_N. \tag{6.57}$$

where X_i is 1 if the ith packet is decoded correctly and 0 otherwise. Because the number N of packets transmitted is random, R is not the usual binomial random variable.

In the preceding examples we can use the methods of Chapter 4 to find the joint PMF $P_{N,R}(n, r)$. However, we are not able to find a simple closed form expression for the PMF $P_R(r)$. On the other hand, we see in the next theorem that it is possible to express the probability model of R as a formula for the moment generating function $\phi_R(s)$.

Theorem 6.12

Exam

Example 6.10 Let $X_1, X_2 \ldots$ be a sequence of independent Gaussian (100,10) random variables. If K is a Poisson (1) random variable independent of $X_1, X_2 \ldots$, find the expected value and variance of $R = X_1 + \cdots + X_K$.

The PDF and MGF of R are complicated. However, Theorem 6.13 simplifies the calculation of the expected value and the variance. From Appendix A, we observe that a Poisson (1) random variable also has variance 1. Thus

$$E[R] = E[X]\,E[K] = 100, \tag{6.69}$$

and

$$\mathrm{Var}[R] = E[K]\,\mathrm{Var}[X] + \mathrm{Var}[K]\,(E[X])^2 = 100 + (100)^2 = 10,100. \tag{6.70}$$

We see that most of the variance is contributed by the randomness in K. This is true because K is very likely to take on the values 0 and 1, and those two choices dramatically affect the sum.

Quiz 6.5 *Let $X_1, X_2, \ldots$ denote a sequence of iid random variables with exponential PDF*

$$f_X(x) = \begin{cases} e^{-x} & x \geq 0, \\ 0 & \textit{otherwise}. \end{cases} \tag{6.71}$$

Let N denote a geometric (1/5) random variable.

(1) What is the MGF of $R = X_1 + \cdots + X_N$?

(2) Find the PDF of R.

6.6 Central Limit Theorem

Probability theory provides us with tools for interpreting observed data. In many practical situations, both discrete PMFs and continuous PDFs approximately follow a *bell-shaped curve*. For example, Figure 6.1 shows the binomial $(n, 1/2)$ PMF for $n = 5$, $n = 10$ and $n = 20$. We see that as n gets larger, the PMF more closely resembles a bell-shaped curve. Recall that in Section 3.5, we encountered a bell-shaped curve as the PDF of a Gaussian random variable. The central limit theorem explains why so many practical phenomena produce data that can be modeled as Gaussian random variables.

We will use the central limit theorem to estimate probabilities associated with the iid sum $W_n = X_1 + \cdots + X_n$. However, as n approaches infinity, $E[W_n] = n\mu_X$ and $\mathrm{Var}[W_n] = n\,\mathrm{Var}[X]$ approach infinity, which makes it difficult to make a mathematical statement about the convergence of the CDF $F_{W_n}(w)$. Hence our formal statement of the central limit theorem will be in terms of the standardized random variable

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu_X}{\sqrt{n\sigma_X^2}}. \tag{6.72}$$

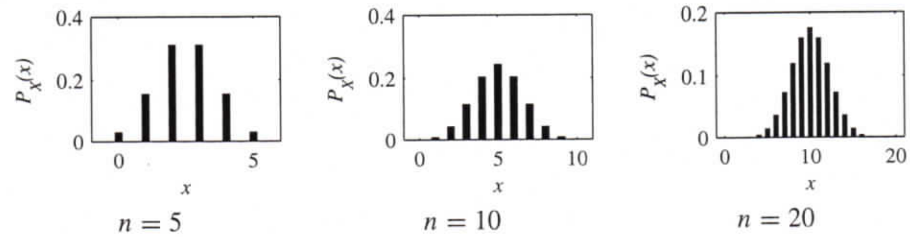


Figure 6.1 The PMF of the X, the number of heads in n coin flips for $n = 5, 10, 20$. As n increases, the PMF more closely resembles a bell-shaped curve.

We say the sum Z_n is standardized since for all n

$$E[Z_n] = 0, \qquad \text{Var}[Z_n] = 1. \tag{6.73}$$

Theorem 6.14 ***Central Limit Theorem***

Given $X_1, X_2, \ldots$, a sequence of iid random variables with expected value μ_X and variance σ_X^2, the CDF of $Z_n = (\sum_{i=1}^{n} X_i - n\mu_X)/\sqrt{n\sigma_X^2}$ has the property

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z).$$

The proof of this theorem is beyond the scope of this text. In addition to Theorem 6.14, there are other central limit theorems, each with its own statement of the sums W_n. One remarkable aspect of Theorem 6.14 and its relatives is the fact that there are no restrictions on the nature of the random variables X_i in the sum. They can be continuous, discrete, or mixed. In all cases the CDF of their sum more and more resembles a Gaussian CDF as the number of terms in the sum increases. Some versions of the central limit theorem apply to sums of sequences X_i that are not even iid.

To use the central limit theorem, we observe that we can express the iid sum $W_n = X_1 + \cdots + X_n$ as

$$W_n = \sqrt{n\sigma_X^2}\, Z_n + n\mu_X. \tag{6.74}$$

The CDF of W_n can be expressed in terms of the CDF of Z_n as

$$F_{W_n}(w) = P\left[\sqrt{n\sigma_X^2}\, Z_n + n\mu_X \le w\right] = F_{Z_n}\left(\frac{w - n\mu_X}{\sqrt{n\sigma_X^2}}\right). \tag{6.75}$$

For large n, the central limit theorem says that $F_{Z_n}(z) \approx \Phi(z)$. This approximation is the basis for practical applications of the central limit theorem.

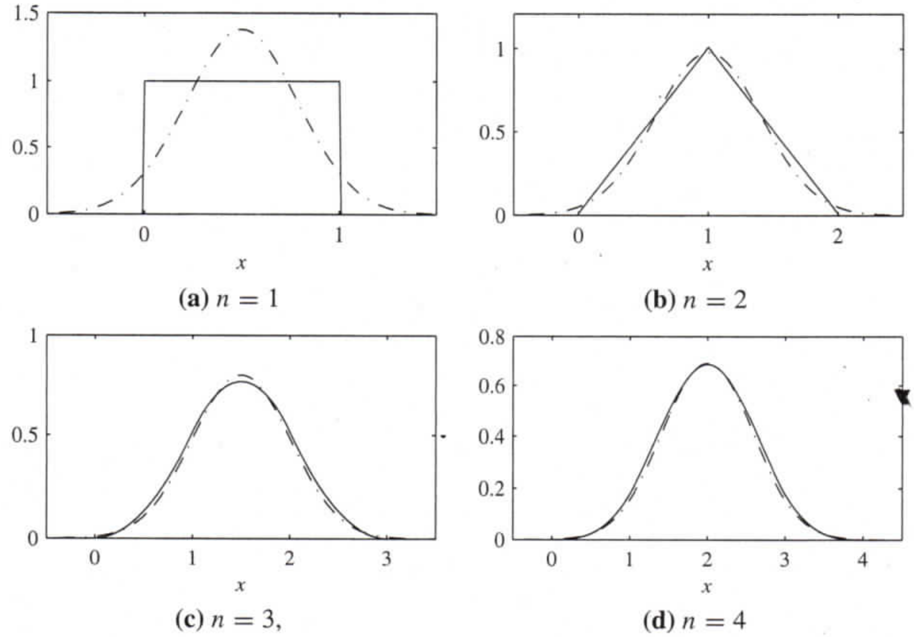


Figure 6.2 The PDF of W_n, the sum of n uniform (0, 1) random variables, and the corresponding central limit theorem approximation for $n = 1, 2, 3, 4$. The solid — line denotes the PDF $f_{W_n}(w)$, while the $- \cdot -$ line denotes the Gaussian approximation.

Definition 6.2 ***Central Limit Theorem Approximation***

Let $W_n = X_1 + \cdots + X_n$ be the sum of n iid random variables, each with $E[X] = \mu_X$ and $\mathrm{Var}[X] = \sigma_X^2$. The central limit theorem approximation to the CDF of W_n is

$$F_{W_n}(w) \approx \Phi\left(\frac{w - n\mu_X}{\sqrt{n\sigma_X^2}}\right).$$

We often call Definition 6.2 a Gaussian approximation for W_n.

Example 6.11 To gain some intuition into the central limit theorem, consider a sequence of iid continuous random variables X_i, where each random variable is uniform (0,1). Let

$$W_n = X_1 + \cdots + X_n. \tag{6.76}$$

Recall that $E[X] = 0.5$ and $\mathrm{Var}[X] = 1/12$. Therefore, W_n has expected value $E[W_n] = n/2$ and variance $n/12$. The central limit theorem says that the CDF of W_n should approach a Gaussian CDF with the same expected value and variance. Moreover, since W_n is a continuous random variable, we would also expect that the PDF of W_n would converge to a Gaussian PDF. In Figure 6.2, we compare the PDF of W_n to the PDF of a Gaussian random variable with the same expected value and variance. First, W_1 is a uniform random variable with the rectangular PDF shown in Figure 6.2(a). This figure also shows the PDF of W_1, a Gaussian random variable with expected

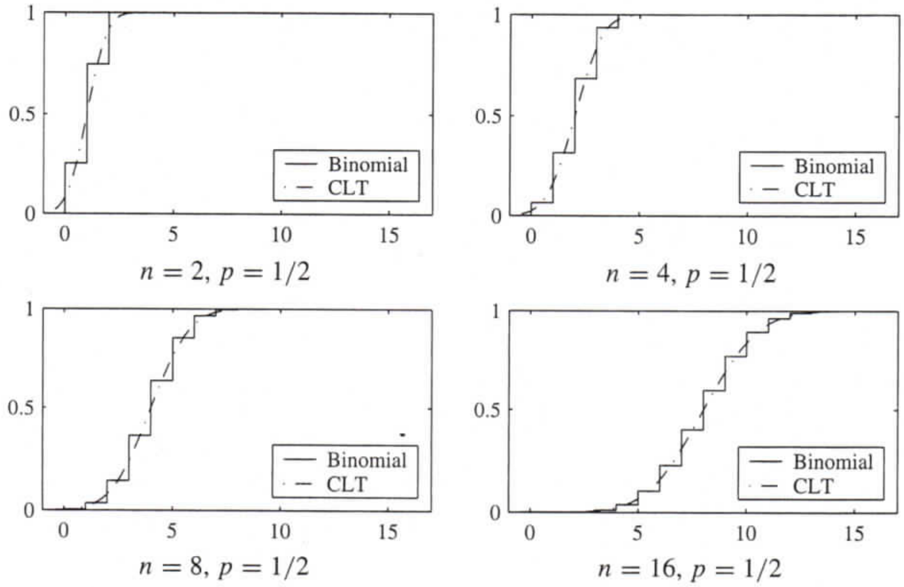


Figure 6.3 The binomial (n, p) CDF and the corresponding central limit theorem approximation for $n = 4, 8, 16, 32$, and $p = 1/2$.

value $\mu = 0.5$ and variance $\sigma^2 = 1/12$. Here the PDFs are very dissimilar. When we consider $n = 2$, we have the situation in Figure 6.2(b). The PDF of W_2 is a triangle with expected value 1 and variance 2/12. The figure shows the corresponding Gaussian PDF. The following figures show the PDFs of $W_3, \ldots, W_6$. The convergence to a bell shape is apparent.

Example 6.12 Now suppose $W_n = X_1 + \cdots + X_n$ is a sum of independent Bernoulli (p) random variables. We know that W_n has the binomial PMF

$$P_{W_n}(w) = \binom{n}{w} p^w (1-p)^{n-w}. \tag{6.77}$$

No matter how large n becomes, W_n is always a discrete random variable and would have a PDF consisting of impulses. However, the central limit theorem says that *the CDF* of W_n converges to a Gaussian CDF. Figure 6.3 demonstrates the convergence of the sequence of binomial CDFs to a Gaussian CDF for $p = 1/2$ and four values of n, the number of Bernoulli random variables that are added to produce a binomial random variable. For $n \geq 32$, Figure 6.3 suggests that approximations based on the Gaussian distribution are very accurate.

Quiz 6.6 *The random variable X milliseconds is the total access time (waiting time + read time) to get one block of information from a computer disk. X is uniformly distributed between 0 and 12 milliseconds. Before performing a certain task, the computer must access 12 different* *blocks of information from the disk. (Access times for different blocks are independent of one another.) The total access time for all the information is a random variable A milliseconds.*

(1) What is $E[X]$, the expected value of the access time?

(2) What is $\text{Var}[X]$, the variance of the access time?

(3) What is $E[A]$, the expected value of the total access time?

(4) What is σ_A, the standard deviation of the total access time?

(5) Use the central limit theorem to estimate $P[A > 75 \text{ ms}]$, the probability that the total access time exceeds 75 ms.

(6) Use the central limit theorem to estimate $P[A < 48 \text{ ms}]$, the probability that the total access time is less than 48 ms.

6.7 Applications of the Central Limit Theorem

In addition to helping us understand why we observe bell-shaped curves in so many situations, the central limit theorem makes it possible to perform quick, accurate calculations that would otherwise be extremely complex and time consuming. In these calculations, the random variable of interest is a sum of other random variables, and we calculate the probabilities of events by referring to the corresponding Gaussian random variable. In the following example, the random variable of interest is the average of eight iid uniform random variables. The expected value and variance of the average are easy to obtain. However, a complete probability model is extremely complex (it consists of segments of eighth-order polynomials).

Example 6.13 A compact disc (CD) contains digitized samples of an acoustic waveform. In a CD player with a "one bit digital to analog converter," each digital sample is represented to an accuracy of ± 0.5 mV. The CD player "oversamples" the waveform by making eight independent measurements corresponding to each sample. The CD player obtains a waveform sample by calculating the average (sample mean) of the eight measurements. What is the probability that the error in the waveform sample is greater than 0.1 mV?

The measurements $X_1, X_2, \ldots, X_8$ all have a uniform distribution between $v - 0.5$ mV and $v + 0.5$ mV, where v mV is the exact value of the waveform sample. The compact disk player produces the output $U = W_8/8$, where

$$W_8 = \sum_{i=1}^{8} X_i. \tag{6.78}$$

To find $P[|U - v| > 0.1]$ exactly, we would have to find an exact probability model for W_8, either by computing an eightfold convolution of the uniform PDF of X_i or by using the moment generating function. Either way, the process is extremely complex. Alternatively, we can use the central limit theorem to model W_8 as a Gaussian random variable with $E[W_8] = 8\mu_X = 8v$ mV and variance $\text{Var}[W_8] = 8\,\text{Var}[X] = 8/12$.

Therefore, U is approximately Gaussian with $E[U] = E[W_8]/8 = v$ and variance $\text{Var}[W_8]/64 = 1/96$. Finally, the error, $U - v$ in the output waveform sample is approximately Gaussian with expected value 0 and variance $1/96$. It follows that

$$P[|U - v| > 0.1] = 2\left[1 - \Phi\left(0.1/\sqrt{1/96}\right)\right] = 0.3272. \tag{6.79}$$

The central limit theorem is particularly useful in calculating events related to binomial random variables. Figure 6.3 from Example 6.12 indicates how the CDF of a sum of n Bernoulli random variables converges to a Gaussian CDF. When n is very high, as in the next two examples, probabilities of events of interest are sums of thousands of terms of a binomial CDF. By contrast, each of the Gaussian approximations requires looking up only one value of the Gaussian CDF $\Phi(x)$.

Example 6.14 A modem transmits one million bits. Each bit is 0 or 1 independently with equal probability. Estimate the probability of at least 502,000 ones.

Let X_i be the value of bit i (either 0 or 1). The number of ones in one million bits is $W = \sum_{i=1}^{10^6} X_i$. Because X_i is a Bernoulli (0.5) random variable, $E[X_i] = 0.5$ and $\text{Var}[X_i] = 0.25$ for all i. Note that $E[W] = 10^6 E[X_i] = 500{,}000$ and $\text{Var}[W] = 10^6\,\text{Var}[X_i] = 250{,}000$. Therefore, $\sigma_W = 500$. By the central limit theorem approximation,

$$P[W \geq 502{,}000] = 1 - P[W \leq 502{,}000] \tag{6.80}$$

$$\approx 1 - \Phi\left(\frac{502{,}000 - 500{,}000}{500}\right) = 1 - \Phi(4). \tag{6.81}$$

Using Table 3.1, we observe that $1 - \Phi(4) = Q(4) = 3.17 \times 10^{-5}$.

Example 6.15 Transmit one million bits. Let A denote the event that there are at least 499,000 ones but no more than 501,000 ones. What is $P[A]$?

As in Example 6.14, $E[W] = 500{,}000$ and $\sigma_W = 500$. By the central limit theorem approximation,

$$P[A] = P[W \leq 501{,}000] - P[W < 499{,}000] \tag{6.82}$$

$$\approx \Phi\left(\frac{501{,}000 - 500{,}000}{500}\right) - \Phi\left(\frac{499{,}000 - 500{,}000}{500}\right) \tag{6.83}$$

$$= \Phi(2) - \Phi(-2) = 0.9544 \tag{6.84}$$

These examples of using a Gaussian approximation to a binomial probability model contain events that consist of thousands of outcomes. When the events of interest contain a small number of outcomes, the accuracy of the approximation can be improved by accounting for the fact that the Gaussian random variable is continuous whereas the corresponding binomial random variable is discrete.

In fact, using a Gaussian approximation to a discrete random variable is fairly common. We recall that the sum of n Bernoulli random variables is binomial, the sum of n geometric random variables is Pascal, and the sum of n Bernoulli random variables (each with success probability λ/n) approaches a Poisson random variable in the limit as $n \to \infty$. Thus a Gaussian approximation can be accurate for a random variable K that is binomial, Pascal, or Poisson.

In general, suppose K is a discrete random variable and that the range of K is $S_K \subset \{n\tau | n = 0, \pm 1, \pm 2 \ldots\}$. For example, when K is binomial, Poisson, or Pascal, $\tau = 1$ and $S_K = \{0, 1, 2 \ldots\}$. We wish to estimate the probability of the event $A = \{k_1 \leq K \leq k_2\}$, where k_1 and k_2 are integers. A Gaussian approximation to $P[A]$ is often poor when k_1 and k_2 are close to one another. In this case, we can improve our approximation by accounting for the discrete nature of K. Consider the Gaussian random variable, X with expected value $E[K]$ and variance $\text{Var}[K]$. An accurate approximation to the probability of the event A is

$$P[A] \approx P[k_1 - \tau/2 \leq X \leq k_2 + \tau/2] \tag{6.85}$$

$$= \Phi\left(\frac{k_2 + \tau/2 - E[K]}{\sqrt{\text{Var}[K]}}\right) - \Phi\left(\frac{k_1 - \tau/2 - E[K]}{\sqrt{\text{Var}[K]}}\right). \tag{6.86}$$

When K is a binomial random variable for n trials and success probability p, $E[K] = np$, and $\text{Var}[K] = np(1-p)$. The formula that corresponds to this statement is known as the De Moivre–Laplace formula. It corresponds to the formula for $P[A]$ with $\tau = 1$.

Definition 6.3 ***De Moivre–Laplace Formula***

For a binomial (n, p) random variable K,

$$P[k_1 \leq K \leq k_2] \approx \Phi\left(\frac{k_2 + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 0.5 - np}{\sqrt{np(1-p)}}\right).$$

To appreciate why the ± 0.5 terms increase the accuracy of approximation, consider the following simple but dramatic example in which $k_1 = k_2$.

Example 6.16 Let K be a binomial $(n = 20, p = 0.4)$ random variable. What is $P[K = 8]$?

Since $E[K] = np = 8$ and $\text{Var}[K] = np(1-p) = 4.8$, the central limit theorem approximation to K is a Gaussian random variable X with $E[X] = 8$ and $\text{Var}[X] = 4.8$. Because X is a continuous random variable, $P[X = 8] = 0$, a useless approximation to $P[K = 8]$. On the other hand, the De Moivre–Laplace formula produces

$$P[8 \leq K \leq 8] \approx P[7.5 \leq X \leq 8.5] \tag{6.87}$$

$$= \Phi\left(\frac{0.5}{\sqrt{4.8}}\right) - \Phi\left(\frac{-0.5}{\sqrt{4.8}}\right) = 0.1803. \tag{6.88}$$

The exact value is $\binom{20}{8}(0.4)^8(1-0.4)^{12} = 0.1797$.

Example 6.17 K is the number of heads in 100 flips of a fair coin. What is $P[50 \leq K \leq 51]$?

Since K is a binomial $(n = 100, p = 1/2)$ random variable,

$$P[50 \leq K \leq 51] = P_K(50) + P_K(51) \tag{6.89}$$

$$= \binom{100}{50}\left(\frac{1}{2}\right)^{100} + \binom{100}{51}\left(\frac{1}{2}\right)^{100} = 0.1576. \tag{6.90}$$

Since $E[K] = 50$ and $\sigma_K = \sqrt{np(1-p)} = 5$, the ordinary central limit theorem approximation produces

$$P[50 \leq K \leq 51] \approx \Phi\left(\frac{51-50}{5}\right) - \Phi\left(\frac{50-50}{5}\right) = 0.0793. \qquad (6.91)$$

This approximation error of roughly 50% occurs because the ordinary central limit theorem approximation ignores the fact that the discrete random variable K has two probability masses in an interval of length 1. As we see next, the De Moivre–Laplace approximation is far more accurate.

$$P[50 \leq K \leq 51] \approx \Phi\left(\frac{51+0.5-50}{5}\right) - \Phi\left(\frac{50-0.5-50}{5}\right) \qquad (6.92)$$

$$= \Phi(0.3) - \Phi(-0.1) = 0.1577. \qquad (6.93)$$

Although the central limit theorem approximation provides a useful means of calculating events related to complicated probability models, it has to be used with caution. When the events of interest are confined to outcomes at the edge of the range of a random variable, the central limit theorem approximation can be quite inaccurate. In all of the examples in this section, the random variable of interest has finite range. By contrast, the corresponding Gaussian models have finite probabilities for any range of numbers between $-\infty$ and ∞. Thus in Example 6.13, $P[U - v > 0.5] = 0$, while the Gaussian approximation suggests that $P[U - v > 0.5] = Q(0.5/\sqrt{1/96}) \approx 5 \times 10^{-7}$. Although this is a low probability, there are many applications in which the events of interest have very low probabilities or probabilities very close to 1. In these applications, it is necessary to resort to more complicated methods than a central limit theorem approximation to obtain useful results. In particular, it is often desirable to provide guarantees in the form of an upper bound rather than the approximation offered by the central limit theorem. In the next section, we describe one such method based on the moment generating function.

Quiz 6.7

Telephone calls can be classified as voice (V) if someone is speaking or data (D) if there is a modem or fax transmission. Based on a lot of observations taken by the telephone company, we have the following probability model: $P[V] = 3/4$, $P[D] = 1/4$. Data calls and voice calls occur independently of one another. The random variable K_n is the number of voice calls in a collection of n phone calls.

(1) What is $E[K_{48}]$, the expected number of voice calls in a set of 48 calls?

(2) What is $\sigma_{K_{48}}$, the standard deviation of the number of voice calls in a set of 48 calls?

(3) Use the central limit theorem to estimate $P[30 \leq K_{48} \leq 42]$, the probability of between 30 and 42 voice calls in a set of 48 calls.

(4) Use the De Moivre–Laplace formula to estimate $P[30 \leq K_{48} \leq 42]$.

6.8 The Chernoff Bound

We now describe an inequality called the Chernoff bound. By referring to the MGF of a random variable, the Chernoff bound provides a way to guarantee that the probability of an unusual event is small.

Theorem 6.15 ***Chernoff Bound***

For an arbitrary random variable X and a constant c,

$$P[X \geq c] \leq \min_{s \geq 0} e^{-sc} \phi_X(s).$$

Proof In terms of the unit step function, $u(x)$, we observe that

$$P[X \geq c] = \int_c^\infty f_X(x)\, dx = \int_{-\infty}^\infty u(x-c) f_X(x)\, dx. \tag{6.94}$$

For all $s \geq 0$, $u(x-c) \leq e^{s(x-c)}$. This implies

$$P[X \geq c] \leq \int_{-\infty}^\infty e^{s(x-c)} f_X(x)\, dx = e^{-sc} \int_{-\infty}^\infty e^{sx} f_X(x)\, dx = e^{-sc}\phi_X(s). \tag{6.95}$$

This inequality is true for any $s \geq 0$. Hence the upper bound must hold when we choose s to minimize $e^{-sc}\phi_X(s)$.

The Chernoff bound can be applied to any random variable. However, for small values of c, $e^{-sc}\phi_X(s)$ will be minimized by a negative value of s. In this case, the minimizing nonnegative s is $s = 0$ and the Chernoff bound gives the trivial answer $P[X \geq c] \leq 1$.

Example 6.18 If the height X, measured in feet, of a randomly chosen adult is a Gaussian (5.5, 1) random variable, use the Chernoff bound to find an upper bound on $P[X \geq 11]$.

In Table 6.1 the MGF of X is

$$\phi_X(s) = e^{(11s+s^2)/2}. \tag{6.96}$$

Thus the Chernoff bound is

$$P[X \geq 11] \leq \min_{s \geq 0} e^{-11s} e^{(11s+s^2)/2} = \min_{s \geq 0} e^{(s^2-11s)/2}. \tag{6.97}$$

To find the minimizing s, it is sufficient to choose s to minimize $h(s) = s^2 - 11s$. Setting the derivative $dh(s)/ds = 2s - 11 = 0$ yields $s = 5.5$. Applying $s = 5.5$ to the bound yields

$$P[X \geq 11] \leq e^{(s^2-11s)/2}\Big|_{s=5.5} = e^{-(5.5)^2/2} = 2.7 \times 10^{-7}. \tag{6.98}$$

Based on our model for adult heights, the actual probability (not shown in Table 3.2) is $Q(11-5.5) = 1.90 \times 10^{-8}$.

Even though the Chernoff bound is 14 times higher than the actual probability, it still conveys the information that the chance of observing someone over 11 feet tall is extremely unlikely. Simpler approximations in Chapter 7 provide bounds of $1/2$ and $1/30$ for $P[X \geq 11]$.