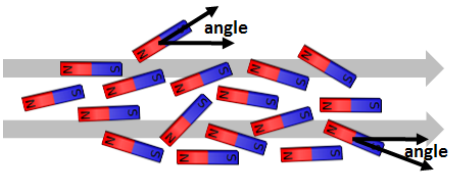
Chapter 6

Sample Statistics

When we think about probability, the first thing that likely comes to mind is flipping a coin, throwing a die, or playing a card game. These are excellent examples of the subject. However, they seldom fit in the context of modern data science, which is concerned with drawing conclusions from data. In our opinion, **the power of probability is its ability to summarize microstates using macro descriptions**. This statement will take us some effort to elaborate. We study probability because we want to analyze the uncertainties. However, when we have many data points, analyzing the uncertainties of each data point (the microstates) is computationally very difficult. Probability is useful here because it allows us to bypass the microstates and summarize the macro behavior. Instead of reporting the states of each individual, we report their sample average. Instead of offering the worst-case guarantee, we offer a probabilistic guarantee. You ask: so what? If we can offer you a performance guarantee at 99.99% confidence but one-tenth of the cost of a 100% performance guarantee, would you consider our offer? The goal of this chapter is to outline the concepts of these probabilistic arguments.

The significance of sample average

Imagine that you have a box containing many tiny magnets. (You can also think of a dataset containing two classes of labels.) In condensed matter physics, these are known as the **spin glasses**. The orientations of the magnets depend on the magnetic field. Under an extreme condition where the magnetic field is strong, all magnets will point in the same direction. When the magnetic field is not as strong, some will align with the field but some will not, as we show in **Figure 6.1**.



Experiment 1	
ID	angle
X_1	10.6°
X_2	7.5°
⋮	⋮
X_{10000}	23.5°
average	$\overline{X}_N$

...

Experiment M	
ID	angle
X_1	21.3°
X_2	17.1°
⋮	⋮
X_{10000}	3.8°
average	$\overline{X}_N$

Figure 6.1: Imagine that we have a box of magnets and we want to measure their orientation angles. The data points have individual randomness and correlations. Studying each one individually could be computationally infeasible, as we need to estimate the joint PDF $f_{X_1,\ldots,X_N}(x_1,\ldots,x_N)$ across all the data points. Probability offers a tool to summarize these individual states using a macro description. For example, we can analyze the sample average $\overline{X}_N$ of the data points and derive conclusions from the PDF of $\overline{X}_N$, i.e., $f_{\overline{X}_N}(x)$. The objective of this chapter is to present a few probabilistic tools to analyze macro descriptions, such as the sample average.

If we try to study every single magnet in this box, the correlation of the magnets

will force us to consider a joint distribution, since if one magnet points to the right it is likely that another magnet will also point to the right. The simultaneous description of all magnets is modeled through a joint probability distribution

$$f_{X_1,X_2,\ldots,X_N}(x_1, x_2, \ldots, x_N).$$

Like any joint PDF, this PDF tells us the probability density that the magnets will take a collection of states simultaneously. If N is large (say, on the order of millions), this joint distribution will be very complicated.

Since the joint PDF is very difficult to obtain computationally, physicists proposed to study the sample statistics. Instead of looking at the individual states, they look at the **sample average** of the states. If we define $X_1, \ldots, X_N$ as the states of the magnets, then the sample average is

$$\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n.$$

Since each magnet is random, the sample average is also random, and therefore it is granted a PDF:

$$f_{\overline{X}_N}(x).$$

Thus, $\overline{X}_N$ has a PDF, a mean, a variance, and so on.

We call $\overline{X}_N$ a sample statistic. It is called a **statistic** because it is a summary of the microstates, and a **sample** statistic because the statistic is based on random samples, not on the underlying theoretical distributions. We are interested in knowing the behavior of $\overline{X}_N$ because it is the summary of the observations. If we know the PDF of $\overline{X}_N$, we will know the mean, the variance, and the value of $\overline{X}_N$ when the magnetic field increases or decreases.

Why study the sample average $\overline{X}_N$?

- Analyzing individual variables is not feasible because the joint PDF can be extremely high-dimensional.
- Sample average is a macro description of the data.
- If you know the behavior of the sample average, you know most of the data.

Probabilistic guarantee versus worst-case guarantee

Besides the sample average, we are also interested in the difference between a probabilistic guarantee and a deterministic guarantee.

Consider the **birthday paradox** (see Chapter 1 for details). Suppose there are 50 students in a room. What is the probability that at least two students have the same birthday? A naive thought would suggest that we need 366 students to guarantee a pair of the same birthday because there are 365 days. So, with only 50 students, it would seem unlikely to have a pair with the same birthday. However, it turns out that with just 50 students, the probability of having at least one pair with the same birthday is more than 97%. **Figure 6.2** below shows a calculation by a computer, where we plot the estimated probability as a function of the number of students. What is more surprising is that with as few as 23 students, the probability is greater than 50%. There is no need for there to be 366 students in order to be nearly certain.

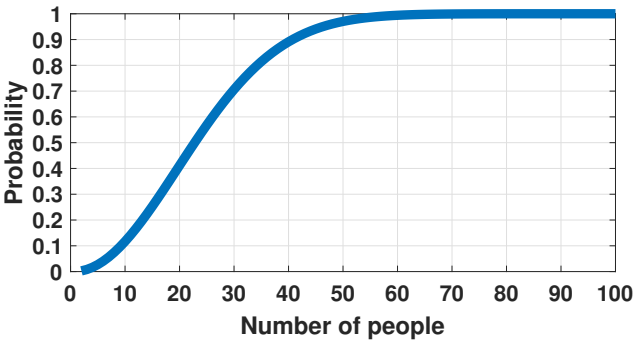


Figure 6.2: The birthday paradox asks the question of how many people we need to ask in order to have at least two of them having the same birthday. While we tend to think that the answer is 366 (because there are 365 days), the actual probability, as we have calculated (see Chapter 1), is more than 97%, even if we have only asked 50 people. The curve above shows the probability of having at least one pair of people having the same birthday as a function of the number of people. The plot highlights the gap between the **worst-case** performance and an **average-case** performance.

Why does this happen? Certainly, we can trace back to the formulae in Chapter 1 and argue through the lens of combinations and permutations. However, the more important message is about the difference between the **worst-case guarantee** and the **average-case guarantee**.

Worst case versus average case

- Worst-case guarantee: You need to ensure that the worst one is protected. This requires an exhaustive search until hitting 100%. It is a deterministic guarantee.
- Average-case guarantee: You guarantee that with a high probability (e.g., 99.99%), the undesirable event does not happen. This is a probabilistic guarantee.

Is there a difference between 99.99% and 100%? If the probability is 99.99%, there is one failure every 10,000 trials on average. You are unlikely to fail, but it is still possible. A 100% guarantee says that no matter how many trials you make you will not fail. The 99.99% guarantee is much weaker (yes, much weaker, not just a little bit weaker) than the deterministic guarantee. However, in practice, people might be willing to pay for the risk in exchange for efficiency. This is the principle behind insurance. Automobile manufacturing also uses this principle — your chance of purchasing a defective car is non-zero, but if the manufacturer can sell enough cars to compensate for the maintenance cost of fixing your car, they might be willing to offer a limited warranty in exchange for a lower selling price.

How do we analyze the probabilistic guarantee, e.g., for the sample average? Remember that the sample average $\overline{X}_N$ is a random variable. Since it is a random variable, it has a mean, variance, and PDF.[1] To measure the probabilistic guarantee, we consider the event

$$\mathcal{B} \stackrel{\text{def}}{=} \{|\overline{X}_N - \mu| \geq \epsilon\},$$

[1]Not all random variables have mean and variance, e.g., a Cauchy random variable, but most of them do.

where $\mu = \mathbb{E}[\overline{X}_N]$ is the true population mean, and $\epsilon > 0$ is a constant. This probability is illustrated in **Figure 6.3**, assuming that $\overline{X}_N$ has the PDF of a Gaussian. The probability of $\mathcal{B}$ is the two tails under the PDF. Therefore, $\mathcal{B}$ is a **bad** event because in principle $\overline{X}_N$ should be close to μ. The probability $\mathbb{P}[\mathcal{B}]$ measures situations where $\overline{X}_N$ stays very far from μ. If we can show that $\mathbb{P}[\mathcal{B}]$ is small (e.g., $< 0.01\%$), then we can say that we have obtained a probabilistic guarantee at 99.99%.

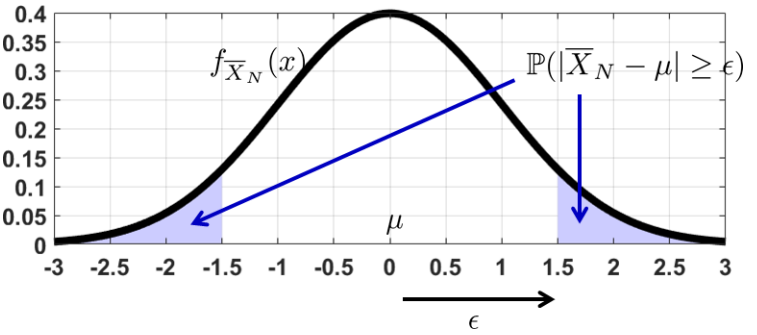


Figure 6.3: The probabilistic guarantee of a sample average $\overline{X}_N$ is established by computing the probability of the tails. In this example, we assume that $f_{\overline{X}_N}(x)$ takes a Gaussian shape, and we define $\epsilon = 1$. Anything belonging to $|\overline{X}_N - \mu| \geq \epsilon$ is called an **undesired** event $\mathcal{B}$. If the probability of an undesired event is small, we say that we can offer a probabilistic guarantee.

The moment we compute $\mathbb{P}[|\overline{X}_N - \mu| \geq \epsilon]$, we enter the race of probabilistic guarantee (e.g., 99.99%). Why? If the probability $\mathbb{P}[|\overline{X}_N - \mu| \geq \epsilon]$ is less than 0.01%, it still does not exclude the possibility that something bad will happen once every 10,000 trials on average. The chance is low, but it is still possible. We will learn some mathematical tools for analyzing this type of probabilistic guarantee.

Plan for this chapter

With these two main themes in mind, we now discuss the organization of this chapter. There are four sections: two for mathematical tools and two for main results.

- **Moment-generating functions**: We have seen in Chapter 5 that the PDF of a sum of two random variables $X + Y$ is the **convolution** of the two PDFs $f_X * f_Y$. Convolutions are non-trivial, especially when we have more random variables to sum. The moment-generating functions provide a convenient way of summing N random variables. They are the transform domain techniques (e.g., Fourier transforms). Since convolutions in time are multiplications in frequency, the moment-generating functions allow us to multiply PDFs in the transformed space. In this way, we can sum as many random variables as we want. We will discuss this idea in Section 6.1.

  > **Key Concept 1: Why study moment-generating functions?**
  > Moment-generating functions help us determine the PDF of $X_1 + X_2 + \cdots + X_N$.

- **Probability inequalities**: When analyzing sample statistics such as $\overline{X}_N$, evaluating the exact probability could be difficult because it requires integrating the PDFs. However,

if our ultimate goal is to estimate the probability, deriving an upper bound might be sufficient to achieve the goal. The probability inequalities are designed for this purpose. In Section 6.2, we discuss several of the most basic probability inequalities. We will use some of them to prove the law of large numbers.

> **Key Concept 2: How can probability inequalities be useful?**
> Probability inequalities help us upper-bound the bad event $\mathbb{P}[|\overline{X}_N - \mu| \geq \epsilon]$.

- **Law of large numbers**: This is the first main result of the chapter. The law of large numbers says that the sample average $\overline{X}_N$ converges to the population mean μ when the number of samples grows to infinity. The law of large numbers comes in two versions: the **weak** law of large numbers and the **strong** law of large numbers. The difference is the type of convergence they guarantee. The weak law is based on **convergence in probability**, whereas the strong law is based on **almost sure convergence**. We will discuss these types of convergence in Section 6.3.

> **Key Concept 3: What is the law of large numbers?**
> There is a weak law and a strong law of large numbers. The weak law of large numbers says that $\overline{X}_N$ converges to the true mean μ, as N grows:
>
> $$\lim_{N\to\infty} \mathbb{P}[|\overline{X}_N - \mu| > \epsilon] = 0.$$

- **Central Limit Theorem**: The Central Limit Theorem says that the probability of $\overline{X}_N$ can be approximated by the probability of a Gaussian. You can also think of this as saying that the CDF of $\overline{X}_N$ is converging to a distribution that can be well approximated by a bell-shaped Gaussian. If we have many random variables and their sum is becoming a Gaussian, we can ignore the individual PDFs and focus on the Gaussian. Thus it explains why Gaussian is so popular. We will discuss this theorem in detail in Section 6.4.

> **Key Concept 4: What is the Central Limit Theorem?**
> The CDF of $\overline{X}_N$ can be approximated by the CDF of a Gaussian, as N grows.

6.1 Moment-Generating and Characteristic Functions

Consider two independent random variables X and Y with PDFs $f_X(x)$ and $f_Y(y)$, respectively. Let $Z = X + Y$ be the sum of the two random variables. We know from Chapter 5 that the PDF of Z, f_Z, is the convolution of f_X and f_Y. However, we think you will agree that convolutions are not easy to compute. Especially when the sum involves more random variables, computing the convolution would be tedious. So how should we proceed in this case? One approach is to use some kind of "frequency domain" method that transforms the PDFs to another domain and then perform multiplication instead of the convolution to make the calculations easy or at least easier. The moment-generating functions and the characteristic functions are designed for this purpose.

6.1.1 Moment-generating function

Definition 6.1

For any random variable X, the **moment-generating function** (MGF) $M_X(s)$ is

$$M_X(s) = \mathbb{E}\left[e^{sX}\right]. \tag{6.1}$$

The definition says that the moment-generating function (MGF) is the expectation of the random variable taken to the power e^{sX} for some s. Effectively, it is the expectation of a **function** of random variables. The meaning of the expectation can be seen by writing out the definition. For the discrete case, the MGF is

$$M_X(s) = \sum_{x\in\Omega} e^{sx} p_X(x), \tag{6.2}$$

whereas in the continuous case, the MGF is

$$M_X(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx. \tag{6.3}$$

The continuous case should remind us of the definition of a Laplace transform. For any function $f(t)$, the Laplace transform is

$$\mathcal{L}[f](s) = \int_0^{\infty} f(t)e^{-st}\, dt.$$

From this perspective, we can interpret the MGF as the **Laplace transform** of the PDF. The argument s of the output can be regarded as the coordinate in the Laplace space. If $s = -j\omega$, then $M_X(-j\omega)$ becomes the Fourier transform of the PDF.

Example 6.1

Consider a random variable X with three states $0, 1, 2$ and with probability masses $\frac{2}{6}, \frac{3}{6}, \frac{1}{6}$ respectively. Find the MGF.

Solution. The moment-generating function is

$$\begin{aligned} M_X(s) = \mathbb{E}[e^{sX}] &= e^{s0}\cdot\frac{2}{6} + e^{s1}\cdot\frac{3}{6} + e^{s2}\cdot\frac{1}{6} \\ &= \frac{1}{3} + \frac{e^s}{2} + \frac{e^{2s}}{6}. \end{aligned}$$

Practice Exercise 6.1

Find the MGF for a Poisson random variable.

Solution. The MGF of Poisson random variable can be found as

$$M_X(s) = \mathbb{E}[e^{sX}] = \sum_{x=0}^{\infty} e^{sx}\frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=0}^{\infty}\frac{(\lambda e^s)^x}{x!}e^{-\lambda} = e^{\lambda e^s}e^{-\lambda}.$$

Practice Exercise 6.2

Find the MGF for an exponential random variable.

Solution. The MGF of an exponential random variable can be found as

$$M_X(s) = \mathbb{E}[e^{sX}] = \int_0^\infty e^{sx}\lambda e^{-\lambda x}\,dx = \int_0^\infty \lambda e^{(s-\lambda)x}\,dx = \frac{\lambda}{\lambda - s}, \qquad \text{if } \lambda > s.$$

Why are moment-generating functions so called? The following theorem reveals the reason.

Theorem 6.1

The MGF has the properties that

- $M_X(0) = 1$,
- $\frac{d}{ds}M_X(s)|_{s=0} = \mathbb{E}[X]$, $\frac{d^2}{ds^2}M_X(s)|_{s=0} = \mathbb{E}[X^2]$,
- $\frac{d^k}{ds^k}M_X(s)|_{s=0} = \mathbb{E}[X^k]$, for any positive integer k.

Proof

The first property can be proved by noting that

$$M_X(0) = \mathbb{E}[e^{0X}] = \mathbb{E}[1] = 1.$$

The third property holds because

$$\frac{d^k}{ds^k}M_X(s) = \int_{-\infty}^\infty \frac{d^k}{ds^k}e^{sx}f_X(x)\,dx = \int_{-\infty}^\infty x^k e^{sx}f_X(x)\,dx.$$

Setting $s = 0$ yields

$$\frac{d^k}{ds^k}M_X(s)|_{s=0} = \int_{-\infty}^\infty x^k f_X(x)\,dx = \mathbb{E}[X^k].$$

The second property is a special case of the third property.

The theorem tells us that if we take the derivative of the MGF and set $s = 0$, we will obtain the moment. The order of the moment depends on the order of the derivative. As a result, the MGF can "generate moments" by taking derivatives. This happens because of the exponential function e^{sx}. Since $\frac{d}{ds}e^{sx} = xe^{sx}$, the variable x appears whenever we take the derivative.

Practice Exercise 6.3

Let X be a Bernoulli random variable with parameter p. Find the first two moments using MGF.

Solution. The MGF of a Bernoulli random variable is

$$\begin{aligned} M_X(s) &= \mathbb{E}[e^{sX}] \\ &= e^{s0}p_X(0) + e^{s1}p_X(1) \\ &= (1)(1-p) + (e^s)(p) \\ &= 1 - p + pe^s. \end{aligned}$$

The first and the second moment, using the derivative approach, are

$$\mathbb{E}[X] = \frac{d}{ds}M_X(s)\bigg|_{s=0} = \frac{d}{ds}\left(1 - p + pe^s\right)\bigg|_{s=0} = pe^s\bigg|_{s=0} = p,$$
$$\mathbb{E}[X^2] = \frac{d^2}{ds^2}M_X(s)\bigg|_{s=0} = \frac{d^2}{ds^2}\left(1 - p + pe^s\right)\bigg|_{s=0} = pe^s\bigg|_{s=0} = p.$$

To facilitate our discussions of MGF, we summarize a few MGFs in the table below.

Distribution	PMF / PDF	$\mathbb{E}[X]$	$\text{Var}[X]$	$M_X(s)$
Bernoulli	$p_X(1) = p$ and $p_X(0) = 1-p$	p	$p(1-p)$	$1 - p + pe^s$
Binomial	$p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	$(1 - p + pe^s)^n$
Geometric	$p_X(k) = p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pe^s}{1-(1-p)e^s}$
Poisson	$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$e^{\lambda(e^s-1)}$
Gaussian	$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	μ	σ^2	$\exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}$
Exponential	$f_X(x) = \lambda\exp\{-\lambda x\}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - s}$
Uniform	$f_X(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb} - e^{sa}}{s(b-a)}$

Table 6.1: Moment-generating functions of common random variables.

6.1.2 Sum of independent variables via MGF

MGFs are most useful when analyzing the PDF of a sum of two random variables. The following theorem highlights the result.

Theorem 6.2

Let X and Y be independent random variables. Let $Z = X + Y$. Then

$$M_Z(s) = M_X(s)M_Y(s). \tag{6.4}$$

Proof

By the definition of MGF, we have that

$$M_Z(s) = \mathbb{E}\left[e^{s(X+Y)}\right] \overset{(a)}{=} \mathbb{E}\left[e^{sX}\right]\mathbb{E}\left[e^{sY}\right] = M_X(s)M_Y(s),$$

where (a) is valid because X and Y are independent.

Corollary 6.1

Consider independent random variables $X_1, \ldots, X_N$. Let $Z = \sum_{n=1}^{N} X_n$ be the sum of random variables. Then the MGF of Z is

$$M_Z(s) = \prod_{n=1}^{N} M_{X_n}(s). \tag{6.5}$$

If these random variables are further assumed to be identically distributed, the MGF is

$$M_Z(s) = \left(M_{X_1}(s)\right)^N. \tag{6.6}$$

Proof

This follows immediately from the previous theorem:

$$M_Z(s) = \mathbb{E}[e^{s(X_1+\cdots+X_N)}] = \mathbb{E}[e^{sX_1}]\mathbb{E}[e^{sX_2}]\cdots\mathbb{E}[e^{sX_N}] = \prod_{n=1}^{N} M_{X_n}(s).$$

If the random variables $X_1, \ldots, X_N$ are i.i.d., then the product simplifies to

$$\prod_{n=1}^{N} M_{X_n}(s) = \prod_{n=1}^{N} M_{X_1}(s) = \left(M_{X_1}(s)\right)^N.$$

Theorem 6.3 (Sum of Bernoulli = binomial)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. Bernoulli random variables with parameter p. Let $Z = X_1 + \cdots + X_N$ be the sum. Then Z is a binomial random variable with parameters (N, p).

Proof

Let us consider a sequence of i.i.d. Bernoulli random variables $X_n \sim \text{Bernoulli}(p)$ for $n = 1, \ldots, N$. Let $Z = X_1 + \cdots + X_N$. The moment-generating function of Z is

$$\begin{aligned} M_Z(s) = \mathbb{E}[e^{s(X_1+\cdots+X_N)}] &= \prod_{n=1}^{N} \mathbb{E}[e^{sX_n}] \\ &= \prod_{n=1}^{N} \left(pe^{s1} + (1-p)e^{s0}\right) = \left(pe^s + (1-p)\right)^N. \end{aligned}$$

Now, let us check the moment-generating function of a binomial random variable: If $Z \sim \text{Binomial}(N, p)$, then

$$\begin{aligned} M_Z(s) = \mathbb{E}[e^{sZ}] &= \sum_{k=0}^{N} e^{sk} \binom{N}{k} p^k (1-p)^{N-k} \\ &= \sum_{k=0}^{N} \binom{N}{k} (pe^s)^k (1-p)^{N-k} = (pe^s + (1-p))^N, \end{aligned}$$

where the last equality holds because $\sum_{k=0}^{N} \binom{N}{k} a^k b^{N-k} = (a+b)^N$. Therefore, the two moment-generating functions are identical.

Theorem 6.4 (Sum of binomial = binomial)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. binomial random variables with parameters (n, p). Let $Z = X_1 + \cdots + X_N$ be the sum. Then Z is a binomial random variable with parameters (Nn, p).

Proof

The MGF of a binomial random variable is

$$M_{X_i}(s) = (pe^s + (1-p))^n.$$

If we have N of these random variables, then $Z = X_1 + \cdots + X_N$ will have the MGF

$$M_Z(s) = \prod_{i=1}^{N} M_{X_i}(s) = (pe^s + (1-p))^{Nn}.$$

Note that this is just the MGF of another binomial random variable with parameter (Nn, p).

Theorem 6.5 (Sum of Poisson = Poisson)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. Poisson random variables with parameter λ. Let $Z = X_1 + \cdots + X_N$ be the sum. Then Z is a Poisson random variable with parameter $N\lambda$.

Proof

The MGF of a Poisson random variable is

$$\begin{aligned} M_X(s) = \mathbb{E}[e^{sX}] &= \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^s)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^s} = e^{\lambda(e^s - 1)}. \end{aligned}$$

Assume that we have a sum of N i.i.d. Poisson random variables. Then, by the main theorem, we have that

$$M_Z(s) = [M_X(s)]^N = e^{N\lambda(e^s-1)}.$$

Therefore, the resulting random variable Z is a Poisson with parameter $N\lambda$.

Theorem 6.6 (Sum of Gaussian = Gaussian)

Let $X_1, \ldots, X_N$ be a sequence of independent Gaussian random variables with parameters $(\mu_1, \sigma_1^2), \ldots, (\mu_N, \sigma_N^2)$. Let $Z = X_1 + \cdots + X_N$ be the sum. Then Z is a Gaussian random variable:

$$Z = \text{Gaussian}\left(\sum_{n=1}^{N} \mu_n, \sum_{n=1}^{N} \sigma_n^2\right). \tag{6.7}$$

Proof

We skip the proof of the MGF of a Gaussian. It can be shown that

$$M_X(s) = \exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}.$$

When we have a sequence of Gaussian random variables, then

$$\begin{aligned} M_Z(s) &= \mathbb{E}[e^{s(X_1+\cdots+X_N)}] \\ &= M_{X_1}(s) \cdots M_{X_N}(s) \\ &= \left(\exp\left\{\mu_1 s + \frac{\sigma_1^2 s^2}{2}\right\}\right) \cdots \left(\exp\left\{\mu_N s + \frac{\sigma_N^2 s^2}{2}\right\}\right) \\ &= \exp\left\{\left(\sum_{n=1}^{N} \mu_n\right) s + \left(\sum_{n=1}^{N} \sigma_n^2\right) \frac{s^2}{2}\right\}. \end{aligned}$$

Therefore, the resulting random variable Z is also a Gaussian. The mean and variance of Z are $\sum_{n=1}^{N} \mu_n$ and $\sum_{n=1}^{N} \sigma_n^2$, respectively.

6.1.3 Characteristic functions

Moment-generating functions are the Laplace transforms of the PDFs. However, the defining integral $\int e^{sx} f_X(x)\, dx$ does not converge for every PDF, so not every random variable has an MGF. One way to mitigate this problem is to restrict s to the imaginary axis, $s = j\omega$, which makes $|e^{sx}| = 1$ and hence guarantees convergence. This will give us the **characteristic function**.

Definition 6.2 (Usual definition)

The **characteristic function** of a random variable X is

$$\Phi_X(j\omega) = \mathbb{E}[e^{j\omega X}]. \tag{6.8}$$

However, we note that since ω can take any value in $(-\infty, \infty)$, it does not matter if we consider $\mathbb{E}[e^{-j\omega X}]$ or $\mathbb{E}[e^{j\omega X}]$. This leads to the following equivalent definition of the characteristic function:

Definition 6.3 (Alternative definition (for this book))

The **characteristic function** of a random variable X is

$$\Phi_X(j\omega) = \mathbb{E}[e^{-j\omega X}]. \tag{6.9}$$

If we follow this definition, we see that the characteristic function can be written as

$$\Phi_X(j\omega) = \mathbb{E}[e^{-j\omega X}] = \int_{-\infty}^{\infty} e^{-j\omega x} f_X(x)\, dx. \tag{6.10}$$

This is exactly the **Fourier transform** of the PDF. The reason for introducing this alternative characteristic function is that $\mathbb{E}[e^{-j\omega X}]$ is the Fourier transform of $f_X(x)$ but $\mathbb{E}[e^{j\omega X}]$ is the inverse Fourier transform of $f_X(x)$. The former is more convenient (in terms of notation) for students who have taken a course in signals and systems. However, we should stress that the usual way of defining the characteristic function is $\mathbb{E}[e^{j\omega X}]$.

A list of common Fourier transforms is shown in the table below. Additional identities can be found in standard signals and systems textbooks.

Example 6.2

Let X be a random variable with PDF $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Find the characteristic function.

Solution. The Fourier transform pair is

$$\lambda e^{-\lambda x} \longrightarrow \lambda \cdot \mathcal{F}\left\{e^{-\lambda x}\right\} = \lambda \cdot \frac{1}{\lambda + j\omega}.$$

Therefore, the characteristic function is $\Phi_X(j\omega) = \frac{\lambda}{\lambda + j\omega}$.

Example 6.3

Let X and Y be independent, and let

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$$

Find the PDF of $Z = X + Y$.

Solution. The characteristic function of X and Y can be found from the Fourier table:

$$\Phi_X(j\omega) = \frac{\lambda}{\lambda + j\omega} \qquad \text{and} \qquad \Phi_Y(j\omega) = \frac{\lambda}{\lambda + j\omega}.$$

Fourier Transforms $f(t) \longleftrightarrow F(\omega)$

1.	$e^{-at}u(t) \longleftrightarrow \frac{1}{a+j\omega}, a>0$	10.	$\text{sinc}^2(\frac{Wt}{2}) \longleftrightarrow \frac{2\pi}{W}\Delta(\frac{\omega}{2W})$
2.	$e^{at}u(-t) \longleftrightarrow \frac{1}{a-j\omega}, a>0$	11.	$e^{-at}\sin(\omega_0 t)u(t) \longleftrightarrow \frac{\omega_0}{(a+j\omega)^2+\omega_0^2}, a>0$
3.	$e^{-a\lvert t\rvert} \longleftrightarrow \frac{2a}{a^2+\omega^2}, a>0$	12.	$e^{-at}\cos(\omega_0 t)u(t) \longleftrightarrow \frac{a+j\omega}{(a+j\omega)^2+\omega_0^2}, a>0$
4.	$\frac{a^2}{a^2+t^2} \longleftrightarrow \pi a e^{-a\lvert\omega\rvert}, a>0$	13.	$e^{-\frac{t^2}{2\sigma^2}} \longleftrightarrow \sqrt{2\pi}\sigma e^{-\frac{\sigma^2\omega^2}{2}}$
5.	$te^{-at}u(t) \longleftrightarrow \frac{1}{(a+j\omega)^2}, a>0$	14.	$\delta(t) \longleftrightarrow 1$
6.	$t^n e^{-at}u(t) \longleftrightarrow \frac{n!}{(a+j\omega)^{n+1}}, a>0$	15.	$1 \longleftrightarrow 2\pi\delta(\omega)$
7.	$\text{rect}(\frac{t}{\tau}) \longleftrightarrow \tau\text{sinc}(\frac{\omega\tau}{2})$	16.	$\delta(t-t_0) \longleftrightarrow e^{-j\omega t_0}$
8.	$\text{sinc}(Wt) \longleftrightarrow \frac{\pi}{W}\text{rect}(\frac{\omega}{2W})$	17.	$e^{j\omega_0 t} \longleftrightarrow 2\pi\delta(\omega-\omega_0)$
9.	$\Delta(\frac{t}{\tau}) \longleftrightarrow \frac{\tau}{2}\text{sinc}^2(\frac{\omega\tau}{4})$	18.	$f(t)e^{j\omega_0 t} \longleftrightarrow F(\omega-\omega_0)$

$$\text{sinc}(t) = \frac{\sin(t)}{t} \qquad \text{rect}(t) = \begin{cases} 1, & -0.5 \le t \le 0.5, \\ 0, & \text{otherwise.} \end{cases} \qquad \Delta(t) = \begin{cases} 1-2|t|, & -0.5 \le t \le 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

Table 6.2: Fourier transform pairs of commonly used functions.

Therefore, the characteristic function of Z is

$$\Phi_Z(j\omega) = \Phi_X(j\omega)\Phi_Y(j\omega) = \frac{\lambda^2}{(\lambda+j\omega)^2}.$$

By inverse Fourier transform, we have that

$$f_Z(z) = \mathcal{F}^{-1}\left\{\frac{\lambda^2}{(\lambda+j\omega)^2}\right\} = \lambda^2 z e^{-\lambda z}, \quad z \ge 0.$$

Why $\Phi_X(j\omega)$ but not $M_X(s)$? As we said, the MGF is not always defined. Recall that the expectation $\mathbb{E}[X]$ exists only when X is absolutely integrable, i.e., $\mathbb{E}[|X|] < \infty$. For a characteristic function, the expectation is valid because $\mathbb{E}[|e^{j\omega X}|] = \mathbb{E}[1] = 1$. However, for a moment-generating function, $\mathbb{E}[|e^{sX}|]$ could be unbounded. To see a counterexample, we consider the Cauchy distribution.

Theorem 6.7

Consider the Cauchy distribution with PDF

$$f_X(x) = \frac{1}{\pi(x^2+1)}. \tag{6.11}$$

The MGF of X is undefined but the characteristic function is well defined.

Proof

Take any $s > 0$. The MGF is

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{\infty} e^{sx} \frac{1}{\pi(x^2+1)}\, dx \geq \int_1^{\infty} e^{sx} \frac{1}{\pi(x^2+1)}\, dx \\ &\geq \int_1^{\infty} \frac{(sx)^3}{6\pi(x^2+1)}\, dx, \quad \text{because} \quad e^{sx} \geq \frac{(sx)^3}{6} \\ &\geq \int_1^{\infty} \frac{(sx)^3}{6\pi(2x^2)}\, dx = \frac{s^3}{12\pi} \int_1^{\infty} x\, dx = \infty. \end{aligned}$$

By symmetry the same divergence occurs on the left tail for $s < 0$. Therefore, the MGF is infinite for every $s \neq 0$ — it exists only at the single point $s = 0$, where it trivially equals 1, so it is useless. On the other hand, by the Fourier table we know that

$$\Phi_X(j\omega) = \mathcal{F}\left\{ \frac{1}{\pi(x^2+1)} \right\} = e^{-|\omega|}.$$

Example 6.4

Let $X_0, X_1, \ldots$ be a sequence of independent random variables with PDF

$$f_{X_k}(x) = \frac{a_k}{\pi(a_k^2 + x^2)}, \quad a_k = \frac{1}{2^{k+1}} \quad \text{for } k = 0, 1, \ldots.$$

Find the PDF of Y, where $Y = \sum_{k=0}^{\infty} X_k$.
Solution. From the Fourier transform table, we know that

$$\frac{a_k}{\pi(a_k^2 + x^2)} = \frac{1}{a_k \pi} \cdot \frac{a_k^2}{(a_k^2 + x^2)} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{a_k \pi} \cdot \pi a_k e^{-a_k|\omega|} = e^{-a_k|\omega|}.$$

The characteristic function of Y is

$$\Phi_Y(j\omega) = \prod_{k=0}^{\infty} \Phi_{X_k}(j\omega) = \exp\left\{ -|\omega| \sum_{k=0}^{\infty} a_k \right\}.$$

Since $\sum_{k=0}^{\infty} a_k = \sum_{k=0}^{\infty} \frac{1}{2^{k+1}} = \frac{1}{2} + \frac{1}{4} + \cdots = 1$, the characteristic function becomes $\Phi_Y(j\omega) = e^{-|\omega|}$. The inverse Fourier transform gives us

$$e^{-|\omega|} = \frac{1}{\pi} \cdot \pi e^{-|\omega|} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{\pi} \cdot \frac{1}{1+x^2}.$$

Therefore the PDF of Y is

$$f_Y(y) = \frac{1}{\pi(1+y^2)}.$$

Example 6.5

Two random variables X and Y have the PDFs

$$f_X(x) = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} e^{-y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$$

Find the PDF of $Z = \max(X, Y) - \min(X, Y)$.
Solution. We first show that

$$Z = \max(X, Y) - \min(X, Y) = |X - Y|.$$

Suppose $X > Y$, then $\max(X, Y) = X$ and $\min(X, Y) = Y$. So $Z = X - Y$. If $X < Y$, then $\max(X, Y) = Y$ and $\min(X, Y) = X$. So $Z = Y - X$. Combining the two cases gives us $Z = |X - Y|$. Now, consider the Fourier transform of the PDFs:

$$e^{-x} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \frac{1}{1+j\omega}.$$

Let $U = X - Y$, and let $Z = |U|$. The characteristic function is

$$\begin{aligned} \Phi_U(j\omega) &= \mathbb{E}[e^{-j\omega(X-Y)}] = \mathbb{E}[e^{-j\omega X}]\mathbb{E}[e^{j\omega Y}] \\ &= \frac{1}{1+j\omega} \cdot \frac{1}{1-j\omega} = \frac{1}{1+\omega^2} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad f_U(u) = \frac{1}{2}e^{-|u|}. \end{aligned}$$

With the PDF of U, we can find the CDF of Z:

$$\begin{aligned} F_Z(z) = \mathbb{P}[Z \leq z] &= \mathbb{P}[|U| \leq z] \\ &= \int_{-z}^{z} f_U(u)\, du \\ &= \int_{-z}^{z} \frac{1}{2}e^{-|u|}\, du \\ &= 2\int_{0}^{z} \frac{1}{2}e^{-u}\, du = 1 - e^{-z}. \end{aligned}$$

Hence, the PDF is

$$f_Z(z) = \frac{d}{dz}F_Z(z) = e^{-z}.$$

Closing remark. Moment-generating functions and characteristic functions are useful tools. In this section, we have confined our discussion to using them to compute the sum of two random variables. Later sections and chapters will explain further uses for these functions. For example, we use the MGFs when proving Chernoff's bound and proving the Central

Limit Theorem.

6.2 Probability Inequalities

Moment-generating functions and characteristic functions are powerful tools for handling the sum of random variables. We now introduce another set of tools, known as the **probability inequalities**, that allow us to do approximations. We will highlight a few basic probability inequalities in this section.

6.2.1 Union bound

The first inequality is the union bound we had introduced when we discussed the axioms of probability. The union bound states the following:

Theorem 6.8 (Union Bound)

Let $A_1, \ldots, A_N$ be a collection of sets. Then

$$\mathbb{P}\left[\bigcup_{n=1}^{N} A_n\right] \leq \sum_{n=1}^{N} \mathbb{P}[A_n]. \tag{6.12}$$

Proof

We can prove this by induction. First, if $N = 2$,

$$\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2] - \mathbb{P}[A_1 \cap A_2] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2],$$

because $\mathbb{P}[A_1 \cap A_2]$ is a probability and so it must be non-negative. Thus we have proved the base case. Assume that the statement is true for $N = K$. We need to prove that the statement is also true for $N = K + 1$. To this end, we note that

$$\begin{aligned}
\mathbb{P}\left[\bigcup_{n=1}^{K+1} A_n\right] &= \mathbb{P}\left[\left(\bigcup_{n=1}^{K} A_n\right) \cup A_{K+1}\right] \\
&= \mathbb{P}\left[\bigcup_{n=1}^{K} A_n\right] + \mathbb{P}[A_{K+1}] - \mathbb{P}\left[\left(\bigcup_{n=1}^{K} A_n\right) \cap A_{K+1}\right] \\
&\leq \mathbb{P}\left[\bigcup_{n=1}^{K} A_n\right] + \mathbb{P}[A_{K+1}].
\end{aligned}$$

Then, according to our hypothesis for $N = K$, it follows that

$$\mathbb{P}\left[\bigcup_{n=1}^{K} A_n\right] \leq \sum_{n=1}^{K} \mathbb{P}[A_n].$$

Putting these together,

$$\mathbb{P}\left[\bigcup_{n=1}^{K+1} A_n\right] \leq \sum_{n=1}^{K} \mathbb{P}[A_n] + \mathbb{P}[A_{K+1}] = \sum_{n=1}^{K+1} \mathbb{P}[A_n].$$

Therefore, by the principle of induction, we have proved the statement.

Remark. The tightness of the union bound depends on the amount of **overlapping** between the events $A_1, \ldots, A_N$, as illustrated in **Figure 6.4**. If the events are disjoint, the union bound is tight. If the events are overlapping significantly, the union is loose. The idea of the union bound is the principle of divide and conquer. We decompose the system into smaller events for a system of n variables and use the union bound to upper-limit the overall probability. If the probability of each event is small, the union bound tells us that the overall probability of the system will also be small.

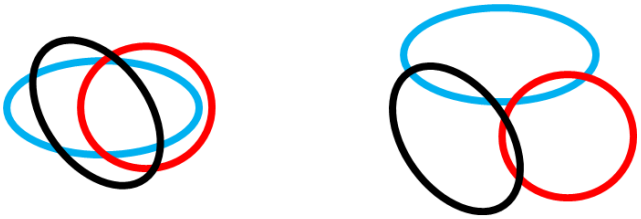


Figure 6.4: Conditions under which the union bound is loose or tight. [Left] The union bound is loose when the sets are overlapping. [Right] The union bound is tight when the sets are (nearly) disjoint.

Example 6.6

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with CDF $F_{X_n}(x)$ and let $Z = \min(X_1, \ldots, X_N)$. Find an upper bound on the CDF of Z.
Solution. Note that $Z = \min(X_1, \ldots, X_N) \leq z$ is equivalent to at least one of the X_n's being less than z. Thus, we have that

$$Z = \min(X_1, \ldots, X_N) \leq z \quad \Leftrightarrow \quad X_1 \leq z \cup \cdots \cup X_N \leq z.$$

Substituting this result into the CDF,

$$\begin{aligned} F_Z(z) &= \mathbb{P}[Z \leq z] \\ &= \mathbb{P}[\min(X_1, \ldots, X_N) \leq z] \\ &= \mathbb{P}[X_1 \leq z \cup \cdots \cup X_N \leq z] \\ &\leq \mathbb{P}[X_1 \leq z] + \cdots + \mathbb{P}[X_N \leq z] \\ &= N \cdot F_X(z). \end{aligned}$$

6.2.2 The Cauchy-Schwarz inequality

The second inequality we study here is the **Cauchy-Schwarz inequality**, which we previously mentioned in Chapter 5. We review it for the sake of completeness.

Theorem 6.9 (Cauchy-Schwarz inequality)

Let X and Y be two random variables. Then

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \tag{6.13}$$

Proof

Let $f(s) = \mathbb{E}[(sX + Y)^2]$ for any real s. Then

$$\begin{aligned} f(s) &= \mathbb{E}[(sX + Y)^2] \\ &= \mathbb{E}[s^2X^2 + 2sXY + Y^2] \\ &= \mathbb{E}[X^2]s^2 + 2\mathbb{E}[XY]s + \mathbb{E}[Y^2]. \end{aligned}$$

This is a quadratic equation, and $f(s) \geq 0$ for all s because $\mathbb{E}[(sX + Y)^2] \geq 0$. Recall that for a quadratic equation $\phi(x) = ax^2 + bx + c$, if the function $\phi(x) \geq 0$ then $b^2 - 4ac \leq 0$. Substituting this result into our problem, we show that

$$(2\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0.$$

This implies that

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2],$$

which completes the proof.

Remark. As shown in Chapter 5, the Cauchy-Schwarz inequality is useful in analyzing $\mathbb{E}[XY]$. For example, we can use the Cauchy-Schwarz inequality to prove that the correlation coefficient ρ is bounded between -1 and 1.

6.2.3 Jensen's inequality

Our next inequality is **Jensen's inequality**. To motivate the inequality, we recall that

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Since $\text{Var}[X] \geq 0$ for any X, it follows that

$$\underbrace{\mathbb{E}[X^2]}_{=\mathbb{E}[g(X)]} \quad \geq \quad \underbrace{\mathbb{E}[X]^2}_{=g(\mathbb{E}[X])}. \tag{6.14}$$

Jensen's inequality is a generalization of the above result by recognizing that the inequality does not only hold for the function $g(X) = X^2$ but also for any **convex** function g. The theorem is stated as follows:

Theorem 6.10 (Jensen's inequality)

Let X be a random variable, and let $g : \mathbb{R} \to \mathbb{R}$ be a **convex** function. Then

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]). \tag{6.15}$$

If the function g is **concave**, then the inequality sign is flipped: $\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$. The way to remember this result is to remember that $\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \text{Var}[X] \geq 0$.

Now, what is a convex function? Informally, a function g is **convex** if, when we pick any two points on the function and connect them with a straight line, the line will be above the function for that segment. This definition is illustrated in **Figure 6.5**. Consider an interval $[x, y]$, and the line segment connecting $g(x)$ and $g(y)$. If the function $g(\cdot)$ is convex, then the entire line segment should be above the curve.

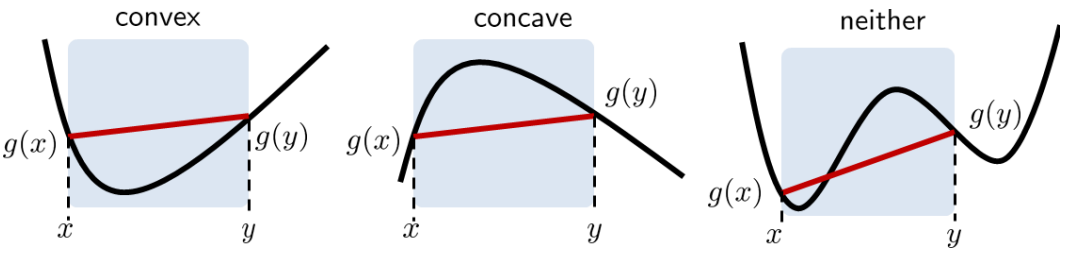


Figure 6.5: Illustration of a convex function, a concave function, and a function that is neither convex nor concave.

The definition of a convex function essentially follows the above picture:

Definition 6.4

A function g is **convex** if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y), \tag{6.16}$$

for any $0 \leq \lambda \leq 1$.

Here λ represents a "sweeping" constant that goes from 0 to 1. When $\lambda = 1$ then $\lambda x + (1-\lambda)y$ simplifies to x, and when $\lambda = 0$ then $\lambda x + (1-\lambda)y$ simplifies to y.

The definition is easy to understand. The left-hand side $g(\lambda x + (1-\lambda)y)$ is the function evaluated at any points in the interval $[x, y]$. The right-hand side is the red straight line we plotted in **Figure 6.5**. It connects the two points $g(x)$ and $g(y)$. Convexity means that the red line is entirely above the curve.

For twice-differentiable 1D functions, convexity can be described by the curvature of the function. A function is convex if

$$g''(x) \geq 0. \tag{6.17}$$

This is self-explanatory because if the curvature is non-negative for all x, then the slope of g has to keep increasing.

Example 6.7

The following functions are convex or concave:

- $g(x) = \log x$ is concave, because $g'(x) = \frac{1}{x}$ and $g''(x) = -\frac{1}{x^2} \leq 0$ for all x.
- $g(x) = x^2$ is convex, because $g'(x) = 2x$ and $g''(x) = 2$ is positive.
- $g(x) = e^{-x}$ is convex, because $g'(x) = -e^{-x}$ and $g''(x) = e^{-x} \geq 0$.

Why is Jensen's inequality valid for a convex function? Consider the illustration in **Figure 6.6**. Suppose we have a random variable X taking some PDF $f_X(x)$. There is a convex function $g(\cdot)$ that maps the random variable X to $g(X)$. Since $g(\cdot)$ is convex, a PDF like the one we see in **Figure 6.6** will become skewed. (You can map the left tail to the new left tail, the peak to the new peak, and the right tail to the new right tail.) As you can see from the figure, the new random variable $g(X)$ has a mean $\mathbb{E}[g(X)]$ that is greater than the mapped old mean $g(\mathbb{E}[X])$. Jensen's inequality captures this phenomenon by stating that $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ for any convex function $g(\cdot)$.

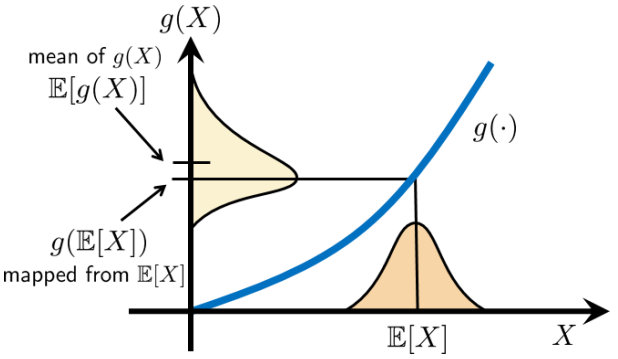


Figure 6.6: Jensen's inequality states that if there is a convex function $g(\cdot)$ that maps a random variable X to a new random variable $g(X)$, the new mean $\mathbb{E}[g(X)]$ will be greater than the mapped old mean $g(\mathbb{E}[X])$.

Proving Jensen's inequality is straightforward for a two-state discrete random variable. Define a random variable X with states x and y. The probabilities for these two states are $\mathbb{P}[X = x] = \lambda$ and $\mathbb{P}[X = y] = 1 - \lambda$. Then

$$\mathbb{E}[X] = \sum_{x' \in \{x,y\}} x' p_X(x') = \lambda x + (1 - \lambda)y.$$

Now, let $g(\cdot)$ be a convex function. We know from the expectation that

$$\mathbb{E}[g(X)] = \sum_{x' \in \{x,y\}} g(x') p_X(x') = g(x)\lambda + (1 - \lambda)g(y).$$

By convexity of the function $g(\cdot)$, it follows that

$$\underbrace{g(\lambda x + (1 - \lambda)y)}_{= g(\mathbb{E}[X])} \leq \underbrace{\lambda g(x) + (1 - \lambda)g(y)}_{= \mathbb{E}[g(X)]},$$

where in the underbrace we substitute the definitions using the expectation. Therefore, for any two-state discrete random variables, the proof of Jensen's inequality follows directly from the convexity. If the discrete random variable takes more than two states, we can prove the theorem by induction. For continuous random variables, we can prove the theorem using the following approach.

You may skip the proof of Jensen's inequality if this is your first time reading the book.

Here we present an alternative proof of Jensen's inequality that does not require proof by induction. The idea is to recognize that if the function g is convex we can find a tangent line $L(X) = aX + b$ at the point $\mathbb{E}[X]$ that is uniformly lower than $g(X)$, i.e., $g(X) \geq L(X)$ for all X. Then we can prove the result with a simple geometric argument. **Figure 6.7** illustrates this idea.

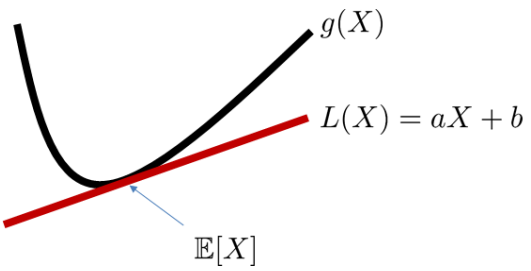


Figure 6.7: Geometric illustration of the proof of Jensen's inequality. Suppose $g(\cdot)$ is a convex function. For any point X on $g(\cdot)$, we can find a tangent line $L(X) = aX + b$. Since the black curve is always above the tangent, it follows that $\mathbb{E}[g(X)] \geq \mathbb{E}[L(X)]$ for any X. Also, note that at a particular point $\mathbb{E}[X]$, the black curve and the red line touch, and so we have $L(\mathbb{E}[X]) = g(\mathbb{E}[X])$.

Proof of Jensen's inequality

Consider $L(X)$ as defined above. Since g is convex, $g(X) \geq L(X)$ for all X. Therefore,

$$\begin{aligned}\mathbb{E}[g(X)] &\geq \mathbb{E}[L(X)] \\ &= \mathbb{E}[aX + b] \\ &= a\mathbb{E}[X] + b \\ &= L(\mathbb{E}[X]) = g(\mathbb{E}[X]),\end{aligned}$$

where the last equality holds because L is a tangent line to g where they meet at $\mathbb{E}[X]$.

What are (a, b) in the proof? By Taylor expansion,

$$\begin{aligned}g(X) &\approx g(\mathbb{E}[X]) + g'(\mathbb{E}[X])(X - \mathbb{E}[X]) \\ &\stackrel{\text{def}}{=} L(X).\end{aligned}$$

Therefore, if we want to be precise, then $a = g'(\mathbb{E}[X])$ and $b = g(\mathbb{E}[X]) - g'(\mathbb{E}[X])\mathbb{E}[X]$.

The end of the proof.

Example 6.8

By Jensen's inequality, we have that

(a) $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$, because $g(x) = x^2$ is convex.

(b) If $X > 0$, then $\mathbb{E}\left[\frac{1}{X}\right] \geq \frac{1}{\mathbb{E}[X]}$, because $g(x) = \frac{1}{x}$ is convex on $x > 0$.

(c) $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$, because $g(x) = \log x$ is concave.

6.2.4 Markov's inequality

Our next inequality, **Markov's inequality**, is an elementary inequality that links probability and expectation.

Theorem 6.11 (Markov's inequality)

Let $X \geq 0$ be a non-negative random variable. Then, for any $\varepsilon > 0$, we have

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}. \tag{6.18}$$

Markov's inequality concerns the **tail** of the random variable. As illustrated in **Figure 6.8**, $\mathbb{P}[X \geq \varepsilon]$ measures the probability that the random variable takes a value greater than ε. Markov's inequality asserts that this probability $\mathbb{P}[X \geq \varepsilon]$ is upper-bounded by the ratio $\mathbb{E}[X]/\varepsilon$. This result is useful because it relates the probability and the expectation. In many problems the probability $\mathbb{P}[X \geq \varepsilon]$ could be difficult to evaluate if the PDF is complicated. The expectation, on the other hand, is usually easier to evaluate.

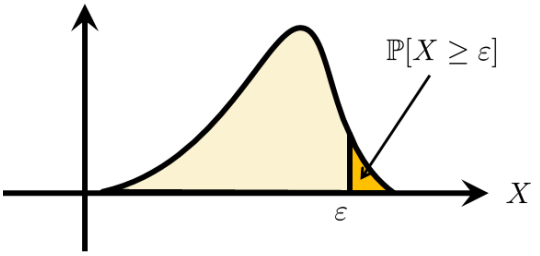


Figure 6.8: Markov's inequality provides an upper bound to the tail of a random variable. The inequality states that the probability $\mathbb{P}[X \geq \varepsilon]$ is upper bounded by the ratio $\mathbb{E}[X]/\varepsilon$.

Proof

Consider $\varepsilon\mathbb{P}[X \geq \varepsilon]$. It follows that

$$\varepsilon\mathbb{P}[X \geq \varepsilon] = \int_{\varepsilon}^{\infty} \varepsilon\, f_X(x)\, dx \leq \int_{\varepsilon}^{\infty} x f_X(x)\, dx,$$

where the inequality is valid because on the region of integration we have $x \geq \varepsilon$, and $f_X(x) \geq 0$, so replacing the constant ε by x can only increase the integrand. It then follows that

$$\int_{\varepsilon}^{\infty} x f_X(x)\, dx \leq \int_{0}^{\infty} x f_X(x)\, dx = \mathbb{E}[X].$$

A pictorial interpretation of Markov's inequality is shown in **Figure 6.9**. For $X > 0$, it is not difficult to show that $\mathbb{E}[X] = \int_0^\infty 1 - F_X(x)\, dx$. Then, in the CDF plot, we see that $\varepsilon \cdot \mathbb{P}[X \geq \varepsilon]$ is a rectangle covering the top left corner. This area is clearly smaller than the area covered by the function $1 - F_X(x)$.

Practice Exercise 6.4

Prove that if $X > 0$, then $\mathbb{E}[X] = \int_0^\infty 1 - F_X(x)\, dx$.

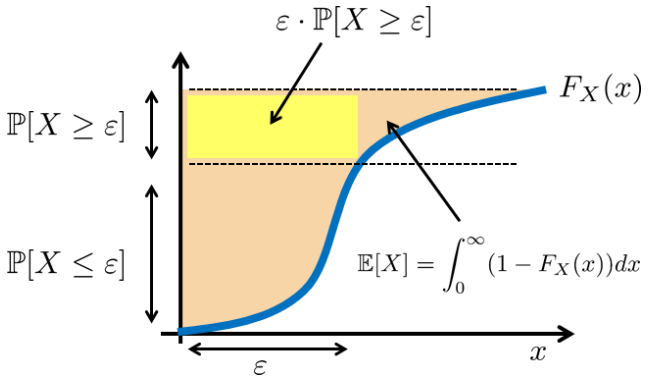
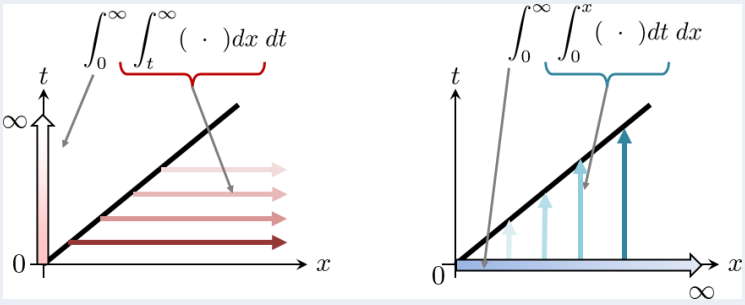


Figure 6.9: The proof of Markov's inequality follows from the fact that $\varepsilon \cdot \mathbb{P}[X \geq \varepsilon]$ occupies the top left corner marked by the yellow rectangle. The expectation is the area above the CDF so that $\mathbb{E}[X] = \int_0^\infty 1 - F_X(x)\ dx$. Since the yellow rectangle is smaller than the orange shaded area, it follows that $\varepsilon \cdot \mathbb{P}[X \geq \varepsilon] \leq \mathbb{E}[X]$, which is Markov's inequality.

Solution. We start from the right-hand side:

$$\begin{aligned}
\int_0^\infty 1 - F_X(x)\ dx &= \int_0^\infty 1 - \mathbb{P}[X \leq x]\ dx \\
&= \int_0^\infty \mathbb{P}[X \geq x]\ dx \\
&= \int_0^\infty \int_x^\infty f_X(t)\ dt\ dx \\
&= \int_0^\infty \int_0^t f_X(t)\ dx\ dt \\
&= \int_0^\infty t f_X(t)\ dt = \mathbb{E}[X].
\end{aligned}$$

The change in the integration order is illustrated below.



How tight is Markov's inequality? It is possible to create a random variable such that the equality is met (see Exercise 6.14). However, in general, the estimate provided by the upper bound is not tight. Here is an example.

Practice Exercise 6.5

Let $X \sim \text{Uniform}(0, 4)$. Verify Markov's inequality for $\mathbb{P}[X \geq 2]$, $\mathbb{P}[X \geq 3]$ and $\mathbb{P}[X \geq 4]$.

Solution. First, we observe that $\mathbb{E}[X] = 2$. Then

$$\begin{aligned} \mathbb{P}[X \geq 2] &= 0.5, & \frac{\mathbb{E}[X]}{2} &= 1, \\ \mathbb{P}[X \geq 3] &= 0.25, & \frac{\mathbb{E}[X]}{3} &= 0.67, \\ \mathbb{P}[X \geq 4] &= 0, & \frac{\mathbb{E}[X]}{4} &= 0.5. \end{aligned}$$

Therefore, although the upper bounds are all valid, they are very loose.

If Markov's inequality is not tight, why is it useful? It turns out that while Markov's inequality is not tight, its variations can be powerful. We will come back to this point when we discuss Chernoff's bound.

6.2.5 Chebyshev's inequality

The next inequality is a simple extension of Markov's inequality. The result is known as Chebyshev's inequality.

Theorem 6.12 (Chebyshev's inequality)

Let X be a random variable with mean μ. Then for any $\varepsilon > 0$ we have

$$\mathbb{P}[|X - \mu| \geq \varepsilon] \leq \frac{\text{Var}[X]}{\varepsilon^2}. \tag{6.19}$$

The tail measured by Chebyshev's inequality is illustrated in **Figure 6.10**. Since the event $|X - \mu| \geq \varepsilon$ involves an absolute value, the probability measures the two-sided tail. Chebyshev's inequality states that this tail probability is upper-bounded by $\text{Var}[X]/\varepsilon^2$.

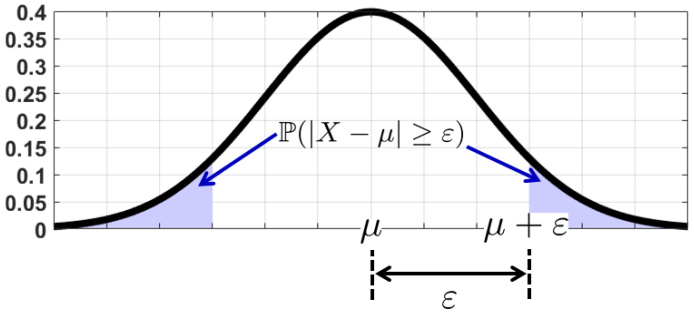


Figure 6.10: Chebyshev's inequality states that the two-sided tail probability $\mathbb{P}[|X - \mu| \geq \varepsilon]$ is upper-bounded by $\text{Var}[X]/\varepsilon^2$

Proof

We apply Markov's inequality to show that

$$\begin{aligned}\mathbb{P}[|X-\mu| \geq \varepsilon] &= \mathbb{P}[(X-\mu)^2 \geq \varepsilon^2] \\ &\leq \frac{\mathbb{E}[(X-\mu)^2]}{\varepsilon^2} = \frac{\text{Var}[X]}{\varepsilon^2}.\end{aligned}$$

An alternative form of Chebyshev's inequality is obtained by letting $\varepsilon = k\sigma$. In this case, we have

$$\mathbb{P}[|X-\mu| \geq k\sigma] \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}.$$

Therefore, if a random variable is k times the standard deviation away from the mean, then the probability bound drops to $1/k^2$.

Practice Exercise 6.6

Let $X \sim \text{Uniform}(0, 2\sqrt{2})$. Find the bound of Chebyshev's inequality for the probability $\mathbb{P}[|X-\mu| \geq 1]$.

Solution. Note that $\mathbb{E}[X] = \sqrt{2}$ and $\sigma^2 = (2\sqrt{2})^2/12 = 2/3$. Therefore, we have

$$\mathbb{P}[|X-\mu| \geq 1] \leq \frac{\sigma^2}{\varepsilon^2} = \frac{2}{3},$$

which is a valid upper bound, but not very useful.

Practice Exercise 6.7

Let $X \sim \text{Exponential}(1)$. Find the bound of Chebyshev's inequality for the probability $\mathbb{P}[X \geq \varepsilon]$.

Solution. Note that $\mathbb{E}[X] = 1$ and $\sigma^2 = 1$. Assume $\varepsilon > 1$ (otherwise the bound below is vacuous, since a probability bound greater than 1 says nothing). Thus we have

$$\begin{aligned}\mathbb{P}[X \geq \varepsilon] = \mathbb{P}[X-\mu \geq \varepsilon-\mu] &\leq \mathbb{P}[|X-\mu| \geq \varepsilon-\mu] \\ &\leq \frac{\sigma^2}{(\varepsilon-\mu)^2} = \frac{1}{(\varepsilon-1)^2}.\end{aligned}$$

We can compare this with the exact probability, which is

$$\mathbb{P}[X \geq \varepsilon] = 1 - F_X(\varepsilon) = e^{-\varepsilon}.$$

Again, the estimate given by Chebyshev's inequality is acceptable but too conservative.

Corollary 6.2

Let $X_1, \ldots, X_N$ be i.i.d. random variables with mean $\mathbb{E}[X_n] = \mu$ and variance

$\mathrm{Var}[X_n] = \sigma^2$. Let $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$ be the sample mean. Then

$$\mathbb{P}\left[\left|\overline{X}_N - \mu\right| > \epsilon\right] \leq \frac{\sigma^2}{N\epsilon^2}. \tag{6.20}$$

Proof

We can first show that $\mathbb{E}[\overline{X}_N] = \mu$ and $\mathrm{Var}[\overline{X}_N]$ satisfies

$$\mathrm{Var}[\overline{X}_N] = \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}[X_n] = \frac{\sigma^2}{N}.$$

Then by Chebyshev's inequality,

$$\mathbb{P}\left[\left|\overline{X}_N - \mu\right| > \epsilon\right] \leq \frac{\mathrm{Var}[\overline{X}_N]}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2}.$$

The consequence of this corollary is that the upper bound $\sigma^2/(N\epsilon^2)$ will converge to zero as $N \to \infty$. Therefore, the probability of getting the event $\{|\overline{X}_N - \mu| > \epsilon\}$ is vanishing. It means that the sample average $\overline{X}_N$ is converging to the true population mean μ, in the sense that the probability of failing is shrinking.

6.2.6 Chernoff's bound

We now introduce a powerful inequality or a set of general procedures that gives us some highly useful inequalities. The idea is named for Herman Chernoff, although it was actually due to his colleague Herman Rubin.

Theorem 6.13 (Chernoff's bound)

Let X be a random variable. Then, for any $\varepsilon \geq 0$, we have that

$$\mathbb{P}[X \geq \varepsilon] \leq e^{-\varphi(\varepsilon)}, \tag{6.21}$$

where[a]

$$\varphi(\varepsilon) = \max_{s>0}\left\{s\varepsilon - \log M_X(s)\right\}, \tag{6.22}$$

and $M_X(s)$ is the moment-generating function.

[a] $\varphi(\varepsilon)$ is called the Fenchel-Legendre dual function of $\log M_X$. See references [6-14].

Proof

There are two tricks in the proof of Chernoff's bound. The first trick is a nonlinear transformation. Since e^{sx} is an increasing function for any $s > 0$ and x, we have that

$$\begin{aligned}\mathbb{P}[X \geq \varepsilon] &= \mathbb{P}[e^{sX} \geq e^{s\varepsilon}] \\ &\overset{(a)}{\leq} \frac{\mathbb{E}[e^{sX}]}{e^{s\varepsilon}} \\ &\overset{(b)}{=} e^{-s\varepsilon} M_X(s) \\ &= e^{-s\varepsilon + \log M_X(s)},\end{aligned}$$

where the inequality (a) is due to Markov's inequality. Step (b) just uses the definition of MGF that $\mathbb{E}[e^{sX}] = M_X(s)$.
Now for the second trick. Note that the above result holds for all s. That means it must also hold for the s that minimizes $e^{-s\varepsilon + \log M_X(s)}$. This implies that

$$\mathbb{P}[X \geq \varepsilon] \leq \min_{s>0} \left\{ e^{-s\varepsilon + \log M_X(s)} \right\}.$$

Again, since e^x is increasing, the minimizer of the above probability is also the maximizer of this function:

$$\varphi(\varepsilon) = \max_{s>0} \left\{ s\varepsilon - \log M_X(s) \right\}.$$

Thus, we conclude that $\mathbb{P}[X \geq \varepsilon] \leq e^{-\varphi(\varepsilon)}$.

6.2.7 Comparing Chernoff and Chebyshev

Let's consider an example of how Chernoff's bound can be useful.

Suppose that we have a random variable $X \sim \text{Gaussian}(0, \sigma^2/N)$. The number N can be regarded as the number of samples. For example, if $Y_1, \ldots, Y_N$ are N Gaussian random variables with mean 0 and variance σ^2, then the average $X = \frac{1}{N}\sum_{n=1}^{N} Y_n$ will have mean 0 and variance σ^2/N. Therefore, as N grows, the variance of X will become smaller and smaller.

First, since the random variable is Gaussian, we can show the following:

Lemma 6.1

Let $X \sim \text{Gaussian}(0, \frac{\sigma^2}{N})$ be a Gaussian random variable. Then, for any $\varepsilon > 0$,

$$\mathbb{P}[X \geq \varepsilon] = 1 - \Phi\left(\frac{\sqrt{N}\varepsilon}{\sigma}\right), \tag{6.23}$$

where Φ is the standard Gaussian's CDF.

Note that this is the **exact** result: If you tell me ε, N, and σ, then the probability $\mathbb{P}[X \geq \varepsilon]$ is exactly the one shown on the right-hand side. No approximation, no randomness.

Proof

Since X is Gaussian, the probability is

$$\begin{aligned}\mathbb{P}[X \geq \varepsilon] &= \int_{\varepsilon}^{\infty} \frac{1}{\sqrt{2\pi(\sigma^2/N)}} \exp\left\{-\frac{x^2}{2(\sigma^2/N)}\right\} dx \\ &= 1 - \int_{-\infty}^{\varepsilon} \frac{1}{\sqrt{2\pi(\sigma^2/N)}} \exp\left\{-\frac{x^2}{2(\sigma^2/N)}\right\} dx \\ &= 1 - \int_{-\infty}^{\frac{\varepsilon}{\sqrt{\sigma^2/N}}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = 1 - \Phi\left(\frac{\varepsilon}{\sqrt{\sigma^2/N}}\right) = 1 - \Phi\left(\frac{\sqrt{N}\varepsilon}{\sigma}\right).\end{aligned}$$

Let us compute the bound given by Chebyshev's inequality.

Lemma 6.2

Let $X \sim \text{Gaussian}(0, \frac{\sigma^2}{N})$ be a Gaussian random variable. Then, for any $\varepsilon > 0$, Chebyshev's inequality implies that

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\sigma^2}{N\varepsilon^2}. \tag{6.24}$$

Proof

We apply Chebyshev's inequality by assuming that $\mu = 0$:

$$\begin{aligned}\mathbb{P}[X \geq \varepsilon] = \mathbb{P}[X - \mu \geq \varepsilon - \mu] &\leq \mathbb{P}[|X - \mu| \geq \varepsilon - \mu] \\ &\leq \frac{\mathbb{E}[(X-\mu)^2]}{(\varepsilon - \mu)^2} = \frac{\sigma^2}{N\varepsilon^2}.\end{aligned}$$

We now compute Chernoff's bound.

Theorem 6.14

Let $X \sim \text{Gaussian}(0, \frac{\sigma^2}{N})$ be a Gaussian random variable. Then, for any $\varepsilon > 0$, Chernoff's bound implies that

$$\mathbb{P}\left[X \geq \varepsilon\right] \leq \exp\left\{-\frac{\varepsilon^2 N}{2\sigma^2}\right\}. \tag{6.25}$$

Proof

The MGF of a zero-mean Gaussian random variable with variance σ^2/N is $M_X(s) = \exp\left\{\frac{\sigma^2 s^2}{2N}\right\}$. Therefore, the function φ can be written as

$$\begin{aligned}\varphi(\varepsilon) &= \max_{s>0} \left\{s\varepsilon - \log M_X(s)\right\} \\ &= \max_{s>0} \left\{s\varepsilon - \frac{\sigma^2 s^2}{2N}\right\}.\end{aligned}$$

To maximize the function we take the derivative and set it to zero. This yields

$$\frac{d}{ds}\left\{s\varepsilon - \frac{\sigma^2 s^2}{2N}\right\} = 0 \quad \Rightarrow \quad s^* = \frac{N\varepsilon}{\sigma^2}.$$

Note that this s^* is a maximizer because $s\varepsilon - \frac{\sigma^2 s^2}{2N}$ is a concave function. Substituting s^* into $\varphi(\varepsilon)$,

$$\begin{aligned}\varphi(\varepsilon) &= \max_{s>0}\left\{s\varepsilon - \frac{\sigma^2 s^2}{2N}\right\} \\ &= s^*\varepsilon - \frac{\sigma^2 (s^*)^2}{2N} = \left(\frac{N\varepsilon}{\sigma^2}\right)\varepsilon - \frac{\sigma^2}{2N}\left(\frac{N\varepsilon}{\sigma^2}\right)^2 = \frac{\varepsilon^2 N}{2\sigma^2},\end{aligned}$$

and hence

$$\mathbb{P}[X \geq \varepsilon] \leq e^{-\varphi(\varepsilon)} = \exp\left\{-\frac{\varepsilon^2 N}{2\sigma^2}\right\}.$$

Figure 6.11 shows the comparison between the exact probability, the bound provided by Chebyshev's inequality, and Chernoff's bound:

- **Exact**: $\mathbb{P}[X \geq \varepsilon] = 1 - \Phi\left(\frac{\sqrt{N}\varepsilon}{\sigma}\right)$.
- **Chebyshev**: $\mathbb{P}[X \geq \varepsilon] \leq \frac{\sigma^2}{N\varepsilon^2}$,
- **Chernoff**: $\mathbb{P}\left[X \geq \varepsilon\right] \leq \exp\left\{-\frac{\varepsilon^2 N}{2\sigma^2}\right\}$.

In this numerical experiment, we set $\varepsilon = 0.1$, and $\sigma = 1$. We vary the number N. As we can see from the figure, the bound provided by Chebyshev is valid but very loose. It does not even capture the tail as N grows. On the other hand, Chernoff's bound is reasonably tight. However, one should note that the tightness of Chernoff is only valid for **large** N. When N is small, it is possible to construct random variables such that Chebyshev is tighter.

The MATLAB code used to generate this plot is illustrated below.

```
% MATLAB code to compare the probability bounds
epsilon = 0.1;
sigma   = 1;
N       = logspace(1,3.9,50);
p_exact = 1-normcdf(sqrt(N)*epsilon/sigma);
p_cheby = sigma^2./(epsilon^2*N);
p_chern = exp(-epsilon^2*N/(2*sigma^2));

loglog(N, p_exact, '-o', 'Color', [1 0.5 0], 'LineWidth', 2); hold on;
loglog(N, p_cheby, '-', 'Color', [0.2 0.7 0.1], 'LineWidth', 2);
loglog(N, p_chern, '-', 'Color', [0.2 0.0 0.8], 'LineWidth', 2);
```

What could go wrong if we insist on using Chebyshev's inequality? Consider the following example.

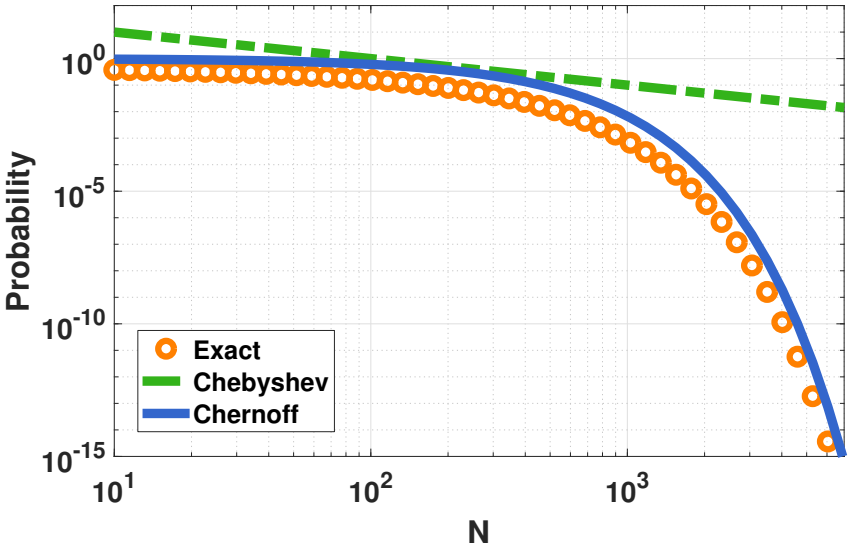


Figure 6.11: Comparison between Chernoff's bound and Chebyshev's bound. The random variable we use is $X \sim \text{Gaussian}(0, \sigma^2/N)$. As N grows, we show the probability bounds predicted by the two methods.

Example 6.9

Let $X \sim \text{Gaussian}(0, \sigma^2/N)$. Suppose that we want the probability to be no greater than a confidence level of α:

$$\mathbb{P}[X \geq \varepsilon] \leq \alpha.$$

Let $\alpha = 0.05$, $\varepsilon = 0.1$, and $\sigma = 1$. Find the N using (i) Chebyshev's inequality and (ii) Chernoff's inequality.

Solution: (i) Chebyshev's inequality implies that

$$\mathbb{P}[X \geq \varepsilon] \leq \frac{\sigma^2}{N\varepsilon^2} \leq \alpha,$$

which means that

$$N \geq \frac{\sigma^2}{\alpha\varepsilon^2}.$$

If we plug in $\alpha = 0.05$, $\varepsilon = 0.1$, and $\sigma = 1$, then $N \geq 2000$.

(ii) For Chernoff's inequality, it holds that

$$\mathbb{P}[X \geq \varepsilon] \leq \exp\left\{-\frac{\varepsilon^2 N}{2\sigma^2}\right\} \leq \alpha,$$

which means that

$$N \geq -\frac{2\sigma^2}{\varepsilon^2}\log\alpha$$

Plugging in $\alpha = 0.05$, $\varepsilon = 0.1$, and $\sigma = 1$, we have that $N \geq 600$. This is more than 3 times smaller than the one predicted by Chebyshev's inequality. Which one is

correct? Both are correct but Chebyshev's inequality is overly conservative. If $N \geq 600$ can make $\mathbb{P}[X \geq \varepsilon] \leq \alpha$, then certainly $N \geq 2000$ will work too. However, $N \geq 2000$ is too loose.

6.2.8 Hoeffding's inequality

Chernoff's bound can be used to derive many powerful inequalities. Here we present an inequality for bounded random variables. This result is known as Hoeffding's inequality.

Theorem 6.15 (Hoeffding's inequality)

Let $X_1, \ldots, X_N$ be i.i.d. random variables with $0 \leq X_n \leq 1$, and $\mathbb{E}[X_n] = \mu$. Then

$$\mathbb{P}\left[\left|\overline{X}_N - \mu\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}, \tag{6.26}$$

where $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$.

You may skip the proof of Hoeffding's inequality if this is your first time reading the book.

Proof

(Hoeffding's inequality) First, we show that

$$\begin{aligned}
\mathbb{P}\left[\overline{X}_N - \mu > \epsilon\right] &= \mathbb{P}\left[\frac{1}{N}\sum_{n=1}^{N} X_n - \mu > \epsilon\right] = \mathbb{P}\left[\sum_{n=1}^{N}(X_n - \mu) > N\epsilon\right] \\
&= \mathbb{P}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)} \geq e^{s\epsilon N}\right] \\
&\leq \frac{\mathbb{E}\left[e^{s\sum_{n=1}^{N}(X_n-\mu)}\right]}{e^{s\epsilon N}} = \left(\frac{\mathbb{E}\left[e^{s(X_n-\mu)}\right]}{e^{s\epsilon}}\right)^N.
\end{aligned}$$

Let $Z_n = X_n - \mu$. Then $-\mu \leq Z_n \leq 1 - \mu$. At this point we use Hoeffding Lemma (see below) that $\mathbb{E}[e^{sZ_n}] \leq e^{\frac{s^2}{8}}$ because $b - a = (1-\mu) - (-\mu) = 1$. Thus,

$$\mathbb{P}\left[\overline{X}_N - \mu > \epsilon\right] \leq \left(\frac{\mathbb{E}\left[e^{sZ_n}\right]}{e^{s\epsilon}}\right)^N \leq \left(\frac{e^{\frac{s^2}{8}}}{e^{s\epsilon}}\right)^N = e^{\frac{s^2 N}{8} - s\epsilon N}, \qquad \forall s.$$

This result holds for all s, and thus it holds for the s that minimizes the right-hand side. This implies that

$$\mathbb{P}\left[\overline{X}_N - \mu > \epsilon\right] \leq \min_s \left\{\exp\left\{\frac{s^2 N}{8} - s\epsilon N\right\}\right\}.$$

Minimizing the exponent gives $\frac{d}{ds}\left\{\frac{s^2N}{8} - s\epsilon N\right\} = \frac{sN}{4} - \epsilon N = 0$. Thus we have $s = 4\epsilon$. Hence,

$$\mathbb{P}\left[\overline{X}_N - \mu > \epsilon\right] \leq \exp\left\{\frac{(4\epsilon)^2 N}{8} - (4\epsilon)\epsilon N\right\} = e^{-2\epsilon^2 N}.$$

By symmetry, $\mathbb{P}\left[\overline{X}_N - \mu < -\epsilon\right] \leq e^{-2\epsilon^2 N}$. Since the two tail events are disjoint, their probabilities add, and so

$$\begin{aligned}\mathbb{P}\left[|\overline{X}_N - \mu| > \epsilon\right] &= \mathbb{P}\left[\overline{X}_N - \mu > \epsilon\right] + \mathbb{P}\left[\overline{X}_N - \mu < -\epsilon\right] \\ &\leq e^{-2\epsilon^2 N} + e^{-2\epsilon^2 N} \\ &= 2e^{-2\epsilon^2 N}.\end{aligned}$$

Lemma 6.3 (Hoeffding's lemma)

Let $a \leq X \leq b$ be a random variable with $\mathbb{E}[X] = 0$. Then

$$M_X(s) \stackrel{\text{def}}{=} \mathbb{E}\left[e^{sX}\right] \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}. \tag{6.27}$$

Proof

Since $a \leq X \leq b$, we can write X as a linear combination of a and b:

$$X = \lambda b + (1-\lambda)a,$$

where $\lambda = \frac{X-a}{b-a}$. Since $\exp(\cdot)$ is a convex function, it follows that $e^{\lambda b + (1-\lambda)a} \leq \lambda e^b + (1-\lambda)e^a$. (Recall that h is convex if $h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y)$.) Therefore, we have

$$\begin{aligned}e^{sX} &\leq \lambda e^{sb} + (1-\lambda)e^{sa} \\ &= \frac{X-a}{b-a}e^{sb} + \frac{b-X}{b-a}e^{sa}.\end{aligned}$$

Taking expectations on both sides of the equation,

$$\mathbb{E}[e^{sX}] \leq \frac{-a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa},$$

because $\mathbb{E}[X] = 0$. Now, if we let $\theta = -\frac{a}{b-a}$, then

$$\begin{aligned}\frac{-a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} &= \theta e^{sb} + (1-\theta)e^{sa} \\ &= e^{sa}\left(1 - \theta + \theta e^{s(b-a)}\right) = \left(1 - \theta + \theta e^{s(b-a)}\right)e^{-s\theta(b-a)} \\ &= \left(1 - \theta + \theta e^u\right)e^{-\theta u} = e^{-\theta u + \log(1-\theta+\theta e^u)},\end{aligned}$$

where we let $u = s(b-a)$. This can be simplified as $\mathbb{E}[e^{sX}] \leq e^{\phi(u)}$ by defining

$$\phi(u) = -\theta u + \log(1 - \theta + \theta e^u).$$

The final step is to approximate $\phi(u)$. To this end, we use Taylor approximation:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\xi),$$

for some ξ between 0 and u. Since $\phi(0) = 0$, $\phi'(0) = 0$, and $\phi''(u) \leq \frac{1}{4}$ for all u, it follows that

$$\phi(u) = \frac{u^2}{2}\phi''(\xi) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

End of the proof.

What is so special about the Hoeffding's inequality?

- Since Hoeffding's inequality is derived from Chernoff's bound, it inherits the tightness. Hoeffding's inequality is much stronger than Chebyshev's inequality in bounding the tail distributions.
- Hoeffding's inequality is one of the few inequalities that do not require $\mathbb{E}[X]$ and $\text{Var}[X]$ on the right-hand side.
- A downside of the inequality is that boundedness is not always easy to satisfy. For example, if X_n is a Gaussian random variable, Hoeffding does not apply. There are more advanced inequalities for situations like these.

Interpreting Hoeffding's inequality. One way to interpret Hoeffding's inequality is to write the equation as

$$\mathbb{P}\left[\left|\overline{X}_N - \mu\right| > \epsilon\right] \leq \underbrace{2e^{-2\epsilon^2 N}}_{\delta},$$

which is equivalent to

$$\mathbb{P}\left[\left|\overline{X}_N - \mu\right| \leq \epsilon\right] \geq 1 - \delta.$$

This means that with a probability at least $1 - \delta$, we have

$$\overline{X}_N - \epsilon \leq \mu \leq \overline{X}_N + \epsilon.$$

If we let $\delta = 2e^{-2\epsilon^2 N}$, this becomes

$$\overline{X}_N - \sqrt{\frac{1}{2N}\log\frac{2}{\delta}} \leq \mu \leq \overline{X}_N + \sqrt{\frac{1}{2N}\log\frac{2}{\delta}}. \tag{6.28}$$

This inequality is a **confidence interval** (see Chapter 9). It says that with probability at least $1 - \delta$, the interval $[\overline{X}_N - \epsilon,\ \overline{X}_N + \epsilon]$ includes the true population mean μ.

There are two questions one can ask about the confidence interval:

- Given N and δ, what is the confidence interval? Equation (6.28) tells us that if we know N, to achieve a probability of at least $1-\delta$ the confidence interval will follow Equation (6.28). For example, if $N = 10{,}000$ and $\delta = 0.01$, $\sqrt{\frac{1}{2N}\log\frac{2}{\delta}} = 0.016$. Therefore, with a probability at least 99%, the true population mean μ will be included in the interval
$$\overline{X}_N - 0.016 \leq \mu \leq \overline{X}_N + 0.016.$$
- If we want to achieve a certain confidence interval, what is the N we need? If we are given ϵ and δ, the N we need is
$$\delta \leq 2e^{-2\epsilon^2 N} \quad \Rightarrow \quad N \geq \frac{\log\frac{2}{\delta}}{2\epsilon^2}.$$
For example, if $\delta = 0.01$ and $\epsilon = 0.01$, the N we need is $N \geq 26{,}500$.

When is Hoeffding's inequality used? Hoeffding's inequality is fundamental in modern machine learning theory. In this field, one often wants to quantify how well a learning algorithm performs with respect to the complexity of the model and the number of training samples. For example, if we choose a complex model, we should expect to use more training samples or overfit otherwise. Hoeffding's inequality provides an asymptotic description of the training error, testing error, and the number of training samples. The inequality is often used to compare the theoretical performance limit of one model versus another model. Therefore, although we do not need to use Hoeffding's inequality in this book, we hope you appreciate its tightness.

Closing Remark. We close this section by providing the historic context of Chernoff's inequality. Herman Chernoff, the discoverer of Chernoff's inequality, wrote the following many years after the publication of the original paper in 1952.

"*In working on an artificial example, I discovered that I was using the Central Limit Theorem for large deviations where it did not apply. This led me to derive the asymptotic upper and lower bounds that were needed for the tail probabilities. [Herman] Rubin claimed he could get these bounds with much less work, and I challenged him. He produced a rather simple argument, using Markov's inequality, for the upper bound. Since that seemed to be a minor lemma in the ensuing paper I published (Chernoff, 1952), I neglected to give him credit. I now consider it a serious error in judgment, especially because his result is stronger for the upper bound than the asymptotic result I had derived.*" — Herman Chernoff, "A career in statistics," in Lin et al., *Past, Present, and Future of Statistical Science* (2014), p. 35.

6.3 Law of Large Numbers

In this section, we present our first main result: the law of large numbers. We will discuss two versions of the law: the weak law and the strong law. We will also introduce two forms of convergence: convergence in probability and almost sure convergence.

6.3.1 Sample average

The law of large numbers is a probabilistic statement about the **sample average**. Suppose that we have a collection of i.i.d. random variables $X_1, \ldots, X_N$. The sample average of these N random variables is defined as follows:

Definition 6.5

The **sample average** of a sequence of random variables $X_1, \ldots, X_N$ is

$$\overline{X}_N = \frac{1}{N} \sum_{n=1}^{N} X_n. \tag{6.29}$$

If the random variables $X_1, \ldots, X_N$ are i.i.d. so that they have the same **population mean** $\mathbb{E}[X_n] = \mu$ (for $n = 1, \ldots, N$), then by the linearity of the expectation,

$$\mathbb{E}\left[\overline{X}_N\right] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[X_n] = \mu.$$

Therefore, the mean of $\overline{X}_N$ is the population mean μ.

The sample average, $\overline{X}_N$, plays an important role in statistics. For example, by surveying 10,000 Americans, we can find a sample average of their ages. Since we never have access to the true population mean, the sample average is an estimate, and since $\overline{X}_N$ is only an estimate, we need to ask how good the estimate is.

One reason we ask this question is that $\overline{X}_N$ is a finite-sample "approximation" of μ. More importantly, the root of the problem is that $\overline{X}_N$ itself is a random variable because $X_1, \ldots, X_N$ are all random variables. Since $\overline{X}_N$ is a random variable, there is a PDF of $\overline{X}_N$; there is a CDF of $\overline{X}_N$; there is $\mathbb{E}[\overline{X}_N]$; and there is $\mathrm{Var}[\overline{X}_N]$. Since $\overline{X}_N$ is a random variable, it has uncertainty. To say that we are confident about $\overline{X}_N$, we need to ensure that the uncertainty is within some tolerable range.

How do we control the uncertainty? We can compute the variance. If $X_1, \ldots, X_N$ are i.i.d. random variables with the same variance $\mathrm{Var}[X_n] = \sigma^2$ (for $n = 1, \ldots, N$), then

$$\mathrm{Var}\left[\overline{X}_N\right] = \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{Var}[X_n] = \frac{1}{N^2} \sum_{n=1}^{N} \sigma^2 = \frac{\sigma^2}{N}.$$

Therefore, the variance will shrink to 0 as N grows. In other words, the more samples we use to construct the sample average, the less deviation the random variable $\overline{X}_N$ will have.

Visualizing the sample average

To help you visualize the randomness of $\overline{X}_N$, we consider an experiment of drawing N Bernoulli random variables $X_1, \ldots, X_N$ with parameter $p = 1/2$. Since X_n is Bernoulli, it follows that

$$\mathbb{E}[X_n] = p \qquad \text{and} \qquad \mathrm{Var}[X_n] = p(1-p).$$

We construct a sample average $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Since X_n is a Bernoulli random variable, we know everything about $\overline{X}_N$. First, $N\overline{X}_N = \sum_{n=1}^{N} X_n$ is a binomial random variable, so $\overline{X}_N$ is a scaled binomial taking values in $\{0, \frac{1}{N}, \ldots, 1\}$. Second, the mean and variance of $\overline{X}_N$ are respectively

$$\mu_{\overline{X}_N} \overset{\text{def}}{=} \mathbb{E}[\overline{X}_N] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[X_n] = p,$$

$$\sigma^2_{\overline{X}_N} \overset{\text{def}}{=} \mathrm{Var}[\overline{X}_N] = \frac{1}{N^2} \sum_{n=1}^{N} \mathrm{Var}[X_n] = \frac{p(1-p)}{N}.$$

In **Figure 6.12**, we plot the random variables $\overline{X}_N$ (the black crosses) for every N. You can see that at each N, e.g., $N = 100$, there are many possible observations for $\overline{X}_N$ because $\overline{X}_N$ itself is a random variable. As N increases, we see that the deviation of the random variables becomes smaller. In the same plot, we show the bounds $\mu \pm 3\sigma_{\overline{X}_N}$, which are three standard deviations from the mean. We can see clearly that the bounds provide a very good envelope covering the random variables. As N goes to infinity, we can see that the standard deviation goes to zero, and so $\overline{X}_N$ approaches the true mean.

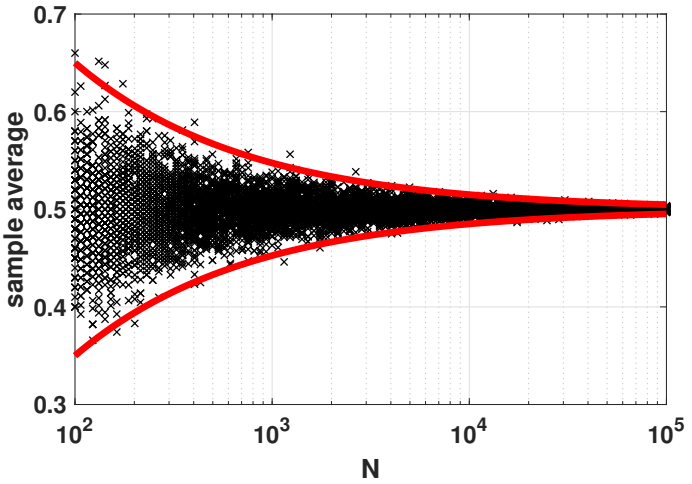


Figure 6.12: The weak law of large numbers. In this plot, we assume that $X_1, \ldots, X_N$ are i.i.d. Bernoulli random variables with a parameter p. The black crosses in the plot are the sample averages $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. The red curves are the ideal bounds $\mu_{\overline{X}_N} \pm 3\sigma_{\overline{X}_N}$, where $\mu_{\overline{X}_N} = p$ and $\sigma_{\overline{X}_N} = \sqrt{p(1-p)/N}$. As N grows, we observe that the variance shrinks to zero. Therefore, the sample average is converging to the true population mean.

For your reference, the MATLAB code and the Python code we used to generate the plot are shown below.

```
% MATLAB code to illustrate the weak law of large numbers
Nset = round(logspace(2,5,100));
for i=1:length(Nset)
    N = Nset(i);
    p = 0.5;
    x(:,i) = binornd(N, p, 1000,1)/N;
end
y = x(1:10:end,:)';
semilogx(Nset, y, 'kx'); hold on;
semilogx(Nset, p+3*sqrt(p*(1-p)./Nset), 'r', 'LineWidth', 4);
semilogx(Nset, p-3*sqrt(p*(1-p)./Nset), 'r', 'LineWidth', 4);
```

```
# Python code to illustrate the weak law of large numbers
import numpy as np
import matplotlib.pyplot as plt
```

```
import scipy.stats as stats
p = 0.5
Nset = np.round(np.logspace(2,5,100)).astype(int)
x = np.zeros((1000,Nset.size))
for i in range(Nset.size):
  N = Nset[i]
  x[:,i] = stats.binom.rvs(N, p, size=1000)/N
Nset_grid = np.broadcast_to(Nset, (1000, Nset.size))

plt.semilogx(Nset_grid, x,'ko');
plt.semilogx(Nset, p + 3*np.sqrt((p*(1-p))/Nset), 'r', linewidth=6)
plt.semilogx(Nset, p - 3*np.sqrt((p*(1-p))/Nset), 'r', linewidth=6)
```

Note the outliers for each N in **Figure 6.12**. For example, at $N = 10^2$ we see a point located near 0.7 on the y-axis. This point is outside three standard deviations. Is it normal? Yes. Being outside three standard deviations only says that the probability of having this outlier is **small**. It does not say that the outlier is **impossible**. Having a small probability does not exclude the possibility. By contrast, if you say that something will surely not happen you mean that there is not even a small probability. The former is a weaker statement than the latter. Therefore, even though we establish a three standard deviation envelope, there are points falling outside the envelope. As N grows, the chance of having a bad outlier becomes smaller. Therefore, the greater the N, the smaller the chance we will get an outlier.

If the random variables X_n are i.i.d., the above phenomenon is universal. Below is an example of the Poisson case.

Practice Exercise 6.8

Let $X_n \sim \text{Poisson}(\lambda)$. Define the sample average as $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Find the mean and variance of $\overline{X}_N$.

Solution. Since X_n is Poisson, we know that $\mathbb{E}[X_n] = \lambda$ and $\text{Var}[X_n] = \lambda$. So

$$\mathbb{E}[\overline{X}_N] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n] = \frac{1}{N}\sum_{n=1}^{N}\lambda = \lambda,$$

$$\text{Var}[\overline{X}_N] = \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}[X_n] = \frac{1}{N^2}\sum_{n=1}^{N}\lambda = \frac{\lambda}{N}.$$

Therefore, as $N \to \infty$, the variance $\text{Var}[\overline{X}_N] \to 0$.

6.3.2 Weak law of large numbers (WLLN)

The analysis of **Figure 6.12** shows us something important, namely that the convergence in a probabilistic way is different from that in a deterministic way. We now describe one fundamental result related to probabilistic convergence, known as the weak law of large numbers.

Theorem 6.16 (Weak law of large numbers)

Let $X_1, \ldots, X_N$ be a set of i.i.d. random variables with mean μ and *finite* variance σ^2, i.e., $\mathbb{E}[X_n^2] < \infty$. Let $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Then for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] = 0. \tag{6.30}$$

Proof

By Chebyshev's inequality,

$$\mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] \leq \frac{\mathrm{Var}[\overline{X}_N]}{\varepsilon^2} = \frac{\mathrm{Var}[X_n]}{N\varepsilon^2}.$$

Therefore, setting $N \to \infty$ we have

$$\lim_{N\to\infty} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] = \lim_{N\to\infty} \frac{\mathrm{Var}[X_n]}{N\varepsilon^2} = 0.$$

Example 6.10

Consider a set of i.i.d. random variables $X_1, \ldots, X_N$ where

$$X_n \sim \text{Gaussian}(\mu, \sigma^2).$$

Verify that the sample average $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$ follows the weak law of large numbers.

Solution: Since X_n is a Gaussian, the sample average $\overline{X}_N$ is also a Gaussian:

$$\overline{X}_N \sim \text{Gaussian}\left(\mu, \frac{\sigma^2}{N}\right).$$

Consider the probability $\mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right]$ for each N:

$$\begin{aligned}
\delta_N &\overset{\text{def}}{=} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] \\
&= \mathbb{P}\left[\overline{X}_N - \mu > \varepsilon\right] + \mathbb{P}\left[\overline{X}_N - \mu < -\varepsilon\right] \\
&= 1 - \Phi\left(\frac{\varepsilon\sqrt{N}}{\sigma}\right) + \Phi\left(-\frac{\varepsilon\sqrt{N}}{\sigma}\right) \\
&= 2\Phi\left(-\frac{\varepsilon\sqrt{N}}{\sigma}\right).
\end{aligned}$$

If we set $\sigma = 1$ and $\varepsilon = 0.1$, then

$$\delta_1 = 2\Phi\left(-\frac{0.1 \cdot 1}{1}\right) = 0.9203, \qquad \delta_5 = 2\Phi\left(-\frac{0.1 \cdot \sqrt{5}}{1}\right) = 0.8231,$$

$$\delta_{10} = 2\Phi\left(-\frac{0.1 \cdot \sqrt{10}}{1}\right) = 0.7518, \qquad \delta_{100} = 2\Phi\left(-\frac{0.1 \cdot \sqrt{100}}{1}\right) = 0.3173,$$

$$\delta_{1000} = 2\Phi\left(-\frac{0.1 \cdot \sqrt{1000}}{1}\right) = 0.0016.$$

As you can see, the sequence $\delta_1, \delta_2, \ldots, \delta_N, \ldots$ rapidly converges to 0 as N grows. In fact, since $\Phi(z)$ is an increasing function for $z < 0$ with $\Phi(-\infty) = 0$, it follows that

$$\lim_{N\to\infty} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] = \lim_{N\to\infty} 2\Phi\left(-\frac{\varepsilon\sqrt{N}}{\sigma}\right) = 0.$$

The weak law of large numbers is portrayed graphically in **Figure 6.13**. In this figure we draw several PDFs of the sample average $\overline{X}_N$. The shapes of the PDFs are getting narrower as the variance of the random variable shrinks. Since the PDFs become narrower, the probability $\mathbb{P}[|\overline{X}_N - \mu| > \varepsilon]$ becomes more unlikely. At the limit when $N \to \infty$, the probability vanishes. The weak law of large numbers asserts that this happens for any set of i.i.d. random variables. It says that the sequence of probability values $\delta_N \stackrel{\text{def}}{=} \mathbb{P}[|\overline{X}_N - \mu| > \varepsilon]$ will converge to zero.

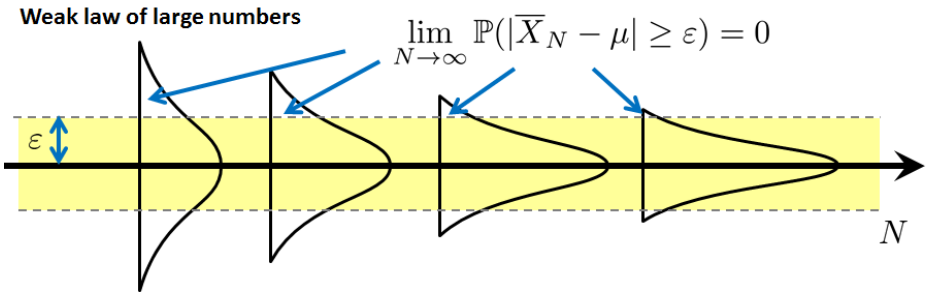


Figure 6.13: The weak law of large numbers states that as N increases, the variance of the sample average $\overline{X}_N$ shrinks. As a result, the probability $\mathbb{P}[|\overline{X}_N - \mu| > \varepsilon]$ decreases and eventually vanishes. Note that the convergence here is that of the sequence of probabilities $\mathbb{P}[|\overline{X}_N - \mu| > \varepsilon]$, which is just a sequence of numbers.

What is the weak law of large numbers?

Let $\overline{X}_N$ be the sample average of i.i.d. random variables $X_1, \ldots, X_N$.

$$\lim_{N\to\infty} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] = 0. \tag{6.31}$$

- For details, see Theorem 6.16.

- The WLLN concerns the sequence of probability **values** $\delta_N = \mathbb{P}[|\overline{X}_N - \mu| > \varepsilon]$.
- The probabilities converge to zero as N grows.
- It is weak because having a small probability does not exclude the possibility of happening.

6.3.3 Convergence in probability

The example above tells us that in order to show convergence, we need to first compute the probability δ_n of each event and then take the limit of the sequence, e.g., the one shown in the table below:

δ_1	δ_5	δ_{10}	δ_{100}	δ_{1000}	δ_{10000}
0.9203	0.8231	0.7518	0.3173	0.0016	1.5240×10^{-23}

Therefore, the convergence is the convergence of the **probability**. Since $\{\delta_1, \delta_2, \ldots\}$ is a sequence of real numbers (between 0 and 1), any convergence results for real numbers apply here.

Note that the convergence controls only the probabilities. Probability means chance. Therefore, having the limit converging to zero only means that the chance of happening is becoming smaller and smaller. However, at any N, there is still a chance that some bad event can happen.

What do we mean by a bad event? Assume that X_n are fair coins. The sample average $\overline{X}_N = (1/N)\sum_{n=1}^{N} X_n$ is more or less equal to $1/2$ as N grows. However, even if N is a large number, say $N = 1000$, we are still not certain that the sample average is **exactly** $1/2$. It is possible, though very unlikely, that we obtain 1000 heads or 1000 tails (so that the sample average is "1" or "0"). The bottom line is: Having a probability converging to zero only means that for any tolerance level we can always find an N large enough so that the probability is smaller than that tolerance.

The type of convergence described by the weak law of large numbers is known as the **convergence in probability**.

Definition 6.6

A sequence of random variables $A_1, \ldots, A_N$ **converges in probability** to a deterministic number α if for every $\varepsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left[|A_N - \alpha| > \varepsilon\right] = 0. \tag{6.32}$$

We write $A_N \xrightarrow{p} \alpha$ to denote convergence in probability.

The following two examples illustrate how to prove convergence in probability.

Example 6.11

Let $X_1, \ldots, X_N$ be i.i.d. random variables with $X_n \sim \text{Uniform}(0,1)$. Define $A_N = \min(X_1, \ldots, X_N)$. Show that A_N converges in probability to zero.

Solution. (Without determining the PDF of A_N, we notice that as N increases, the value of A_N will likely decrease. Therefore, we should expect A_N to converge to

zero.) Pick an $\varepsilon > 0$. It follows that

$$\begin{aligned}\mathbb{P}[|A_N - 0| \geq \varepsilon] &= \mathbb{P}[\min(X_1, \ldots, X_N) \geq \varepsilon], \qquad \text{because } X_n \geq 0 \\ &= \mathbb{P}[X_1 \geq \varepsilon \text{ and } \cdots \text{ and } X_N \geq \varepsilon] \\ &= \mathbb{P}\big(X_1 \geq \varepsilon\big) \cdots \mathbb{P}\big(X_N \geq \varepsilon\big) = (1-\varepsilon)^N.\end{aligned}$$

Setting the limit of $N \to \infty$, we conclude that

$$\lim_{N\to\infty} \mathbb{P}[|A_N - 0| \geq \varepsilon] = \lim_{N\to\infty} (1-\varepsilon)^N = 0.$$

Therefore, A_N converges to zero in probability.

Practice Exercise 6.9

Let $X \sim \text{Exponential}(1)$. By evaluating the CDF, we know that $\mathbb{P}[X \geq x] = e^{-x}$. Let $A_N = X/N$. Prove that A_N converges to zero in probability.
Solution. For any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}[|A_N - 0| \geq \varepsilon] &= \mathbb{P}[A_N \geq \varepsilon] \\ &= \mathbb{P}[X \geq N\varepsilon] \\ &= e^{-N\varepsilon}.\end{aligned}$$

Putting $N \to \infty$ on both sides of the equation gives us

$$\lim_{N\to\infty} \mathbb{P}[|A_N - 0| \geq \varepsilon] = \lim_{N\to\infty} e^{-N\varepsilon} = 0.$$

Thus, A_N converges to zero in probability.

Example 6.12

Construct an example such that A_N converges in probability to something, but $\mathbb{E}[A_N]$ does not converge to the same thing.
Solution. Consider a sequence of random variables A_N such that

$$\mathbb{P}[A_N = a] = \begin{cases} 1 - \frac{1}{N}, & a = 0, \\ \frac{1}{N}, & a = N^2, \\ 0, & \text{otherwise.} \end{cases}$$

The PDF of the random variable A_N is shown in **Figure 6.14**.

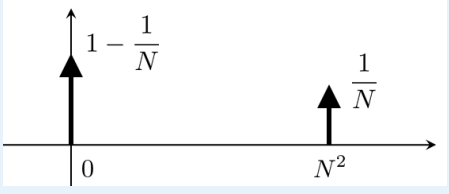


Figure 6.14: Probability density function of the random variable A_N.

We first show that A_N converges in probability to zero. Let $\varepsilon > 0$ be a fixed constant. Since $\varepsilon > 0$,

$$\mathbb{P}[A_N \geq \varepsilon] = \frac{1}{N}$$

for any $N > \sqrt{\varepsilon}$. Therefore, we have that

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}[|A_N - 0| \geq \varepsilon] &= \lim_{N\to\infty} \mathbb{P}[A_N \geq \varepsilon] \\ &= \lim_{N\to\infty} \frac{1}{N} = 0.\end{aligned}$$

Hence, A_N converges to 0 in probability.
However, $\mathbb{E}[A_N]$ does not converge to zero, because

$$\begin{aligned}\mathbb{E}[A_N] &= 0 \cdot \left(1 - \frac{1}{N}\right) + N^2 \cdot \frac{1}{N} \\ &= N.\end{aligned}$$

So $\mathbb{E}[A_N]$ goes to infinity as N grows.

6.3.4 Can we prove WLLN using Chernoff's bound?

The following discussion of using Chernoff's bound to prove WLLN can be skipped if this is your first time reading the book.

In proving WLLN we use Chebyshev's inequality. Can we use Chernoff's inequality (or Hoeffding's) to prove the result? Yes, we can use them. However, notice that the task here is to prove convergence, not to find the *best* convergence. Finding the best convergence means finding the fastest decay rate of the probability sequence. Chernoff's bound (and Hoeffding's inequality) offers a better decay rate. However, Chernoff's bound needs to be customized for individual random variables. For example, Chernoff's bound for Gaussian is different from Chernoff's bound for exponential. This result makes Chebyshev the most convenient bound because it only requires the variance to be bounded.

What if we insist on using Chernoff's bound in proving the WLLN? We can do that for specific random variables. Let's consider two examples. The first example is the Gaussian random variable where $X_n \sim \text{Gaussian}(0, \sigma^2)$. We know that $\overline{X}_N \sim \text{Gaussian}(0, \sigma^2/N)$. Chernoff's bound shows that

$$\mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] \leq 2\exp\left\{-\frac{\varepsilon^2 N}{2\sigma^2}\right\},$$

Taking the limit on both sides, we have

$$\lim_{N\to\infty} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] = \lim_{N\to\infty} 2\exp\left\{-\frac{\varepsilon^2 N}{2\sigma^2}\right\} = 0.$$

Note that the rate of convergence here is exponential. The rate of convergence offered by Chebyshev is only linear. Of course, you may argue that since X_n is Gaussian we have closed-form expressions about the probability, so we do not need Chernoff's bound. This is a legitimate point, and so here is an example where we do not have a closed-form expression for the probability.

Consider a sequence of arbitrary i.i.d. random variables $X_1, \ldots, X_N$ with $0 \leq X_n \leq 1$. Then Hoeffding's inequality tells us that

$$\mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] \leq 2\exp\left\{-2\varepsilon^2 N\right\}.$$

Taking the limit on both sides, we have

$$\lim_{N\to\infty} \mathbb{P}\left[|\overline{X}_N - \mu| > \varepsilon\right] = \lim_{N\to\infty} 2\exp\left\{-2\varepsilon^2 N\right\} = 0.$$

Again, we obtain a WLLN result, this time for i.i.d. random variables $X_1, \ldots, X_N$ with $0 \leq X_n \leq 1$.

As you can see from these two examples, WLLN can be proved in multiple ways depending on how general the random variables need to be.

End of the discussions.

6.3.5 Does the weak law of large numbers always hold?

The following discussion of the failure of the weak law of large numbers can be skipped if this is your first time reading the book.

The weak law of large numbers does not always hold. Recall that when we prove the weak law of large numbers using Chebyshev's inequality, we implicitly require that the variance $\mathrm{Var}[\overline{X}_N]$ is finite. (Look at the condition that $\mathbb{E}[X^2] < \infty$.) Thus for distributions whose variance is unbounded, Chebyshev's inequality does not hold. One example is the Cauchy distribution. The PDF of a Cauchy distribution is

$$f_X(x) = \frac{\gamma}{\pi(\gamma^2 + x^2)},$$

where γ is a parameter. Letting $\gamma = 1$,

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} \frac{x^2}{\pi(1+x^2)}\, dx = \frac{1}{\pi}\int_{-\infty}^{\infty} 1 - \frac{1}{1+x^2}\, dx \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} dx - \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\, dx = \frac{1}{\pi}\left[x - \tan^{-1}(x)\right]\Big|_{x=-\infty}^{\infty} = \infty.\end{aligned}$$

Since the second moment is unbounded, the variance of X will also be unbounded.

A perceptive reader may observe that even if $\mathbb{E}[X^2]$ is unbounded, it does not mean that the tail probability is unbounded. This is correct. However, for Cauchy distributions,

we can show that the sample average $\overline{X}_N$ does not converge to the mean when $N \to \infty$ (and so the WLLN fails). To see this, we note that

$$\frac{1}{\pi(1+x^2)} \leftrightarrow e^{-|\omega|}.$$

So for the sample average $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$, the characteristic function is

$$\mathbb{E}[e^{-j\omega\overline{X}_N}] = \mathbb{E}[e^{-\frac{j\omega}{N}\sum_{n=1}^{N} X_n}] = \prod_{n=1}^{N} \mathbb{E}[e^{-\frac{j\omega}{N}X_n}] = \left[e^{-\frac{|\omega|}{N}}\right]^N = e^{-|\omega|},$$

which remains a Cauchy distribution with $\gamma = 1$. Therefore, we have that

$$\begin{aligned}\mathbb{P}[|\overline{X}_N| \leq \varepsilon] &= \int_{-\varepsilon}^{\varepsilon} \frac{1}{\pi(1+x^2)}\, dx \\ &= 2\int_{0}^{\varepsilon} \frac{1}{\pi(1+x^2)}\, dx = \frac{2}{\pi}\tan^{-1}(\varepsilon).\end{aligned}$$

Thus no matter how many samples we have, $\mathbb{P}[|\overline{X}_N| \leq \varepsilon]$ will never converge to 1 (so $\mathbb{P}[|\overline{X}_N| > \varepsilon]$ will never converge to 0). Therefore, WLLN does not hold.

End of the discussion.

6.3.6 Strong law of large numbers

Since there is a "weak" law of large numbers, you will not be surprised to learn that there is a strong law of large numbers. The strong law is *more restrictive* than the weak law. Any sequence satisfying the strong law will satisfy the weak law, but not vice versa. Since the strong law is "stronger", the proof is more involved.

Theorem 6.17 (Strong law of large numbers)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with common mean μ and variance σ^2. Assume $\mathbb{E}[X_n^4] < \infty$. Let $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$ be the sample average. Then

$$\mathbb{P}\left[\lim_{N\to\infty} \overline{X}_N = \mu\right] = 1. \tag{6.33}$$

The strong law flips the order of limit and probability. As you can see, the difference between the strong law and the weak law is the order of the limit and the probability. In the weak law, the limit is **outside** the probability, whereas, in the strong law, the limit is **inside** the probability. This switch in order makes the interpretation of the result fundamentally different. In the final analysis, the weak law concerns the limit of a sequence of probabilities (which are just real numbers between 0 and 1). However, the strong law concerns the limit of a sequence of random variables. The strong law answers the question, what is the **limiting object** of the sample average as N grows?

The strong law concerns the limiting object, not a sequence of numbers. What is the "limiting object"? If we denote $\overline{X}_N$ as the sample average using N samples, then we know that $\overline{X}_1$ is a random variable, $\overline{X}_2$ is a random variable, and all $\overline{X}_n$'s are random variables.

So we have a sequence of random variables. As N goes to infinity, we can ask about the limiting object $\lim_{N\to\infty} \overline{X}_N$. However, even without any deep analysis, you should be able to see that $\lim_{N\to\infty} \overline{X}_N$ is another random variable. The strong law says that this limiting object will "successfully" become a deterministic number μ, after a finite number of "failures".

The strong law asserts that there are a finite number of failures. Let us explain "success" and "failure". Fix any tolerance $\epsilon > 0$ and call it a *failure* whenever $|\overline{X}_N - \mu| > \epsilon$. $\overline{X}_N$ is a random variable, so it fluctuates and failures do occur. The strong law says that, with probability 1, only *finitely many* of these failures ever happen: there is some (random) index N_0 beyond which $|\overline{X}_N - \mu| \leq \epsilon$ for **every** $N \geq N_0$. Note that this does *not* say that $\overline{X}_N$ ever becomes exactly equal to μ — for a continuous distribution $\mathbb{P}[\overline{X}_N = \mu] = 0$ for every N. What it says is that the whole sequence settles inside any tolerance band you name, and stays there. That is a 100% guarantee about the sequence. The weak law only guarantees that each individual $\mathbb{P}[|\overline{X}_N - \mu| > \epsilon]$ becomes small.

A good example for differentiating the weak law and the strong law is an electronic dictionary that improves itself every time you use it. The weak law says that if you use the dictionary for a long period, the probability of making an error will become small. You will still get an error once in a while, but the probability is very small. This is a 99.99% guarantee, and it is the weak law. The strong law says that the number of failures is finite. After you have gone through this finite number of failures, you will be completely free of error. This is a 100% guarantee by the strong law. When will you hit this magical number? The strong law does not say when; it only asserts the existence of this number. However, this existence is already good enough in many ways. It gives a certificate of assurance, whereas the weak law still has uncertainty.

Strong law $\neq$ deterministic. If the strong law offers a 100% guarantee, does it mean that it is a deterministic guarantee? No, the strong law is still a probabilistic statement because we are still using $\mathbb{P}[\cdot]$ to measure an event. The event can include measure-zero subsets, and the measure-zero subsets can be huge. For example, the set of rational numbers on the real line is a measure-zero set when measuring the probability using an integration. The strong law does not handle those measure-zero subsets.

6.3.7 Almost sure convergence

The discussion below can be skipped if this is your first time reading the book.

The type of convergence used by the strong law of large numbers is the **almost sure convergence**. It is defined formally as follows.

Definition 6.7

A sequence of random variables $A_1, \ldots, A_N$ **converges almost surely** to α if

$$\mathbb{P}\left[\lim_{N\to\infty} A_N = \alpha\right] = 1. \tag{6.34}$$

We write $A_N \overset{a.s.}{\to} \alpha$ to denote almost sure convergence.

To prove almost sure convergence, one needs to show that, with probability 1, the

sequence of *numbers* $A_1(\xi), A_2(\xi), \ldots$ converges to α in the ordinary sense of calculus: for every $\epsilon > 0$ there are only finitely many N with $|A_N - \alpha| > \epsilon$.

Example 6.13

[a] Construct a sequence of events that converges almost surely.

Solution. Let $X_1, \ldots, X_N$ be i.i.d. random variables such that $X_n \sim \text{Uniform}(0,1)$. Define $A_N = \min(X_1, \ldots, X_N)$. Since A_N is nonincreasing and is bounded below by zero, it must have a limit. Let us call this limit

$$A \overset{\text{def}}{=} \lim_{N\to\infty} A_N.$$

Then we can show that

$$\begin{aligned}
\mathbb{P}[A \geq \epsilon] &= \mathbb{P}[\min(X_1, X_2, \ldots) \geq \epsilon] \\
&\overset{(a)}{\leq} \mathbb{P}[\min(X_1, X_2, \ldots, X_N) \geq \epsilon] \\
&\overset{(b)}{=} \mathbb{P}[X_1 \geq \epsilon \text{ and } X_2 \geq \epsilon \text{ and } \cdots \text{ and } X_N \geq \epsilon] \\
&= (1-\epsilon)^N,
\end{aligned}$$

where (a) holds because there are more elements in $(X_1, X_2, \ldots)$ than in $(X_1, X_2, \ldots, X_N)$. Therefore, the minimum value of the former is less than the minimum value of the latter. (b) holds because if $\min(X_1, X_2, \ldots, X_N) \geq \epsilon$, then $X_n \geq \epsilon$ for all n.

Since $\mathbb{P}[A \geq \epsilon] \leq (1-\epsilon)^N$ for any N, the statement still holds as $N \to \infty$. Thus,

$$\mathbb{P}[A \geq \epsilon] \leq \lim_{N\to\infty} (1-\epsilon)^N = 0.$$

This shows $\mathbb{P}[A \geq \epsilon] = 0$ for any positive ϵ. So $\mathbb{P}[A > 0] = 0$, and hence $\mathbb{P}[A = 0] = 1$. Since A is the limit of A_N, we conclude that

$$\mathbb{P}\left[\lim_{N\to\infty} A_N = 0\right] = \mathbb{P}\left[A = 0\right] = 1.$$

So A_N converges to 0 almost surely.

[a] This example is modified from Bertsekas and Tsitsiklis, *Introduction to Probability*, Chapter 5.5.

Example 6.14

[a] Construct an example where a sequence of events converges in probability but does not converge almost surely.

Solution. Consider a discrete time arrival process. The set of times is partitioned into

consecutive intervals of the form

$$\begin{aligned} I_1 &= \{2,3\}, \\ I_2 &= \{4,5,6,7\}, \\ I_3 &= \{8,9,10,\ldots,15\}, \\ &\vdots \\ I_k &= \{2^k, 2^k+1, \ldots, 2^{k+1}-1\}. \end{aligned}$$

Therefore, the length of each interval is $|I_1| = 2$, $|I_2| = 4$, …, $|I_k| = 2^k$.
During each interval, there is exactly one arrival. Define Y_n as a binary random variable such that for every $n \in I_k$,

$$Y_n = \begin{cases} 1, & \text{with probability } \frac{1}{|I_k|}, \\ 0, & \text{with probability } 1 - \frac{1}{|I_k|}. \end{cases}$$

For example, if $n \in \{2,3\}$, then $\mathbb{P}[Y_n = 1] = \frac{1}{2}$. If $n \in \{4,5,6,7\}$, then $\mathbb{P}[Y_n = 1] = \frac{1}{4}$. In general, we have that

$$\lim_{n\to\infty} \mathbb{P}[Y_n = 1] = \lim_{n\to\infty} \frac{1}{|I_k|} = \lim_{n\to\infty} \frac{1}{2^k} = 0,$$

and hence

$$\lim_{n\to\infty} \mathbb{P}[Y_n = 0] = \lim_{n\to\infty} 1 - \frac{1}{2^k} = 1.$$

Therefore, Y_n converges to 0 in probability.
However, when we carry out the experiment, there is exactly one arrival per interval according to the problem conditions. Since we have an infinite number of intervals $I_1, I_2, \ldots$, we will have an infinite number of arrivals in total. As a result, $Y_n = 1$ **infinitely often**. We do not know which Y_n will equal 1 and which Y_n will equal 0. However, we know that there are infinitely many Y_n that are equal to 1 — and, since each interval I_k has length $2^k \geq 2$, also infinitely many Y_n that are equal to 0.
This is exactly what breaks the almost sure convergence. A sequence of numbers that contains infinitely many 1's and infinitely many 0's does not converge at all: no matter how far out you go, you can still find a later 1 and a later 0. So for (almost) every realization the limit $\lim_{n\to\infty} Y_n$ *does not exist*, and in particular

$$\mathbb{P}\left[\lim_{n\to\infty} Y_n = 0\right] = 0.$$

Therefore, Y_n converges to 0 in probability but not almost surely.

[a] This example is modified from Bertsekas and Tsitsiklis, *Introduction to Probability*, Chapter 5.5.

End of the discussions.

6.3.8 Proof of the strong law of large numbers

The strong law of large numbers can be proved in several ways. We present a proof based on Bertsekas and Tsitsiklis, *Introduction to Probability*, Problems 5.16 and 5.17, which require a finite fourth moment $\mathbb{E}[X_n^4] < \infty$. An alternative proof that requires only $\mathbb{E}[X_n] < \infty$ is from Billingsley, *Probability and Measure*, Theorem 22.1.

The proof of the strong law of large numbers is beyond the scope of this book. This section is optional.

Lemma 6.4

Consider non-negative random variables $X_1, \ldots, X_N$. Assume that

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] < \infty. \tag{6.35}$$

Then $X_n \overset{a.s.}{\to} 0$.

Proof

Let $S = \sum_{n=1}^{\infty} X_n$. Note that S is a random variable, and our assumption is that $\mathbb{E}[S] < \infty$. Thus, we argue that $S < \infty$ with probability 1. If not, then S will have a positive probability of being ∞. But if this happens, we will have $\mathbb{E}[S] = \infty$ because (by the law of total expectation):

$$\mathbb{E}[S] = \underbrace{\mathbb{E}[S \mid S = \text{infinite}]}_{=\infty}\mathbb{P}[S = \text{infinite}] + \mathbb{E}[S \mid S = \text{finite}]\mathbb{P}[S = \text{finite}].$$

Now, since S is finite, the sequence $\{X_1, \ldots, X_N, \ldots\}$ must converge to zero. Otherwise, if X_n is converging to some constant $c > 0$, then summing the tail of the sequence (which contains infinitely many terms) gives infinity:

$$S = \underbrace{X_1 + \cdots}_{=\text{finite}} + \underbrace{X_N + \cdots +}_{=\text{infinite}}.$$

Since the probability of S being finite is 1, it follows that $\{X_1, \ldots, X_N, \ldots\}$ is converging to zero with probability 1.

Theorem 6.18 (Strong law of large numbers)

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with common mean μ and variance σ^2. Assume $\mathbb{E}[X_n^4] < \infty$. Let $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$ be the sample average. Then

$$\mathbb{P}\left[\lim_{N\to\infty} \overline{X}_N = \mu\right] = 1. \tag{6.36}$$

Proof

We first prove the case where $\mathbb{E}[X_n] = 0$. To establish that $\overline{X}_N \to 0$ with probability 1, we use the lemma to show that

$$\mathbb{E}\left[\sum_{N=1}^{\infty} |\overline{X}_N|\right] < \infty.$$

But to show $\mathbb{E}[\sum_{N=1}^{\infty} |\overline{X}_N|] < \infty$, we note that $|x| \leq 1 + x^4$. Therefore, $\mathbb{E}[\sum_{N=1}^{\infty} |\overline{X}_N|] \leq 1 + \mathbb{E}[\sum_{N=1}^{\infty} \overline{X}_N^4]$, and hence we just need to show that

$$\mathbb{E}\left[\sum_{N=1}^{\infty} \overline{X}_N^4\right] < \infty.$$

Let us expand the term $\mathbb{E}[\overline{X}_N^4]$ as follows:

$$\mathbb{E}[\overline{X}_N^4] = \frac{1}{N^4} \sum_{n_1=1}^{N} \sum_{n_2=1}^{N} \sum_{n_3=1}^{N} \sum_{n_4=1}^{N} \mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}].$$

There are five possibilities for $\mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}]$:

- All indices are different. Then

$$\mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}] = \mathbb{E}[X_{n_1}]\mathbb{E}[X_{n_2}]\mathbb{E}[X_{n_3}]\mathbb{E}[X_{n_4}] = 0 \cdot 0 \cdot 0 \cdot 0 = 0.$$

- One index is different from other three indices. For example, if n_1 is different from n_2, n_3, n_4, then

$$\mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}] = \mathbb{E}[X_{n_1}]\mathbb{E}[X_{n_2} X_{n_3} X_{n_4}] = 0 \cdot \mathbb{E}[X_{n_2} X_{n_3} X_{n_4}] = 0.$$

- Two indices are identical. For example, if $n_1 = n_3$, and $n_2 = n_4$, then

$$\mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}] = \mathbb{E}[X_{n_1} X_{n_3}]\mathbb{E}[X_{n_2} X_{n_4}] = \mathbb{E}[X_{n_1}^2 X_{n_2}^2].$$

  There are altogether $3N(N-1)$ of these cases: $N(N-1)$ comes from choosing N followed by choosing $N-1$, and 3 accounts for $n_1 = n_2 \neq n_3 = n_4$, $n_1 = n_3 \neq n_2 = n_4$, and $n_1 = n_4 \neq n_2 = n_3$.

- Two indices are identical, and two indices are different. For example, if $n_1 = n_3$ but n_2 and n_4 are different. Then

$$\begin{aligned}\mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}] &= \mathbb{E}[X_{n_1} X_{n_3}]\mathbb{E}[X_{n_2}]\mathbb{E}[X_{n_4}] \\ &= \mathbb{E}[X_{n_1}^2] \cdot 0 \cdot 0 = 0.\end{aligned}$$

- All indices are identical. If $n_1 = n_2 = n_3 = n_4$, then

$$\mathbb{E}[X_{n_1} X_{n_2} X_{n_3} X_{n_4}] = \mathbb{E}[X_{n_1}^4].$$

  There are altogether N cases of this.

Therefore, it follows that

$$\mathbb{E}[\overline{X}_N^4] = \frac{N\mathbb{E}[X_1^4] + 3N(N-1)\mathbb{E}[X_1^2 X_2^2]}{N^4}.$$

Since $xy \leq (x^2 + y^2)/2$, it follows that

$$\begin{aligned}\mathbb{E}[X_1^2 X_2^2] &\leq \mathbb{E}[(X_1^2)^2 + (X_2^2)^2]/2 \\ &= \mathbb{E}[X_1^4 + X_2^4]/2 \\ &= \mathbb{E}[X_1^4].\end{aligned}$$

Substituting into the previous result,

$$\begin{aligned}\mathbb{E}[\overline{X}_N^4] &\leq \frac{N\mathbb{E}[X_1^4] + 3N(N-1)\mathbb{E}[X_1^4]}{N^4} \\ &\leq \frac{3N^2}{N^4}\mathbb{E}[X_1^4] \\ &= \frac{3}{N^2}\mathbb{E}[X_1^4].\end{aligned}$$

Now, let us complete the proof.

$$\mathbb{E}\left[\sum_{N=1}^{\infty} \overline{X}_N^4\right] \leq \mathbb{E}\left[\sum_{N=1}^{\infty} \frac{3}{N^2}\mathbb{E}[X_1^4]\right] < \infty,$$

because $\sum_{N=1}^{\infty}(1/N^2)$ is the Basel problem with a solution that $\sum_{N=1}^{\infty}(1/N^2) = \pi^2/6$. Consequently, we have shown that $\mathbb{E}\left[\sum_{N=1}^{\infty} \overline{X}_N^4\right] < \infty$, which implies $\mathbb{E}\left[\sum_{N=1}^{\infty} |\overline{X}_N|\right] < \infty$. Then, by the lemma, we have $\overline{X}_N$ converging to 0 with probability 1, which proves the result.
If $\mathbb{E}[X_n] = \mu$, then just replace X_n with $Y_n = X_n - \mu$ in the above arguments. Then we can show that $\overline{Y}_N$ converges to 0 with probability 1, which is equivalent to $\overline{X}_N$ converging to μ with probability 1.

End of the proof of strong law of large numbers.

6.4 Central Limit Theorem

The law of large numbers tells us the **mean** of the sample average $\overline{X}_N = (1/N)\sum_{n=1}^{N} X_n$. However, if you recall our experiment of throwing N dice and inspecting the PDF of the sum of the numbers, you may remember that the convolution of an infinite number of uniform distributions gives us a Gaussian distribution. For example, we show a sequence of experiments in **Figure 6.15**. In each experiment, we throw N dice and count the sum. Therefore, if each face of the die is denoted as X_n, then the sum is $X_1 + \cdots + X_N$. We plot

the PDF of the sum. As you can see in the figure, $X_1 + \cdots + X_N$ converges to a Gaussian. This phenomenon is explained by the **Central Limit Theorem (CLT)**.

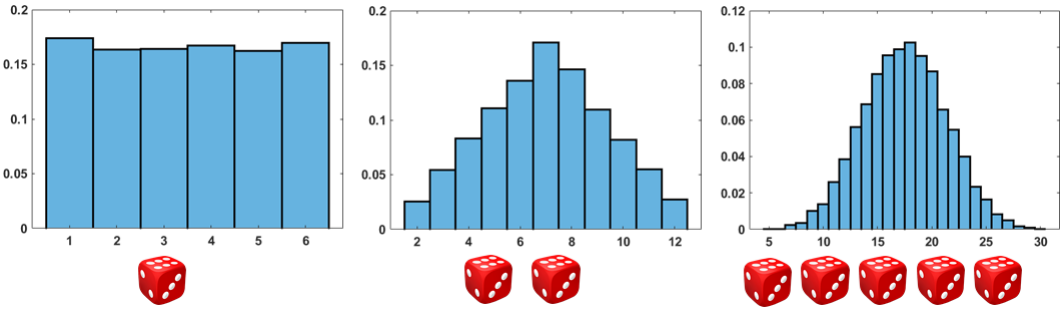


Figure 6.15: Pictorial illustration of the Central Limit Theorem. Suppose we throw a die and record the face. [Left] If we only have one die, then the distribution of the face is uniform. [Middle] If we throw two dice, the distribution is the convolution of two uniform distributions. This will give us a triangle distribution. [Right] If we throw five dice, the distribution is becoming similar to a Gaussian. The Central Limit Theorem says that as N goes to infinity, the distribution of the sum will converge to a Gaussian.

What does the Central Limit Theorem say? Let $\overline{X}_N$ be the sample average, and let $Z_N = \sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right)$ be the normalized variable. The Central Limit Theorem is as follows:

> **Central Limit Theorem**:
>
> The CDF of Z_N is converging pointwise to the CDF of Gaussian(0,1).

Note that we are very careful here. We are not saying that the PDF of Z_N is converging to the PDF of a Gaussian, nor are we saying that the random variable Z_N is converging to a Gaussian random variable. We are only saying that the **values** of the CDF are converging pointwise. The difference is subtle but important.

To understand the difficulty and the core ideas, we first present the concept of convergence in distribution.

6.4.1 Convergence in distribution

Definition 6.8

Let $Z_1, \ldots, Z_N$ be random variables with CDFs $F_{Z_1}, \ldots, F_{Z_N}$ respectively. We say that a sequence of $Z_1, \ldots, Z_N$ **converges in distribution** to a random variable Z with CDF F_Z if

$$\lim_{N\to\infty} F_{Z_N}(z) = F_Z(z), \tag{6.37}$$

for every continuous point z of F_Z. We write $Z_N \xrightarrow{d} Z$ to denote convergence in distribution.

This definition involves many concepts, which we will discuss one by one. However, the definition can be summarized in a nutshell as follows.

Convergence in distribution = values of the CDF converge.

Example 6.15

(**Bernoulli**) Consider flipping a fair coin N times. Denote each coin flip as a Bernoulli random variable $X_n \sim \text{Bernoulli}(p)$, where $n = 1, 2, \ldots, N$. Define Z_N as the sum of N Bernoulli random variables, so that

$$Z_N = \sum_{n=1}^{N} X_n.$$

We know that the resulting random variable Z_N is a binomial random variable with mean Np and variance $Np(1-p)$. Let us plot the PDF $f_{Z_N}(z)$ as shown in **Figure 6.16**.

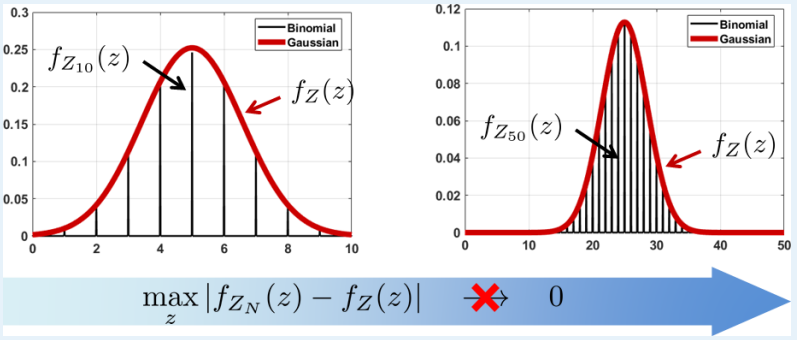


Figure 6.16: Convergence in distribution. The convergence in distribution concerns the convergence of the values of the CDF (not the PDF). In this figure, we let $Z_N = X_1 + \cdots + X_N$, where X_N is a Bernoulli random variable with parameter p. Since a sum of Bernoulli random variables is a binomial, Z_N is a binomial random variable with parameters (N, p). We plot the PDF of Z_N, which is a train of delta functions, and compare it with the Gaussian PDF. Observe that the error, $\max_z |f_{Z_N}(z) - f_Z(z)|$, does *not* converge to 0. The PDF of Z_N is a binomial. A binomial is always a binomial. It will not turn into a Gaussian.

The first thing we notice in the figure is that as N increases, the PDF of the binomial has an envelope that is "very Gaussian". So one temptation is to say that the random variable Z_N is converging to another random variable Z. In addition, we would think that the PDFs converge in the sense that for *all* z,

$$f_{Z_N}(z) = \binom{N}{z} p^z (1-p)^{N-z} \quad \longrightarrow \quad f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(z-\mu)^2}{2\sigma^2} \right\},$$

where $\mu = Np$ and $\sigma^2 = Np(1-p)$.
Unfortunately this argument does not work, because $f_Z(z)$ is continuous but $f_{Z_N}(z)$ is discrete. The sample space of Z_N and the sample space of Z are completely different. In fact, if we write f_{Z_N} as an impulse train, we observe that

$$f_{Z_N}(z) = \sum_{i=0}^{N} \binom{N}{i} p^i (1-p)^{N-i} \delta(z-i).$$

Clearly, no matter how big the N is, the difference $|f_{Z_N}(z) - f_Z(z)|$ will never go to zero for non-integer values of z. Mathematically, we can show that

$$\max_z \; |f_{Z_N}(z) - f_Z(z)| \not\longrightarrow 0,$$

as $N \to \infty$. Z_N is a binomial random variable regardless of N. It will not become a Gaussian.
If $f_{Z_N}(z)$ is not converging to a Gaussian PDF, how do we explain the convergence? The answer is to look at the CDF. For discrete PDFs such as a binomial random variable, the CDF is a staircase function. What we can show is that

$$F_{Z_N}(z) = \sum_{i=0}^{z} \binom{N}{i} p^i (1-p)^{N-i} \quad \longrightarrow \quad F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(t-\mu)^2}{2\sigma^2} \right\} dt.$$

The difference between the PDF convergence and the CDF convergence is that the PDF does not allow a meaningful "distance" between a discrete function and continuous function. For CDF, the distance is well defined by taking the difference between the staircase function and the continuous function. For example, we can compute

$$|F_{Z_N}(z) - F_Z(z)|, \qquad \text{for all continuous points } z \text{ of } F_Z,$$

and show that

$$\max_z |F_{Z_N}(z) - F_Z(z)| \longrightarrow 0.$$

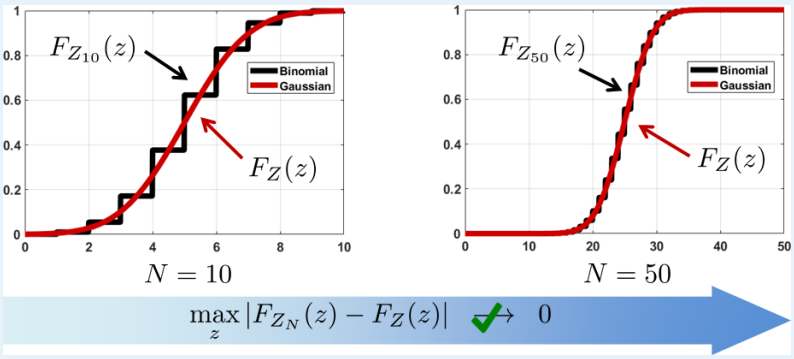


Figure 6.17: Convergence in distribution. This is the same as **Figure 6.16**, but this time we plot the CDF of Z_N. The CDF is a staircase function. We compare it with the Gaussian CDF. Observe that the error, $\max_z |F_{Z_N}(z) - F_Z(z)|$, converges to zero as N grows. Convergence in distribution says that the sequence of CDFs $F_{Z_N}(z)$ will converge to the limiting CDF $F_Z(z)$, at all continuous points of $F_Z(z)$.

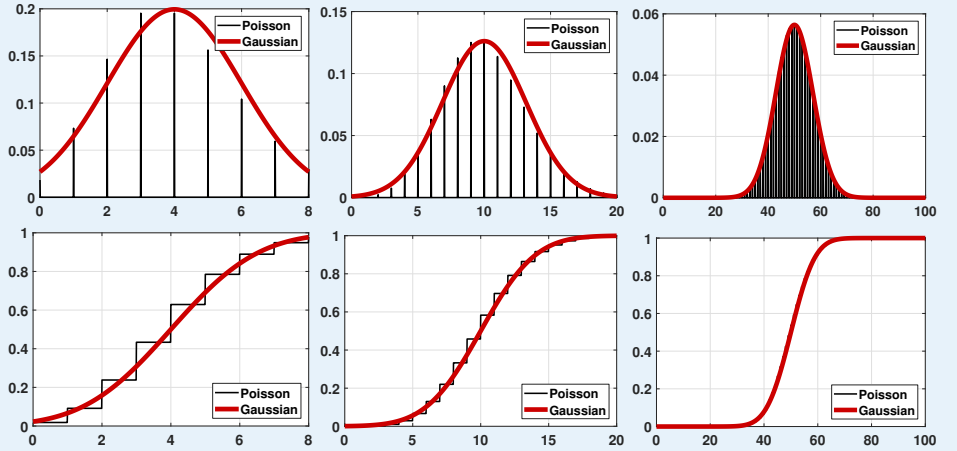
We need to pay attention to the set of z's. We do not evaluate all z's but only the z's that are continuous points of F_Z. If F_Z is Gaussian, this does not matter because all z's are continuous. However, for CDFs containing discontinuous points, our definition of convergence in distribution will ignore these discontinuous points because they have a measure zero.

Example 6.16

(**Poisson**) Consider $X_n \sim \text{Poisson}(\lambda)$, and consider $X_1, \ldots, X_N$. Define $Z_N = \sum_{n=1}^{N} X_n$. It follows that $\mathbb{E}[Z_N] = \sum_{n=1}^{N} \mathbb{E}[X_n] = N\lambda$ and $\text{Var}[Z_N] = \sum_{n=1}^{N} \text{Var}[X_n] = N\lambda$. Moreover, we know that the sum of Poissons remains a Poisson. Therefore, the PDF of Z_N is

$$f_{Z_N}(z) = \sum_{k=0}^{\infty} \frac{(N\lambda)^k}{k!} e^{-N\lambda} \delta(z-k) \qquad \text{and} \qquad f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(z-\mu)^2}{2\sigma^2} \right\},$$

where $\mu = N\lambda$ and $\sigma^2 = N\lambda$. Again, f_{Z_N} does not converge to f_Z. However, if we compare the CDF, we can see from **Figure 6.18** that the CDF of the Poisson is becoming better approximated by the Gaussian.



(a) $N = 4$ (b) $N = 10$ (c) $N = 50$

Figure 6.18: Convergence in distribution for a sum of Poisson random variables. Here we assume that $X_1, \ldots, X_N$ are i.i.d. Poisson with a parameter λ. We let $Z_N = \sum_{n=1}^{N} X_n$ be the sum, and compute the corresponding PDF (top row) and CDFs (bottom row). Just as with the binomial example, the PDFs of the Poisson do not converge but the CDFs of the Poisson converge to the CDF of a Gaussian.

Interpreting "convergence in distribution". After seeing two examples, you should now have some idea of what "convergence in distribution" means. This concept applies to the CDFs. When we write

$$\lim_{N\to\infty} F_{Z_N}(z) = F_Z(z), \tag{6.38}$$

we mean that $F_{Z_N}(z)$ is converging to the value $F_Z(z)$, and this relationship holds for all the continuous z's of F_Z. It does not say that the random variable Z_N is becoming another random variable Z.

$$Z_N \xrightarrow{d} Z \text{ is equivalent to } \lim_{N\to\infty} F_{Z_N}(z) = F_Z(z).$$

Example 6.17

(**Exponential**) So far, we have studied the sum of discrete random variables. Now, let's take a look at continuous random variables. Consider $X_n \sim \text{Exponential}(\lambda)$, and let $X_1, \ldots, X_N$ be i.i.d. copies. Define $Z_N = \sum_{n=1}^{N} X_n$. Then $\mathbb{E}[Z_N] = \sum_{n=1}^{N} \mathbb{E}[X_n] = N/\lambda$ and $\text{Var}[Z_N] = \frac{N}{\lambda^2}$. How about the PDF of Z_N? Using the characteristic functions, we know that

$$f_{X_n}(x) = \lambda e^{-\lambda x} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \Phi_{X_n}(j\omega) = \frac{\lambda}{\lambda + j\omega}.$$

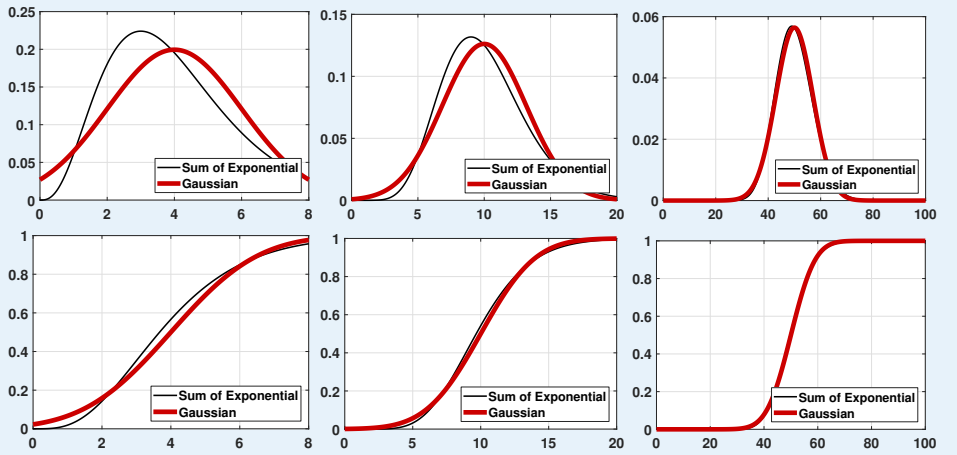
Therefore, the product is

$$\begin{aligned}\Phi_{Z_N}(j\omega) &= \prod_{n=1}^{N} \Phi_{X_n}(j\omega) = \frac{\lambda^N}{(\lambda + j\omega)^N} = \frac{\lambda^N}{(\lambda + j\omega)^N} \times \frac{(N-1)!}{(N-1)!} \\ &= \frac{\lambda^N}{(N-1)!} \cdot \frac{(N-1)!}{(\lambda + j\omega)^N} \overset{\mathcal{F}}{\longleftrightarrow} \frac{\lambda^N}{(N-1)!} z^{N-1} e^{-\lambda z} = f_{Z_N}(z).\end{aligned}$$

This resulting PDF $f_{Z_N}(z) = \frac{\lambda^N}{(N-1)!} z^{N-1} e^{-\lambda z}$ is known as the **Erlang distribution**. The CDF of the Erlang distribution is

$$\begin{aligned}F_{Z_N}(z) &= \int_{-\infty}^{z} f_{Z_N}(t)\, dt \\ &= \int_{0}^{z} \frac{\lambda^N}{(N-1)!} t^{N-1} e^{-\lambda t}\, dt \\ &= \text{Gamma function}(z, N),\end{aligned}$$

where the last integral is known as the incomplete gamma function, evaluated at z. Given all of these, we can now compare the PDF and the CDF of Z_N versus Z. **Figure 6.19** shows the PDFs and the CDFs of the *standardized* sum for various N values. In this experiment we set $\lambda = 1$. As we can see from the experiment, once standardized, the Erlang distribution's PDF *and* CDF both converge to those of a Gaussian. This is the difference from the discrete cases: there, the PDF is a train of deltas and can never converge pointwise to a smooth Gaussian PDF, so only the CDFs converge. We emphasize that even here it is the *distributions* that converge — the random variable Z_N itself does not converge to any particular random variable Z (see the counterexample later in this section).



(a) $N = 4$ (b) $N = 10$ (c) $N = 50$

Figure 6.19: Convergence in distribution for a sum of exponential random variables. Here we assume that $X_1, \ldots, X_N$ are i.i.d. exponentials with a parameter λ. We define $Z_N = \sum_{n=1}^{N} X_n$ to be the sum. It is known that the sum of exponentials is an Erlang. We compute the corresponding PDF (top row) and CDFs (bottom row). Unlike the previous two examples, in this example we see that both PDFs and CDFs of the Erlang distribution are converging to a Gaussian.

Is $\xrightarrow{d}$ stronger than $\xrightarrow{p}$? Convergence in distribution is actually weaker than convergence in probability. Consider a continuous random variable X with a symmetric PDF $f_X(x)$ such that $f_X(x) = f_X(-x)$. It holds that the PDF of $-X$ has the same PDF. If we define the sequence $Z_N = X$ if N is odd and $Z_N = -X$ if N is even, and let $Z = X$, then $F_{Z_N}(z) = F_Z(z)$ for every z because the PDFs of X and $-X$ are identical. Therefore, $Z_N \xrightarrow{d} Z$. However, $Z_N \not\xrightarrow{p} Z$ because Z_N oscillates between the random variables X and $-X$. These two random variables are different (although they have the same CDF) because $\mathbb{P}[X = -X] = \mathbb{P}[\{\omega : X(\omega) = -X(\omega)\}] = \mathbb{P}[\{\omega : X(\omega) = 0\}] = 0$.

6.4.2 Central Limit Theorem

Theorem 6.19 (Central Limit Theorem)

Let $X_1, \ldots, X_N$ be i.i.d. random variables of mean $\mathbb{E}[X_n] = \mu$ and *finite* variance $\mathrm{Var}[X_n] = \sigma^2$. (The classical theorem needs nothing more; the additional assumption $\mathbb{E}[|X_n|^3] < \infty$ is what gives the Berry-Esseen rate discussed below.) Let $\overline{X}_N = (1/N)\sum_{n=1}^{N} X_n$ be the sample average, and let $Z_N = \sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right)$. Then

$$\lim_{N\to\infty} F_{Z_N}(z) = F_Z(z), \tag{6.39}$$

where $Z = \text{Gaussian}(0, 1)$.

In plain words, the Central Limit Theorem says that the sample average (which is a random variable) has a CDF converging to the CDF of a Gaussian. Therefore, if we want to

evaluate probabilities associated with the sample average, we can approximate the probability by the probability of a Gaussian.

As we discussed above, the Central Limit Theorem does not mean that the random variable Z_N is converging to a Gaussian random variable, nor does it mean that the PDF of Z_N is converging to the PDF of a Gaussian. It only means that the CDF of Z_N is converging to the CDF of a Gaussian. Many people think that the Central Limit Theorem means "sample average converges to Gaussian". This is incorrect for the above reasons. However, it is not completely wrong. For continuous random variables where both PDF and CDF are continuous, we will not run into situations where the PDF is a train of delta functions. In this case, convergence in CDF can be translated to convergence in PDF.

The power of the Central Limit Theorem is that the result holds for **any** distribution of $X_1, \ldots, X_N$. That is, regardless of the distribution of $X_1, \ldots, X_N$, the CDF of $\overline{X}_N$ is approaching a Gaussian.

Summary of the Central Limit Theorem

- $X_1, \ldots, X_N$ are i.i.d. random variables, with mean μ and variance σ^2. They are not necessarily Gaussians.
- Define the sample average as $\overline{X}_N = (1/N)\sum_{n=1}^N X_n$, and let $Z_N = \sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right)$.
- The Central Limit Theorem says $Z_N \xrightarrow{d} \text{Gaussian}(0,1)$. Equivalently, the theorem says that $\sqrt{N}(\overline{X}_N - \mu) \xrightarrow{d} \text{Gaussian}(0, \sigma^2)$.
- So if we want to evaluate the probability of $\overline{X}_N \in \mathcal{A}$ for some set $\mathcal{A}$, we can approximate the probability by evaluating the Gaussian:

$$\mathbb{P}[\overline{X}_N \in \mathcal{A}] \approx \int_{\mathcal{A}} \frac{1}{\sqrt{2\pi(\sigma^2/N)}} \exp\left\{-\frac{(y-\mu)^2}{2(\sigma^2/N)}\right\} dy.$$

- CLT does **not** say that the PDF of $\overline{X}_N$ is becoming a Gaussian PDF.
- CLT only says that the CDF of $\overline{X}_N$ is becoming a Gaussian CDF.

If the set $\mathcal{A}$ is an interval, we can use the standard Gaussian CDF to compute the probability.

Corollary 6.3

Let $X_1, \ldots, X_N$ be i.i.d. random variables with mean μ and variance σ^2. Define the sample average as $\overline{X}_N = (1/N)\sum_{n=1}^N X_n$. Then

$$\mathbb{P}[a \le \overline{X}_N \le b] \approx \Phi\left(\sqrt{N}\frac{b-\mu}{\sigma}\right) - \Phi\left(\sqrt{N}\frac{a-\mu}{\sigma}\right), \tag{6.40}$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$ is the CDF of the standard Gaussian.

Proof

By the Central Limit Theorem, we know that $\overline{X}_N \xrightarrow{d} \text{Gaussian}(\mu, \frac{\sigma^2}{N})$. Therefore,

$$\begin{aligned}\mathbb{P}[a \le \overline{X}_N \le b] &\approx \int_a^b \frac{1}{\sqrt{2\pi(\sigma^2/N)}} \exp\left\{-\frac{(y-\mu)^2}{2(\sigma^2/N)}\right\} dy \\ &= \int_{\sqrt{N}\frac{a-\mu}{\sigma}}^{\sqrt{N}\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \Phi\left(\sqrt{N}\frac{b-\mu}{\sigma}\right) - \Phi\left(\sqrt{N}\frac{a-\mu}{\sigma}\right).\end{aligned}$$

A graphical illustration of the CLT is shown in **Figure 6.20**, where we use a binomial random variable (which is the sum of i.i.d. Bernoulli) as an example. The CLT does not say that the binomial random variable is becoming a Gaussian. It only says that the probability covered by the binomial can be approximated by the Gaussian.

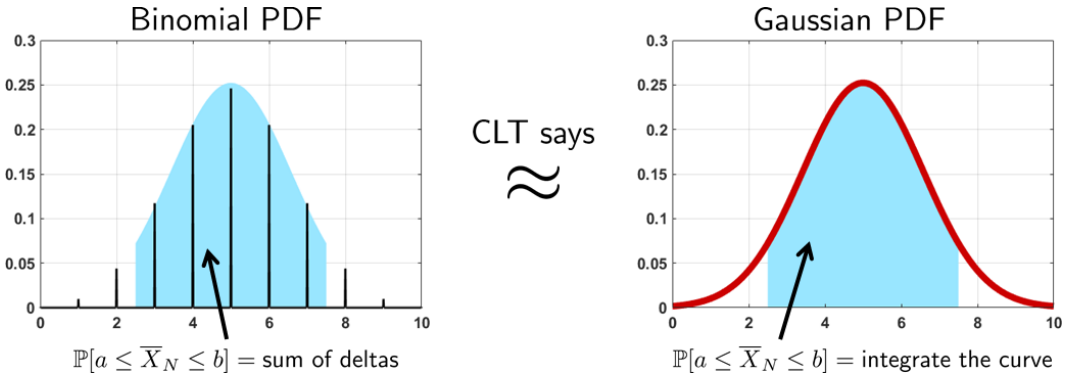


Figure 6.20: The Central Limit Theorem says that if we want to evaluate the probability $\mathbb{P}[a \le \overline{X}_N \le b]$, where $\overline{X}_N = (1/N)\sum_{n=1}^N X_n$ is the sample average of i.i.d. random variables $X_1, \ldots, X_N$, we can approximate the probability by integrating the Gaussian PDF.

The following proof of the Central Limit Theorem can be skipped if this is your first time reading the book.

Proof of the Central Limit Theorem

We now give a "proof" of the Central Limit Theorem. Technically speaking, this proof does not prove the convergence of the CDF as the theorem claims; it only proves that the moment-generating function converges. The actual proof of the CDF convergence is based on the Berry-Esseen Theorem, which is beyond the scope of this book. However, what we prove below is still useful because it gives us some intuition about why Gaussian is the limiting random variable we should consider in the first place.

Let $Z_N = \sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right)$. It follows that $\mathbb{E}[Z_N] = 0$ and $\text{Var}[Z_N] = 1$. Therefore, if we can show that Z_N is converging to a standard Gaussian random variable $Z \sim \text{Gaussian}(0,1)$, then by the linear transformation property of Gaussian, $Y = \frac{\sigma}{\sqrt{N}}Z + \mu$ will be $\text{Gaussian}(\mu, \sigma^2/N)$.

Our proof is based on analyzing the moment-generating function of Z_N. In particular,

$$M_{Z_N}(s) \stackrel{\text{def}}{=} \mathbb{E}[e^{sZ_N}] = \mathbb{E}\left[e^{s\sqrt{N}\left(\frac{\overline{X}_N-\mu}{\sigma}\right)}\right] = \prod_{n=1}^{N} \mathbb{E}\left[e^{\frac{s}{\sigma\sqrt{N}}(X_n-\mu)}\right].$$

Expanding the exponential term using the Taylor expansion (Chapter 1.2),

$$\begin{aligned}
&\prod_{n=1}^{N} \mathbb{E}\left[e^{\frac{s}{\sigma\sqrt{N}}(X_n-\mu)}\right] \\
&= \prod_{n=1}^{N} \mathbb{E}\left[1 + \frac{s}{\sigma\sqrt{N}}(X_n-\mu) + \frac{s^2}{2\sigma^2 N}(X_n-\mu)^2 + \mathcal{O}\left(\frac{(X_n-\mu)^3}{\sigma^3 N\sqrt{N}}\right)\right] \\
&= \prod_{n=1}^{N} \left[1 + \frac{s}{\sigma\sqrt{N}}\mathbb{E}[X_n-\mu] + \frac{s^2}{2\sigma^2 N}\mathbb{E}\left[(X_n-\mu)^2\right]\right] = \left(1+\frac{s^2}{2N}\right)^N.
\end{aligned}$$

It remains to show that $\left(1+\frac{s^2}{2N}\right)^N \to e^{s^2/2}$. If we can show that, we have shown that the MGF of Z_N is also the MGF of Gaussian$(0,1)$. To this end, we consider $\log(1+x)$. By the Taylor approximation, we have that

$$\log(1+x) \approx \log(1) + \left(\frac{d}{dx}\log x|_{x=1}\right)x + \left(\frac{d^2}{dx^2}\log x|_{x=1}\right)\frac{x^2}{2} + \mathcal{O}(x^3).$$

Therefore, we have $\log\left(1+\frac{s^2}{2N}\right) \approx \frac{s^2}{2N} - \frac{s^4}{8N^2}$. As $N \to \infty$, the limit becomes

$$\lim_{N\to\infty} N\log\left(1+\frac{s^2}{2N}\right) \approx \frac{s^2}{2} - \lim_{N\to\infty}\frac{s^4}{8N} = \frac{s^2}{2},$$

and so taking the exponential on both sides yields $\lim_{N\to\infty}\left(1+\frac{s^2}{2N}\right)^N = e^{\frac{s^2}{2}}$. Therefore, we conclude that $\lim_{N\to\infty} M_{Z_N}(s) = e^{\frac{s^2}{2}}$, and so Z_N is converging to a Gaussian.

Limitation of our proof. The limitation of our proof lies in the issue of whether the integration and the limit are interchangeable:

$$\begin{aligned}
\lim_{N\to\infty} M_{Z_N}(s) &= \lim_{N\to\infty}\left\{\int f_{Z_N}(z)e^{sz}\,dz\right\} \\
&\stackrel{?}{=} \int\left(\lim_{N\to\infty} f_{Z_N}(z)\right)e^{sz}\,dz.
\end{aligned}$$

If they were, then proving $\lim_{N\to\infty} M_{Z_N}(s) = M_Z(s)$ is sufficient to claim $f_{Z_N}(z) \to f_Z(z)$. However, we know that the latter is not true in general. For example, if $f_{Z_N}(z)$ is a train of delta functions, then the limit and the integration are not interchangeable.

Berry-Esseen Theorem. The formal way of proving the Central Limit Theorem is to

prove the Berry-Esseen Theorem. The theorem states that

$$\sup_{z\in\mathbb{R}} \left| F_{Z_N}(z) - F_Z(z) \right| \leq C\frac{\beta}{\sigma^3\sqrt{N}},$$

where $\beta = \mathbb{E}[|X_n - \mu|^3]$ is the third absolute central moment of the underlying distribution and C is a *universal* constant (independent of the distribution). Here, you can more or less treat the supremum operator as the maximum. The left-hand side represents the **worst-case** error of the CDF F_{Z_N} compared to the limiting CDF F_Z. On the right-hand side, C, β, and σ do not depend on N.

As N goes to infinity, the right-hand side will converge to zero. Therefore, if we can prove this result, then we have proved the actual Central Limit Theorem. In addition, we have found the rate of convergence since the right-hand side tells us that the error drops at the rate of $1/\sqrt{N}$, which is not particularly fast but is sufficient for our purpose. Unfortunately, proving the Berry-Esseen theorem is not easy. One of the difficulties, for example, is that one needs to deal with the infinite convolutions in the time domain or the frequency domain.

Interpreting our proof. If our proof is not completely valid, why do we mention it? For one thing, it provides us with some useful intuition. For most of the (well-behaving) random variables whose moments are finite, the exponential term in the moment-generating function can be truncated to the second-order polynomial. Since a second-order polynomial is a Gaussian, it naturally concludes that as long as we can perform such truncation the truncated random variable will be Gaussian.

To convince you that the Gaussian MGF is the second-order approximation to other MGFs, we use Bernoulli as an example. Let $X_1, \ldots, X_N$ be i.i.d. Bernoulli with a parameter p. Then the moment-generating function of $\overline{X}_N = (1/N)\sum_{n=1}^N X_n$ would be:

$$\begin{aligned} M_{\overline{X}_N}(s) &= \mathbb{E}[e^{s\overline{X}_N}] = \mathbb{E}[e^{s\frac{1}{N}\sum_{n=1}^N X_n}] = \prod_{n=1}^N \mathbb{E}[e^{\frac{s}{N}X_n}] \\ &= (1-p+pe^{\frac{s}{N}})^N \approx \left(1-p+p\left(1+\frac{s}{N}+\frac{s^2}{2N^2}\right)\right)^N \\ &= \left(1+\frac{sp}{N}+\frac{s^2p}{2N^2}\right)^N. \end{aligned}$$

Using the logarithmic approximation, it follows that

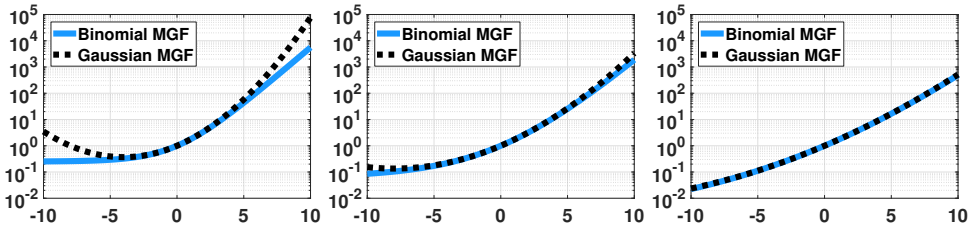
$$\begin{aligned} \log M_{\overline{X}_N}(s) &= N\log\left(1+\frac{sp}{N}+\frac{s^2p}{2N^2}\right) \\ &\approx N\left(\frac{sp}{N}+\frac{s^2p}{2N^2}\right) - \frac{N}{2}\left(\frac{sp}{N}+\frac{s^2p}{2N^2}\right)^2 \\ &\approx sp + \frac{s^2p(1-p)}{2N} \stackrel{\text{def}}{=} \log M_Y(s). \end{aligned}$$

Taking the exponential on both sides, we have that

$$M_Y(s) = \exp\left\{sp + \frac{s^2p(1-p)}{2N}\right\},$$

which is the MGF of a Gaussian random variable $Y \sim \text{Gaussian}\left(p, \frac{p(1-p)}{N}\right)$.

Figure 6.21 shows several MGFs. In each of the subfigures we plot the exact MGF $M_{\overline{X}_N}(s) = (1 - p + pe^{\frac{s}{N}})^N$ as a function of s. (The parameter p in this example is $p = 0.5$.) We vary the number N, and we inspect how the shape of $M_{\overline{X}_N}(s)$ changes. On top of the exact MGFs, we plot the Gaussian approximations $M_Y(s) = \exp\left\{sp + \frac{s^2 p(1-p)}{2N}\right\}$. According to our calculation, this Gaussian approximation is the second-order approximation to the exact MGF. The figures show the effect of the second-order approximation. For example, in (a) when $N = 2$ the Gaussian is a quadratic approximation of the exact MGF. For (b) and (c), as N increases, the approximation improves.



(a) $N = 2$ (b) $N = 4$ (c) $N = 10$

Figure 6.21: Explanation of the Central Limit Theorem using the moment-generating function. In this set of plots, we show the MGF of the random variable $\overline{X}_N = (1/N)\sum_{n=1}^{N} X_n$, where $X_1, \ldots, X_N$ are i.i.d. Bernoulli random variables. The exact MGF of $\overline{X}_N$ is the binomial, whereas the approximated MGF is the Gaussian. We observe that as N increases, the Gaussian approximation to the exact MGF improves.

The reason why the second-order approximation works for Gaussian is that when N increases, the higher order moments of $\overline{X}_N$ vanish and only the leading first two moments survive. The MGFs are becoming flat because $M_Y(s) = \exp\left\{sp + \frac{s^2 p(1-p)}{2N}\right\}$ converges to $\exp\{sp\}$ when $N \to \infty$. Taking the inverse Laplace transform, $M_Y(s) = \exp\{sp\}$ corresponds to a delta function. This makes sense because as N grows, the variance of $\overline{X}_N$ shrinks.

End of the discussion.

6.4.3 Examples

Example 6.18

Prove the equivalence of two statements.

- $\sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right) \xrightarrow{d} \text{Gaussian}(0, 1)$
- $\sqrt{N}(\overline{X}_N - \mu) \xrightarrow{d} \text{Gaussian}(0, \sigma^2)$

Solution. The proof is based on the linear transformation property of Gaussian random variables. For example, if the first statement is true, then the second statement

is also true because

$$\begin{aligned}\lim_{N\to\infty} F_{\sqrt{N}(\overline{X}_N-\mu)}(z) &= \lim_{N\to\infty} \mathbb{P}[\sqrt{N}(\overline{X}_N-\mu)\le z] \\ &= \lim_{N\to\infty} \mathbb{P}\left[\sqrt{N}\left(\frac{\overline{X}_N-\mu}{\sigma}\right)\le \frac{z}{\sigma}\right] \\ &= \int_{-\infty}^{z/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}}\,dt.\end{aligned}$$

The other results can be proved similarly.

Example 6.19

Suppose $X_n \sim \text{Poisson}(10)$ for $n = 1,\dots,N$, and let $\overline{X}_N$ be the sample average. Use the Central Limit Theorem to approximate $\mathbb{P}[9 \le \overline{X}_N \le 11]$ for $N = 20$.

Solution. We first show that

$$\begin{aligned}\mathbb{E}[\overline{X}_N] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[X_n\right] = 10, \\ \text{Var}[\overline{X}_N] &= \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}\left[X_n\right] = \frac{1}{N}\text{Var}[X_n] = \frac{10}{20} = \frac{1}{2}.\end{aligned}$$

Therefore, the Central Limit Theorem implies that $\overline{X}_N \xrightarrow{d} \text{Gaussian}\left(10, \frac{1}{2}\right)$. The probability is

$$\begin{aligned}\mathbb{P}[9 \le \overline{X}_N \le 11] &\approx \Phi\left(\frac{11-10}{\sqrt{1/2}}\right) - \Phi\left(\frac{9-10}{\sqrt{1/2}}\right) \\ &= \Phi\left(\frac{1}{\sqrt{0.5}}\right) - \Phi\left(-\frac{1}{\sqrt{0.5}}\right) = 0.9214 - 0.0786 = 0.8427.\end{aligned}$$

We can also do an exact calculation to verify our approximation. Let $S_N = \sum_{n=1}^{N} X_n$ so that $\overline{X}_N = \frac{S_N}{N}$. Since a sum of Poisson remains a Poisson, it follows that

$$S_N \sim \text{Poisson}(10N) = \text{Poisson}(200).$$

Consequently,

$$\begin{aligned}\mathbb{P}[9 \le \overline{X}_N \le 11] &= \mathbb{P}[180 \le S_N \le 220] \\ &= \sum_{\ell=0}^{220} \frac{200^\ell e^{-200}}{\ell!} - \sum_{\ell=0}^{179} \frac{200^\ell e^{-200}}{\ell!} = 0.9247 - 0.0822 = 0.8425.\end{aligned}$$

Note that this is an exact calculation subject to numerical errors when evaluating the finite sums.

Example 6.20

Suppose you have collected $N = 100$ data points from an unknown distribution. The only thing you know is that the true population mean is $\mu = 500$ and the standard deviation is $\sigma = 80$. (Note that this distribution is not necessarily a Gaussian.)

(a) Find the probability that the sample mean will be inside the interval $(490, 510)$.

(b) Find an interval such that 95% of the sample average is covered.

Solution. To solve (a), we note that $\overline{X}_N \xrightarrow{d}$ Gaussian $\left(500, \left(\frac{80}{\sqrt{100}}\right)^2\right)$. Therefore,

$$\begin{aligned}\mathbb{P}[490 \leq \overline{X}_N \leq 510] &= \Phi\left(\frac{510-500}{\frac{80}{\sqrt{100}}}\right) - \Phi\left(\frac{490-500}{\frac{80}{\sqrt{100}}}\right) \\ &= \Phi(1.25) - \Phi(-1.25) = 0.7888.\end{aligned}$$

To solve (b), we know that $\Phi(x) = 0.025$ implies that $x = -1.96$, and $\Phi(x) = 0.975$ implies that $x = +1.96$. So

$$\frac{y-500}{\frac{80}{\sqrt{100}}} = \pm 1.96 \quad \Rightarrow \quad y = 484.32 \quad \text{or} \quad y = 515.68.$$

Therefore, $\mathbb{P}[484.32 \leq \overline{X}_N \leq 515.68] = 0.95$.

6.4.4 Limitation of the Central Limit Theorem

If we recall the statement of the Central Limit Theorem (Berry-Esseen), we observe that the theorem states only that

$$\lim_{N\to\infty} \mathbb{P}\left[\sqrt{N}\left(\frac{\overline{X}_N - \mu}{\sigma}\right) \leq \varepsilon\right] = \lim_{N\to\infty} F_{Z_N}(\varepsilon) = F_Z(\varepsilon) = \Phi(\varepsilon).$$

Rearranging the terms,

$$\lim_{N\to\infty} \mathbb{P}\left[\overline{X}_N \leq \mu + \frac{\sigma\varepsilon}{\sqrt{N}}\right] = \Phi(\varepsilon).$$

This implies that the approximation is good only when the deviation ε is *small*.

Let us consider an example to illustrate this idea. Consider a set of i.i.d. exponential random variables $X_1, \ldots, X_N$, where $X_n \sim$ Exponential(λ). Let $S_N = X_1 + \cdots + X_N$ be the sum, and let $\overline{X}_N = S_N/N$ be the sample average. Then, according to Chapter 6.4.1, S_N is an Erlang distribution $S_N \sim$ Erlang(N, λ) with a PDF

$$f_{S_N}(x) = \frac{\lambda^N}{(N-1)!} x^{N-1} e^{-\lambda x}.$$

Practice Exercise 6.10

Let $S_N \sim$ Erlang(N, λ) with a PDF $f_{S_N}(x)$. Show that if $Y_N = aS_N + b$ for any con-

stants a and b, then

$$f_{Y_N}(y) = \frac{1}{a} f_{S_N}\left(\frac{y-b}{a}\right).$$

Solution: This is a simple transformation of random variables:

$$F_{Y_N}(y) = \mathbb{P}[Y \le y] = \mathbb{P}[aS_N + b \le y] = \mathbb{P}\left[S_N \le \frac{y-b}{a}\right] = \int_{-\infty}^{\frac{y-b}{a}} f_{S_N}(x)\, dx.$$

Hence, using the fundamental theorem of calculus,

$$f_{Y_N}(y) = \frac{d}{dy}\int_{-\infty}^{\frac{y-b}{a}} f_{S_N}(x)\, dx = \frac{1}{a} f_{S_N}\left(\frac{y-b}{a}\right).$$

We are interested in knowing the statistics of $\overline{X}_N$ and comparing it with a Gaussian. To this end, we construct a normalized variable

$$Z_N = \frac{\overline{X}_N - \mu}{\sigma/\sqrt{N}},$$

where $\mu = \mathbb{E}[X_n] = \frac{1}{\lambda}$ and $\sigma^2 = \mathrm{Var}[X_n] = \frac{1}{\lambda^2}$. Then

$$Z_N = \frac{S_N/N - \mu}{\sigma/\sqrt{N}} = \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\lambda}{\sqrt{N}} S_N - \sqrt{N}$$

Using the result of the practice exercise, by mapping $a = \frac{\lambda}{\sqrt{N}}$ and $b = -\sqrt{N}$, it follows that

$$f_{Z_N}(z) = \frac{\sqrt{N}}{\lambda} f_{S_N}\left(\frac{z+\sqrt{N}}{\frac{\lambda}{\sqrt{N}}}\right).$$

Now we compare Z_N with the standard Gaussian $Z \sim \text{Gaussian}(0,1)$. According to the Central Limit Theorem, the standard Gaussian is a good approximation to the normalized sample average Z_N. To compare the two results, we conduct a numerical experiment. We let $\lambda = 1$ and we vary N. We plot the PDF $f_{Z_N}(z)$ as a function of z, for different N's, in **Figure 6.22**. In addition, we plot the PDF $f_Z(z)$, which is the standard Gaussian.

The plot in **Figure 6.22** shows that while the Central Limit Theorem provides a good approximation, the approximation is only good for values that are close to the mean. For the **tails**, the Gaussian approximation is not as good.

The limitation of the Central Limit Theorem is attributable to the fact that Gaussian is a second-order approximation. If a random variable has a very large third moment, the second-order approximation may not be sufficient. In this case, we need a much larger N to drive the third moment to a small value and make the Gaussian approximation valid.

When will the Central Limit Theorem fail?

- The Central Limit Theorem fails when N is small.
- The Central Limit Theorem needs a finite **variance**. If the variance is infinite —

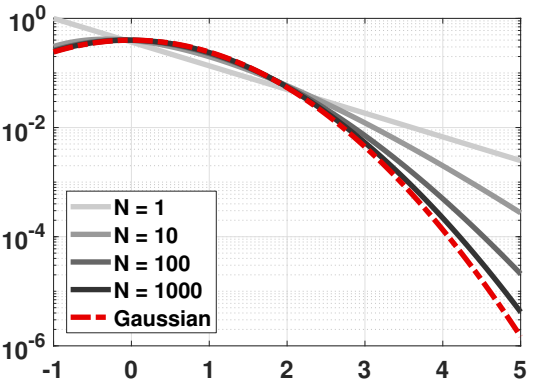


Figure 6.22: CLT fails at the tails. We note that $X_1, \ldots, X_N$ are i.i.d. exponential with a parameter $\lambda = 1$. We plot the PDFs of the normalized sample average $Z_N = \frac{\overline{X}_N - \mu}{\sigma/\sqrt{N}}$ by varying N. We plot the PDF of the standard Gaussian $Z \sim \text{Gaussian}(0,1)$ on the same grid. Note that the Gaussian approximation is good for values that are close to the mean. For the tails, the Gaussian approximation is not very accurate.

as for a Cauchy random variable — the theorem does not apply at all. (Recall from Section 6.3 that even the law of large numbers fails for the Cauchy.)

- Even when the variance is finite, the approximation converges *slowly* if the third absolute moment β is large relative to σ^3, because the Berry-Esseen bound scales as $\beta/(\sigma^3\sqrt{N})$. A large β therefore means you need a much larger N.
- The Central Limit Theorem can only approximate the probability for input values near the mean. It does not approximate the tails, for which we need to use Chernoff's bound.

6.5 Summary

Why do we need to study the sample average? Because it is the summary of the dataset. In machine learning, one of the most frequently asked questions is about the number of training samples required to train a model. The answer can be found by analyzing the average number of successes and failures as the number of training samples grows. For example, if we define f as the classifier that takes a data point $\boldsymbol{x}_n$ and predicts a label $f(\boldsymbol{x}_n)$, we hope that it will match with the true label y_n. If we define an error

$$E_n = \begin{cases} 1, & f(\boldsymbol{x}_n) \neq y_n \quad \text{incorrect classification,} \\ 0, & f(\boldsymbol{x}_n) = y_n \quad \text{correct classification,} \end{cases}$$

then E_n is a Bernoulli random variable, and the total loss $\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N} E_n$ will be the training loss. But what is $\frac{1}{N}\sum_{n=1}^{N} E_n$? It is exactly the sample average of E_n. Therefore, by analyzing the sample average $\mathcal{E}$ we will learn something about the generalization capability of our model.

How should we study the sample average? By understanding the law of large numbers and the Central Limit Theorem, as we have seen in this chapter.

- Law of large numbers: $\overline{X}$ converges to the true mean μ as N grows.
- Central Limit Theorem: The CDF of $\overline{X}$ can be approximated by the CDF of a Gaussian, as N grows.

Performance guarantee? The other topic we discussed in this chapter is the concept of convergence type. There are essentially four types of convergence, ranked in the order of restrictions.

- **Deterministic convergence**: A sequence of deterministic numbers converges to another deterministic number. For example, the sequence $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ converges to 0 deterministically. There is nothing random about it.
- **Almost sure convergence**: Randomness exists, but with probability 1 the whole sequence of random variables converges, in the ordinary calculus sense, to the limit. Equivalently, for any tolerance $\epsilon > 0$ only finitely many terms fall outside the band $[\alpha - \epsilon, \alpha + \epsilon]$.
- **Convergence in probability**: The sequence of probability values converges, i.e., the chance of failure is going to zero. However, you can still fail even if your N is large.
- **Convergence in distribution**: Only the values of the CDF converge, i.e., $F_{Z_N}(z) \to F_Z(z)$ at every continuity point of the limiting CDF F_Z. The limiting distribution need not be Gaussian in general; in the Central Limit Theorem it happens to be.

6.6 References

Moment-Generating and Characteristic Functions

6-1 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 4.4.

6-2 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapters 4.5 and 4.7.

6-3 Athanasios Papoulis and Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapters 5.5 and 7.2.

6-4 Henry Stark and John Woods, *Probability and Random Processes With Applications to Signal Processing*, Prentice Hall, 3rd Edition, 2001. Chapters 4.5 and 4.7.

6-5 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 7.7.

6-6 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapter 4.3.

Basic probability inequality

6-7 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 5.1.

6-8 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapters 6 and 8.

6-9 Athanasios Papoulis and Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 7.4.

6-10 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 8.2.

6-11 Larry Wasserman, *All of Statistics*, Springer 2003. Chapter 4.

Concentration inequalities

6-12 Larry Wasserman, *All of Statistics*, Springer 2003. Chapter 4.

6-13 Martin Wainwright, *High-Dimensional Statistics*, Cambridge University Press, 2019. Chapter 2.1.

6-14 Stephane Boucheron, Gabor Lugosi and Pascal Massart, *Concentration Inequalities*, Oxford University Press, 2013. Chapters 2.1 and 2.2.

Law of large numbers

6-15 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapters 5.2, 5.3, 5.5.

6-16 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapters 7.1, 7.2, 7.4

6-17 Athanasios Papoulis and Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 7.4.

6-18 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 8.2, 8.4.

6-19 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapters 3.3, 14.1, 14.3.

6-20 Larry Wasserman, *All of Statistics*, Springer 2003. Chapter 5.1 - 5.3.

6-21 Patrick Billingsley, *Probability and Measure*, Wiley 1995. Section 22.

Central Limit Theorem

6-22 Dimitri P. Bertsekas and John N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2nd Edition, 2008. Chapter 5.4.

6-23 Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*, Prentice Hall, 3rd Edition, 2008. Chapter 7.3.

6-24 Athanasios Papoulis and Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, 4th Edition, 2001. Chapter 7.4.

6-25 Sheldon Ross, *A First Course in Probability*, Prentice Hall, 8th Edition, 2010. Chapter 8.3.

6-26 John A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, 2006. Chapters 5.6, 14.2.

6-27 Larry Wasserman, *All of Statistics*, Springer 2003. Chapter 5.4.

6-28 Patrick Billingsley, *Probability and Measure*, Wiley 1995. Section 27.

6.7 Exercises

Exercise 1

(True or False.) For each statement, indicate True or False and give a one-sentence explanation.

(a) The moment-generating function $M_X(t)$ uniquely determines the distribution of X (when it exists in a neighborhood of 0).

(b) $M_X(0) = 1$ for any random variable X.

(c) The characteristic function $\varphi_X(\omega) = \mathbb{E}[e^{j\omega X}]$ always exists for any distribution.

(d) If $X \sim \text{Gaussian}(0, 1)$, then $M_X(t) = e^{t^2/2}$.

(e) Markov's inequality requires X to have a symmetric distribution.

(f) Chebyshev's inequality gives a bound of the form $\mathbb{P}[|X - \mu| > \varepsilon] \leq \sigma^2/\varepsilon^2$, which requires no distributional assumption beyond finite variance.

(g) The central limit theorem applies only when the underlying distribution is Gaussian.

(h) If $M_X(t) = e^{\lambda(e^t - 1)}$, then $X \sim \text{Poisson}(\lambda)$.

(i) Hoeffding's inequality applies to bounded random variables without requiring knowledge of their distribution.

(j) Convergence in probability implies convergence almost surely.

Exercise 2

(Short Answer.) Answer each of the following briefly.

(a) Define the MGF and state how to extract the n-th moment from it.

(b) State Markov's inequality.

(c) State Chebyshev's inequality.

(d) If $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, what does the CLT say about $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$?

(e) What is the difference between convergence in probability and convergence in distribution?

(f) State the Weak Law of Large Numbers.

(g) If X and Y are independent, how is $M_{X+Y}(t)$ related to $M_X(t)$ and $M_Y(t)$?

(h) Describe the Chernoff bounding technique in one or two sentences.

(i) State Hoeffding's inequality for the sample mean of bounded i.i.d. random variables.

(j) Define the characteristic function $\Phi_X(j\omega)$, and explain why it is preferred over the MGF $M_X(s)$ for some distributions.

Exercise 3

Let X, Y, Z be three independent random variables:

$$X \sim \text{Bernoulli}(p), \qquad Y \sim \text{Exponential}(\alpha), \qquad Z \sim \text{Poisson}(\lambda)$$

Find the moment-generating function for the following random variables.

(a) $U = Y + Z$

(b) $U = 2Z + 3$

(c) $U = XY$

(d) $U = 2XY + (1 - X)Z$

Exercise 4

Two random variables X and Y have the joint PMF

$$\mathbb{P}(X = n, Y = m) = \frac{\lambda_1^{n+m}\lambda_2^m}{(n+m)!m!}e^{-(\lambda_1+\lambda_2)}, \qquad m = 0, 1, 2, \ldots, \;\; n \geq -m.$$

Let $Z = X + Y$. Find the moment-generating function $M_Z(s)$ and the PMF of Z.

Exercise 5

Let $X_0, X_1, \ldots$ be a sequence of independent random variables with PDF

$$f_{X_k}(x) = \frac{a_k}{\pi(a_k^2 + x^2)}, \qquad a_k = \frac{1}{2^{k+1}},$$

for $k = 0, 1, \ldots$. Find the PDF of Y, where

$$Y = \sum_{k=0}^{\infty} X_k.$$

Hint: You may find the characteristic function useful.

Exercise 6

The random variables X and Y are independent and have PDFs

$$f_X(x) = \begin{cases} e^{-x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad \text{and} \qquad f_Y(y) = \begin{cases} 0, & y > 0, \\ e^{y}, & y \leq 0. \end{cases}$$

Find the PDF of $Z = X + Y$. (Hint: Use the characteristic function and the moment-generating function.)

Exercise 7

A discrete random variable X has a PMF

$$p_X(k) = \frac{1}{2^k}, \qquad k = 1, 2, \ldots.$$

Find the characteristic function $\Phi_X(j\omega)$.

Exercise 8

Let $T_1, T_2, \ldots$ be i.i.d. random variables with PDF

$$f_{T_k}(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0, \end{cases}$$

for $k = 1, 2, 3, \ldots$. Let $S_n = \sum_{k=1}^{n} T_k$. Find the PDF of S_n.

Exercise 9

Chapter 6.2.3 states Jensen's inequality: if g is convex then $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$, and the inequality reverses for concave g.

(a) Let $X > 0$ be a positive random variable. Show that $\mathbb{E}\left[\dfrac{1}{X}\right] \geq \dfrac{1}{\mathbb{E}[X]}$, and state when equality holds. Verify your answer for X uniform on $\{1, 2\}$.

(b) Show that $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$ for $X > 0$. Then, by choosing X uniform on a finite set of positive numbers $a_1, \ldots, a_n$, deduce the arithmetic-mean/geometric-mean inequality

$$(a_1 a_2 \cdots a_n)^{1/n} \;\leq\; \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

(c) An investor's wealth is multiplied each round by an independent copy of a positive random variable R, so that $W_n = W_0 R_1 R_2 \cdots R_n$. Using the strong law of large numbers of Chapter 6.3.7, show that

$$\frac{1}{n} \log \frac{W_n}{W_0} \;\longrightarrow\; \mathbb{E}[\log R] \quad \text{almost surely,}$$

so that the long-run growth rate is governed by $\mathbb{E}[\log R]$, not by $\mathbb{E}[R]$.

(d) Suppose $R = 2$ with probability $1/2$ and $R = 0.4$ with probability $1/2$. Compute $\mathbb{E}[R]$ and $\mathbb{E}[\log R]$. What happens to W_n as $n \to \infty$, and why is this consistent with Jensen's inequality?

Exercise 10

In this exercise we will prove a variant of Chebyshev when the variance σ^2 is unknown but X is bounded between $a \leq X \leq b$.

(a) Let $\gamma \in \mathbb{R}$. Find a γ that minimizes $\mathbb{E}[(X-\gamma)^2]$. Hence, show that $\mathbb{E}[(X-\gamma)^2] \geq \mathrm{Var}[X]$ for any γ.

(b) Let $\gamma = (a+b)/2$. Show that

$$\mathbb{E}[(X-\gamma)^2] = \mathbb{E}[(X-a)(X-b)] + \frac{(b-a)^2}{4}.$$

(c) From (a) and (b), show that $\mathrm{Var}[X] \leq \frac{(b-a)^2}{4}$.

(d) Show that for any $\varepsilon > 0$,

$$\mathbb{P}[|X-\mu| > \varepsilon] \leq \frac{(b-a)^2}{4\varepsilon^2}.$$

Exercise 11

The random variables X and Y are independent with PDFs

$$f_X(x) = \frac{1}{\pi(1+x^2)} \qquad \text{and} \qquad f_Y(y) = \frac{1}{\pi(1+y^2)},$$

respectively. Find the PDF of $Z = X - Y$. (Hint: Use the characteristic function.)

Exercise 12

A random variable X has the characteristic function

$$\Phi_X(j\omega) = e^{j\omega/(1-j\omega)}.$$

Find the mean and variance of X.

Exercise 13

Show that for any random variables X and Y,

$$\mathbb{P}[|X-Y| > \epsilon] \leq \frac{1}{\epsilon^2}\mathbb{E}[(X-Y)^2].$$

Exercise 14

Let X be an exponential random variable with a parameter λ. Let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}[X]$. Compute $\mathbb{P}[|X - \mu| \geq k\sigma]$ for any $k > 1$. Compare this to the bound obtained by Chebyshev's inequality.

Exercise 15

Let $X_1, \ldots, X_N$ be i.i.d. Bernoulli with a parameter p. Let $\alpha > 0$ and define

$$\epsilon = \sqrt{\frac{1}{2N} \log\left(\frac{2}{\alpha}\right)}.$$

Let $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$. Define an interval

$$\mathcal{I} = \left[\overline{X}_N - \epsilon,\ \ \overline{X}_N + \epsilon\right].$$

Use Hoeffding's inequality to show that

$$\mathbb{P}[\mathcal{I} \ \text{ contains } \ p] \geq 1 - \alpha.$$

Exercise 16

Let $Z \sim \text{Gaussian}(0, 1)$. Prove that for any $\epsilon > 0$,

$$\mathbb{P}[|Z| > \epsilon] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{\epsilon^2}{2}}}{\epsilon}.$$

Hint: Note that $\epsilon\mathbb{P}[|Z| > \epsilon] = 2\epsilon\mathbb{P}[Z > \epsilon]$, and then follow the procedure we used to prove Markov's inequality.

Exercise 17

(a) Give a non-negative random variable $X \geq 0$ such that Markov's inequality is met with equality. Hint: Consider a discrete random variable.

(b) Give a random variable X such that Chebyshev's inequality is met with equality.

Exercise 18

Consider a random variable X such that

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2\sigma^2}{2}} \qquad \text{for every real } s.$$

(a) Show that for any $t > 0$,

$$\mathbb{P}[X \geq t] \leq \exp\left\{-\frac{t^2}{2\sigma^2}\right\}.$$

Hint: Use Chernoff's bound.

(b) Show that

$$\mathbb{E}[X^2] \leq 4\sigma^2.$$

Hint: First prove that $\mathbb{E}[X^2] = \int_0^\infty \mathbb{P}[X^2 \geq t]\, dt$. Then use part (a) above. Note that the assumption holds for every real s, so part (a) applies to $-X$ as well as to X; this gives you a two-sided tail bound on $\mathbb{P}[|X| \geq \sqrt{t}\,]$.

Exercise 19

Let $X_1, \ldots, X_N$ be i.i.d. uniform random variables distributed over $[0, 1]$. Suppose $Y_1, \ldots, Y_N$ are defined as follows.

(a) $Y_n = X_n / n$

(b) $Y_n = (X_n)^n$

(c) $Y_n = \max(X_1, \ldots, X_n)$

(d) $Y_n = \min(X_1, \ldots, X_n)$

For (a), (b), (c), and (d), show that Y_n converges in probability to some limit. Identify the limit in each case.

Exercise 20

Let $\lambda_n = \frac{1}{n}$ for $n = 1, 2, \ldots$. Let $X_n \sim \text{Poisson}(\lambda_n)$. Show that X_n converges in probability to 0.

Exercise 21

Let $Y_1, Y_2, \ldots$ be a sequence of random variables such that

$$Y_n = \begin{cases} 0, & \text{with probability } 1 - \frac{1}{n}, \\ 2^n, & \text{with probability } \frac{1}{n}. \end{cases}$$

Does Y_n converge in probability to 0?

Exercise 22

A Laplace random variable has a PDF

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda |x|}, \quad \lambda > 0,$$

and the variance is $\text{Var}[X] = \frac{2}{\lambda^2}$. Let $X_1, \ldots, X_{500}$ be a sequence of i.i.d. Laplace random variables. Let

$$M_{500} = \frac{X_1 + \cdots + X_{500}}{500}.$$

(a) Find $\mathbb{E}[X]$. Express your answer in terms of λ.

(b) Let $\lambda = 10$. Using Chebyshev's inequality, find a lower bound of

$$\mathbb{P}[-0.1 \leq M_{500} \leq 0.1].$$

(c) Let $\lambda = 10$. Using the Central Limit Theorem, find the probability

$$\mathbb{P}\left[-0.1 \leq M_{500} \leq 0.1\right].$$

You may leave your answer in terms of the $\Phi(\cdot)$ function.

Exercise 23

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables such that $X_n = \pm 1$ with equal probability. Let

$$S_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} X_n.$$

Note that S_N is a *standardized sum*, not the sample average $\overline{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$ used elsewhere in this chapter. Prove the Central Limit Theorem for this particular sequence of random variables by showing that

(a) $\mathbb{E}[S_N] = 0$, $\mathrm{Var}[S_N] = 1$.

(b) The moment-generating function of S_N is $M_{S_N}(s) \to e^{\frac{s^2}{2}}$ as $N \to \infty$.

Exercise 24

Let $X_1, \ldots, X_N$ be a sequence of i.i.d. random variables with mean and variance

$$\mathbb{E}[X_n] = \mu \quad \text{and} \quad \mathrm{Var}[X_n] = \sigma^2, \quad n = 1, \ldots, N.$$

The distribution of X_n is unknown. Let

$$M_N = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

Use the Central Limit Theorem to estimate the probability $\mathbb{P}[M_N > 2\mu]$.

Exercise 25

Chapter 6.4.4 warns that the Central Limit Theorem is trustworthy only for *small* deviations. This exercise measures how badly it fails in the tail. Let $X_1, \ldots, X_N$ be i.i.d. Exponential(1), and let $S_N = \sum_{n=1}^{N} X_n$, so that $S_N \sim \mathrm{Erlang}(N, 1)$ with $\mathbb{E}[S_N] = N$ and $\mathrm{Var}[S_N] = N$.

(a) Write the Central Limit Theorem approximation for $\mathbb{P}\left[S_N > N + t\sqrt{N}\right]$, and write the exact value in terms of the Erlang CDF.

(b) For $N = 10$, compute the exact probability and the CLT approximation for $t =$

1, 2, 3, 4, and tabulate the ratio exact/approximation. Use MATLAB or Python (in Python, `scipy.stats.gamma.sf(x, a=N)` gives the Erlang tail).

(c) Repeat part (b) with $N = 100$.

(d) Using your two tables, explain in what regime the CLT approximation is reliable, in what regime it is not, and why increasing N helps but does not remove the problem. Which direction does the CLT err in here, and what feature of the exponential distribution causes it?

Exercise 26

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables uniformly distributed on $[0, \theta]$ for unknown $\theta > 0$. Define $M_n = \max(X_1, \ldots, X_n)$.

(a) Find the CDF and PDF of M_n.

(b) Show that $M_n \xrightarrow{p} \theta$ as $n \to \infty$.

(c) Apply Markov's inequality to bound $\mathbb{P}[M_n \leq (1-\varepsilon)\theta]$ for $\varepsilon \in (0, 1)$, and compare with the exact probability.

Exercise 27

Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli(p) random variables with $0 < p < 1$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(a) Apply Chebyshev's inequality to bound $\mathbb{P}[|\bar{X}_n - p| \geq \varepsilon]$.

(b) Use the MGF-based Chernoff approach to find an exponential upper bound on $\mathbb{P}[\bar{X}_n \geq p + \varepsilon]$ for $\varepsilon > 0$. Optimize over the free parameter $t > 0$ to simplify the bound.

(c) State which bound is tighter for large n and explain why.