1

Preliminaries on Random Variables

In this chapter we recall some basic concepts and results of probability theory. The reader should already be familiar with most of this material, which is routinely taught in introductory probability courses.

Expectation, variance, and moments of random variables are introduced in Section 1.1. Some classical inequalities can be found in Section 1.2. The two fundamental limit theorems of probability – the law of large numbers and the central limit theorem – are recalled in Section 1.3.

1.1 Basic Quantities Associated with Random Variables

In basic courses in probability theory, one learns about the two most important quantities associated with a random variable X, namely the *expectation*[1] (also called the *mean*) and *variance*. They will be denoted in this book by[2]

$$\mathbb{E}\, X \quad \text{and} \quad \operatorname{Var}(X) = \mathbb{E}(X - \mathbb{E}\, X)^2.$$

Let us recall some other classical quantities and functions that describe probability distributions. The *moment generating function of X* is defined as

$$M_X(t) = \mathbb{E}\, e^{tX}, \quad t \in \mathbb{R}.$$

For $p > 0$, the *pth moment* of X is defined as $\mathbb{E}\, X^p$, and the *pth absolute moment* is $\mathbb{E}\, |X|^p$.

It is useful to take the pth root of the moments, which leads to the notion of the L^p *norm* of a random variable:

$$\|X\|_{L^p} = (\mathbb{E}\, |X|^p)^{1/p}, \quad p \in (0, \infty).$$

This definition can be extended to $p = \infty$ by the essential supremum of $|X|$:

$$\|X\|_{L^\infty} = \operatorname{ess\,sup} |X|.$$

For fixed p and a given probability space $(\Omega, \Sigma, \mathbb{P})$, the classical vector space $L^p = L^p(\Omega, \Sigma, \mathbb{P})$ consists of all random variables X on Ω with finite L^p norm, that is,

[1] If you have studied measure theory, you will recall that the expectation $\mathbb{E}\, X$ of a random variable X on a probability space $(\Omega, \Sigma, \mathbb{P})$ is, by definition, the Lebesgue integral of the function $X : \Omega \to \mathbb{R}$. This makes all theorems on Lebesgue integration applicable in probability theory for expectations of random variables.

[2] Throughout this book, we omit brackets and simply write $\mathbb{E}\, f(X)$. Thus, nonlinear functions bind before an expectation.

https://doi.org/10.1017/9781108231596.004 Published online by Cambridge University Press

$$L^p = \{X : \|X\|_{L^p} < \infty\}.$$

If $p \in [1, \infty]$, the quantity $\|X\|_{L^p}$ is a norm and L^p is a *Banach space*. This fact follows from Minkowski's inequality, which we recall in (1.4). For $p < 1$, the triangle inequality fails and $\|X\|_{L^p}$ is not a norm.

The exponent $p = 2$ is special in that L^2 is not only a Banach space but also a *Hilbert space*. The inner product and the corresponding norm on L^2 are given by

$$\langle X, Y\rangle_{L^2} = \mathbb{E}\, XY, \quad \|X\|_{L^2} = (\mathbb{E}\,|X|^2)^{1/2}. \tag{1.1}$$

Then the *standard deviation* of X can be expressed as

$$\|X - \mathbb{E}\, X\|_{L^2} = \sqrt{\mathrm{Var}(X)} = \sigma(X).$$

Similarly, we can express the *covariance* of random variables X and Y as

$$\mathrm{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}\, X)(Y - \mathbb{E}\, Y)) = \langle X - \mathbb{E}\, X, Y - \mathbb{E}\, Y\rangle_{L^2}. \tag{1.2}$$

Remark 1.1.1 (Geometry of random variables) When we consider random variables as vectors in the Hilbert space L^2, the identity (1.2) gives a *geometric interpretation* of the notion of covariance: the more the vectors $X - \mathbb{E}\, X$ and $Y - \mathbb{E}\, Y$ are aligned with each other, the larger are their inner product and covariance.

1.2 Some Classical Inequalities

Jensen's inequality states that for any random variable X and a *convex*[3] function $\varphi : \mathbb{R} \to \mathbb{R}$, we have

$$\varphi(\mathbb{E}\, X) \le \mathbb{E}\, \varphi(X).$$

As a simple consequence of Jensen's inequality, $\|X\|_{L^p}$ is an *increasing function in* p, that is

$$\|X\|_{L^p} \le \|X\|_{L^q} \quad \text{for any } 0 \le p \le q = \infty. \tag{1.3}$$

This inequality follows since $\phi(x) = x^{q/p}$ is a convex function if $q/p \ge 1$.

Minkowski's inequality states that for any $p \in [1, \infty]$ and any random variables $X, Y \in L^p$, we have

$$\|X + Y\|_{L^p} \le \|X\|_{L^p} + \|Y\|_{L^p}. \tag{1.4}$$

This can be viewed as the *triangle inequality*, which implies that $\|\cdot\|_{L^p}$ is a norm when $p \in [1, \infty]$.

The *Cauchy–Schwarz inequality* states that, for any random variables $X, Y \in L^2$, we have

$$|\mathbb{E}\, XY| \le \|X\|_{L^2} \|Y\|_{L^2}.$$

The more general *Hölder's inequality* states that if $p, q \in (1, \infty)$ are conjugate exponents, that is, $1/p + 1/q = 1$, then the random variables $X \in L^p$ and $Y \in L^q$ satisfy

[3] By definition, a function φ is *convex* if $\varphi(\lambda x + (1-\lambda)y) \le \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all $\lambda \in [0, 1]$ and all vectors x, y in the domain of φ.

$$|\mathbb{E}\, XY| \le \|X\|_{L^p}\, \|Y\|_{L^q}.$$

This inequality also holds for the pair $p = 1, q = \infty$.

As we recall from basic probability concepts, the *distribution* of a random variable X is, intuitively, the information about what values X takes with what probabilities. More rigorously, the distribution of X is determined by the *cumulative distribution function* (CDF) of X, defined as

$$F_X(t) = \mathbb{P}\left\{X \le t\right\}, \quad t \in \mathbb{R}.$$

It is often more convenient to work with the *tails* of random variables, namely with

$$\mathbb{P}\left\{X > t\right\} = 1 - F_X(t).$$

There is an important connection between the tails and the expectation (and more generally, the moments) of a random variable. The following identity is typically used to bound the expectation by the tails.

Lemma 1.2.1 (Integral identity) *Let X be a non-negative random variable X. Then*

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\left\{X > t\right\} dt.$$

The two sides of this identity are either finite or infinite simultaneously.

Proof We can represent any non-negative real number x via the identity[4]

$$x = \int_0^x 1\, dt = \int_0^\infty \mathbf{1}_{\{t<x\}}\, dt.$$

Substitute the random variable X for x and take expectation of both sides. This gives

$$\mathbb{E}\, X = \mathbb{E} \int_0^\infty \mathbf{1}_{\{t<X\}}\, dt = \int_0^\infty \mathbb{E}\, \mathbf{1}_{\{t<X\}}\, dt = \int_0^\infty \mathbb{P}\left\{t < X\right\} dt.$$

To change the order of expectation and integration in the second equality, we used the Fubini–Tonelli theorem. The proof is complete. ■

Exercise 1.2.2 (Generalization of integral identity) Prove the following extension of Lemma 1.2.1, which is valid for any random variable X (not necessarily non-negative):

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\left\{X > t\right\} dt - \int_{-\infty}^0 \mathbb{P}\left\{X < t\right\} dt.$$

Exercise 1.2.3 (pth moment via the tail) Let X be a random variable and $p \in (0, \infty)$. Show that

$$\mathbb{E}\, |X|^p = \int_0^\infty p t^{p-1}\, \mathbb{P}\left\{|X| > t\right\} dt$$

whenever the right-hand side is finite. ☞

[4] Here and later in this book, $\mathbf{1}_E$ denotes the *indicator* of the event E; it is the function that takes the value 1 if E occurs and 0 otherwise.

Another classical tool, Markov's inequality, can be used to bound the tail in terms of the expectation.

Proposition 1.2.4 (Markov's inequality) *For any non-negative random variable X and $t > 0$, we have*

$$\mathbb{P}\left\{X \geq t\right\} \leq \frac{\mathbb{E}\,X}{t}.$$

Proof Fix $t > 0$. We can represent any real number x via the identity

$$x = x\mathbf{1}_{\{x \geq t\}} + x\mathbf{1}_{\{x<t\}}.$$

Substitute the random variable X for x and take the expectation of both sides. This gives

$$\begin{aligned}\mathbb{E}\,X &= \mathbb{E}\,X\mathbf{1}_{\{X \geq t\}} + \mathbb{E}\,X\mathbf{1}_{\{X<t\}} \\ &\geq \mathbb{E}\,t\mathbf{1}_{\{X \geq t\}} + 0 = t\,\mathbb{P}\left\{X \geq t\right\}.\end{aligned}$$

Dividing both sides by t, we complete the proof. ∎

A well-known consequence of Markov's inequality is Chebyshev's inequality. It offers a better, quadratic, dependence on t and, instead of controlling a one-side tail, it quantifies the *concentration* of X about its mean.

Corollary 1.2.5 (Chebyshev's inequality) *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{|X - \mu| \geq t\right\} \leq \frac{\sigma^2}{t^2}.$$

Exercise 1.2.6☕ Deduce Chebyshev's inequality by squaring both sides of the bound $|X - \mu| \geq t$ and applying Markov's inequality.

Remark 1.2.7 In Proposition 2.5.2 we will establish relations among the three basic quantities associated with random variables – the moment generating functions, the L^p norms, and the tails.

1.3 Limit Theorems

The study of *sums of independent random variables* is a core part of classical probability theory. Recall that the identity

$$\mathrm{Var}(X_1 + \cdots + X_N) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_N)$$

holds for any independent random variables $X_1, \ldots, X_N$. If, furthermore, the X_i each have the same distribution, with mean μ and variance σ^2, then dividing both sides by N we see that

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{\sigma^2}{N}. \tag{1.5}$$

https://doi.org/10.1017/9781108231596.004 Published online by Cambridge University Press

Thus, the variance of the *sample mean* $\frac{1}{N}\sum_{i=1}^{N} X_i$ of the sample $\{X_1, \ldots, X_N\}$ shrinks to zero as $N \to \infty$. This indicates that, for large N, we should expect that the sample mean concentrates tightly about its expectation μ. One of the most important results in probability theory – the law of large numbers – states precisely this.

Theorem 1.3.1 (Strong law of large numbers) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ. Consider the sum*

$$S_N = X_1 + \cdots X_N.$$

Then, as $N \to \infty$,

$$\frac{S_N}{N} \to \mu \quad \textit{almost surely}.$$

The next result, the central limit theorem, goes one step further. It identifies the limiting distribution of the (properly scaled) sum of the X_i as the *normal* distribution, also called the *Gaussian* distribution. Recall that the *standard normal* distribution, denoted $N(0, 1)$, has density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \tag{1.6}$$

Theorem 1.3.2 (Lindeberg–Lévy central limit theorem) *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Consider the sum*

$$S_N = X_1 + \cdots + X_N$$

and normalize it to obtain a random variable with zero mean and unit variance as follows:

$$Z_N := \frac{S_N - \mathbb{E}\, S_N}{\sqrt{\operatorname{Var}(S_N)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} (X_i - \mu).$$

Then, as $N \to \infty$,

$$Z_N \to N(0, 1) \quad \textit{in distribution}.$$

Convergence in distribution means that the CDF of the normalized sum converges pointwise to the CDF of the standard normal distribution. We can express this in terms of tails as follows. Thus, for every $t \in \mathbb{R}$ we have

$$\mathbb{P}\left\{Z_N \ge t\right\} \to \mathbb{P}\left\{g \ge t\right\} = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-x^2/2}\, dx$$

as $N \to \infty$, where $g \sim N(0, 1)$ is a standard normal random variable.

Exercise 1.3.3☕ Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and finite variance. Show that

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| = O\left(\frac{1}{\sqrt{N}}\right) \quad \text{as } N \to \infty.$$

One remarkable special case of the central limit theorem occurs when the X_i are Bernoulli random variables with some fixed parameter $p \in (0, 1)$, denoted

$$X_i \sim \mathrm{Ber}(p).$$

Recall that this means that the X_i take the values 1 and 0 with probabilities p and $1 - p$ respectively; also recall that $\mathbb{E}\, X_i = p$ and $\mathrm{Var}(X_i) = p(1 - p)$. The sum

$$S_N := X_1 + \cdots + X_N$$

is said to have the *binomial distribution* $\mathrm{Binom}(N, p)$. The central limit theorem (Theorem 1.3.2) yields that, as $N \to \infty$,

$$\frac{S_N - Np}{\sqrt{Np(1-p)}} \to N(0, 1) \quad \text{in distribution.} \tag{1.7}$$

This special case of the central limit theorem is called the *de Moivre–Laplace theorem*.

Now suppose that $X_i \sim \mathrm{Ber}(p_i)$, with parameters p_i that *decay to zero* as $N \to \infty$ so fast that the sum S_N has mean $O(1)$ instead of being proportional to N. The central limit theorem fails in this regime. A different result, which we are about to state, says that S_N still converges but to the *Poisson* instead of the normal distribution.

Recall that a random variable Z has a *Poisson distribution* with parameter λ, denoted

$$Z \sim \mathrm{Pois}(\lambda),$$

if it takes values in $\{0, 1, 2, \ldots\}$ with probabilities

$$\mathbb{P}\left\{Z = k\right\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{1.8}$$

Theorem 1.3.4 (Poisson limit theorem) *Let $X_{N,i}$, $1 \le i \le N$, be independent random variables $X_{N,i} \sim \mathrm{Ber}(p_{N,i})$, and let $S_N = \sum_{i=1}^N X_{N,i}$. Assume that, as $N \to \infty$,*

$$\max_{i \le N} p_{N,i} \to 0 \quad \textit{and} \quad \mathbb{E}\, S_N = \sum_{i=1}^N p_{N,i} \to \lambda < \infty.$$

Then, as $N \to \infty$,

$$S_N \to \mathrm{Pois}(\lambda) \quad \textit{in distribution.}$$

1.4 Notes

The material presented in this chapter is included in most graduate probability textbooks. In particular, proofs of the strong law of large numbers (Theorem 1.3.1) and the Lindeberg–Lévy central limit theorem (Theorem 1.3.2) can be found e.g. in [70, Sections 1.7 and 2.4] and [22, Sections 6 and 27].