1 An illustrative example

1.1 A simple example

Consider the experiment in which one rolls a fair dice n times, and let X_i be the random variable corresponding to the result on the i^{th} roll. We have the following probability mass function for X_1:

$$p_{X_1}(1) = \frac{1}{6}\ ,\ p_{X_1}(2) = \frac{1}{6}\ ,\ p_{X_1}(3) = \frac{1}{6}\ ,\ p_{X_1}(4) = \frac{1}{6}\ ,\ p_{X_1}(5) = \frac{1}{6}\ ,\ p_{X_1}(6) = \frac{1}{6}$$

Now, let us consider the random variable S_2 defined by

$$S_2 = X_1 + X_2$$

What is the probability mass function of S_2?

S_2 can take the values $2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$, and, as we saw in Lecture 7, has the following probability mass function:

$$\begin{aligned}
&p_{S_2}(2) = \frac{1}{36}\ ,\ p_{S_2}(3) = \frac{2}{36}\ ,\ p_{S_2}(4) = \frac{3}{36}\ ,\ p_{S_2}(5) = \frac{4}{36}\ ,\ p_{S_2}(6) = \frac{5}{36}\ ,\ p_{S_2}(7) = \frac{6}{36}\\
&p_{S_2}(8) = \frac{5}{36}\ ,\ p_{S_2}(9) = \frac{4}{36}\ ,\ p_{S_2}(10) = \frac{3}{36}\ ,\ p_{S_2}(11) = \frac{2}{36}\ ,\ p_{S_2}(12) = \frac{1}{36}
\end{aligned}$$

Finally, let us consider the random variable S_3 defined by

$$S_3 = X_1 + X_2 + X_3 = S_2 + X_3$$

Let us also construct the probability mass function for S_3.

S_3 can take the values $3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18$, and it is a relatively simple exercise to verify that

$$\begin{aligned}
&p_{S_3}(3) = \frac{1}{216}\ ,\ p_{S_3}(4) = \frac{3}{216}\ ,\ p_{S_3}(5) = \frac{6}{216}\ ,\ p_{S_3}(6) = \frac{10}{216}\ ,\ p_{S_3}(7) = \frac{15}{216}\ ,\ p_{S_3}(8) = \frac{21}{216}\\
&p_{S_3}(9) = \frac{25}{216}\ ,\ p_{S_3}(10) = \frac{27}{216}\ ,\ p_{S_3}(11) = \frac{27}{216}\ ,\ p_{S_3}(12) = \frac{25}{216}\ ,\ p_{S_3}(13) = \frac{21}{216}\\
&p_{S_3}(14) = \frac{15}{216}\ ,\ p_{S_3}(15) = \frac{10}{216}\ ,\ p_{S_3}(16) = \frac{6}{216}\ ,\ p_{S_3}(17) = \frac{3}{216}\ ,\ p_{S_3}(18) = \frac{1}{216}
\end{aligned}$$

Let us plot the probability mass functions of $S_1 = X_1$, S_2 and S_3. They are shown in Figure 1, along with the probability mass function of S_5, whose calculation we do not show here. We empirically observe a shift to the right of the probability mass functions, as well as a widening and a smoothing of the functions.

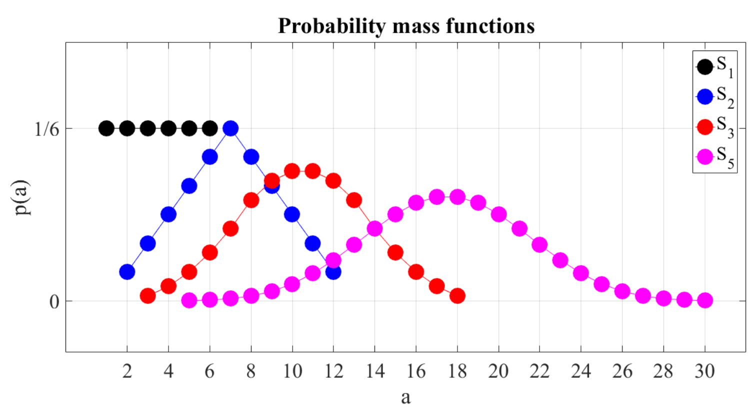

Figure 1: Probability mass functions of the discrete random variables S_1, S_2, S_3 and S_5, corresponding to the sum of the results of 1, 2, 3, and 5 dice respectively, as discussed in the main text. We see that as the number of dice rolled increases, the probability mass function shifts to the right, and its variance increases.

1.2 Generalizing the example

• Let us consider n i.i.d. random variables X_1, X_2,...,X_n, each with mean μ and variance σ^2, and let S_n be the random variable corresponding to the sum

$$S_n = X_1 + X_2 + X_3 + \ldots + X_n$$

Following the same steps as we have done in Lecture 10, since the random variables are independent, we easily obtain:

$$E[S_n] = n\mu \qquad \text{Var}(S_n) = n\sigma^2$$

Thus, S_n is a random variable which is shifted by a factor of n, and with a much wider distribution, as we had observed in Figure 1 corresponding to the elementary example of the previous section.

• On the other hand, in the law of large numbers, we saw that the sample mean

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

had the same expectation μ as the X_i, and a much smaller variance, $\frac{\sigma^2}{n}$. $\overline{X_n}$ therefore has a distribution which is centered on μ and very localized, i.e. with a very small variance. This is sketched in Figure 2 for continuous random variables.

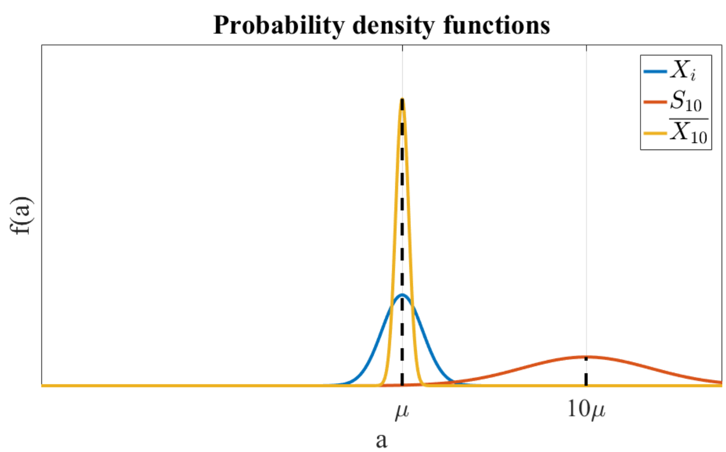

Figure 2: Sketch of the probability density functions for a continuous random variable X_i, the sum S_{10} of 10 continuous random variables which are i.i.d. with X_i, and the sample mean $\overline{X_{10}}$ of these 10 i.i.d. random variables.

As compared to S_n, $\overline{X_n}$ has the advantageous property of keeping the right mean μ, but has a distribution which can sometimes be complicated to deal with as it is nearly singular.

• Finally, let

$$T_n = \frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}$$

Following again the same steps as in Lecture 10, using the i.i.d. property of the random variables X_i, we would readily find that the expectation of T_n is $\sqrt{n}\mu$, and that the variance of T_n is

$$\text{Var}(T_n) = \frac{n}{n}\text{Var}(X_1) = \sigma^2$$

Thus, as compared to $\overline{X_n}$, T_n does not keep the desired mean μ, but it has the advantageous property that it has the same variance as the X_i (although the distribution of T_n may have a different shape than that of the X_i).

The purpose of the second part of this lecture is to show:

1. How one can construct a random variable which has both the desired mean and the desired variance.
2. The remarkable fact that T_n approaches a normal distribution.

2 The central limit theorem

2.1 Standardizing averages

Let $S_n = X_1 + X_2 + \ldots + X_n$ as defined before. By the linearity of expectation, the random variable

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}}$$

has mean zero.

Furthermore,

$$\mathrm{Var}(Z_n) = \mathrm{Var}\left(\frac{S_n}{\sigma_{S_n}} - \frac{E[S_n]}{\sigma_{S_n}}\right) = \frac{1}{\sigma_n^2}\mathrm{Var}(S_n) = 1$$

We say that

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

is a **standardized random variable**, in the sense that it is constructed to have **mean 0 and variance 1**.

Note that we may rewrite Z_n as

$$Z_n = \frac{n\overline{X_n} - n\mu}{\sqrt{n}\sigma} = \frac{\overline{X_n} - E[\overline{X_n}]}{\sqrt{\mathrm{Var}(\overline{X_n})}}$$

so Z_n may also be viewed as the *standardized version* of the sample average $\overline{X_n}$.

The remarkable central limit theorem tells us that whatever the distribution of the X_i is, the cumulative distribution function of Z_n is, in the limit $n \to +\infty$, Φ, i.e. that of a **normal random variable with mean 0 and variance 1**.

2.2 The central limit theorem

Theorem: Let $X_1, X_2, \ldots, X_n$ be any sequence of independent and identically distributed random variables with finite mean μ and finite variance σ^2, and let Z_n be the random variable defined by

$$Z_n = \sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} = \frac{\overline{X_n} - E[\overline{X_n}]}{\sqrt{\mathrm{Var}(\overline{X_n})}}$$

where

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is the sample mean of the n X_i.
Then, for any $x \in \mathbb{R}$,

$$\boxed{\lim_{n \to +\infty} F_{Z_n}(x) = \Phi(x)}$$

where Φ is the cumulative distribution function of the normal distribution $N(0,1)$ with mean 0 and variance 1.

Notes: •

$$Z_n = \frac{\sqrt{n}}{\sigma}(\overline{X_n} - \mu) \;\Leftrightarrow\; \overline{X_n} = \frac{\sigma}{\sqrt{n}}Z_n + \mu$$

Thus, from the rules we derived in lecture 6, if Z_n approximates a normal distribution with mean 0 and variance 1, then $\overline{X_n}$ approaches a normal distribution with mean μ and variance σ^2/n.

As we noted earlier in this lecture, however, as $n \to +\infty$, $\overline{X_n}$ does not approach a normal distribution in the standard way: it is a distribution with vanishing variance, and thus unbounded maximum (i.e. a maximum which is finite for finite n, but going to infinity in the limit $n \to +\infty$). Mathematicians like to treat such special functions with much care, *which is why the central limit theorem is expressed and proved in terms of the standardized Z_n.*

In practice, however, since we will mostly deal with situations in which n is finite, we *will often view* $\overline{X_n}$ *as a normal distribution with mean* μ *and variance* σ^2/n.

• It is remarkable that the central limit theorem applies to *any* sequence of i.i.d. random variables, whatever their distribution (as long as the mean and variance of the distribution are finite).

• We will see that in a large number of situations, the central limit theorem gives accurate results even for n relatively small. This is convenient, as computing the distribution of S_n or $\overline{X_n}$ can be tedious, as we had a glimpse of in the case of the dice.

Knowing μ and σ^2, it is much faster to construct an approximation of these distributions through Z_n and the use of the central limit theorem.

• The central limit theorem is a key reason why the normal distribution plays such an important role in probability and statistics.

2.3 Applications of the central limit theorem

• Let us return to our polling problem, first treated in Lecture 10. In that problem, we were looking for n such that

$$P(|\overline{X_n} - \mu| \geq 0.01) \leq 0.05$$

Now,

$$\begin{aligned} |\overline{X_n} - \mu| \geq 0.01 &\Leftrightarrow \left| \frac{\sigma}{\sqrt{n}} Z_n + \mu - \mu \right| \geq 0.01 \\ &\Leftrightarrow \left| \frac{\sigma}{\sqrt{n}} Z_n \right| \geq 0.01 \\ &\Leftrightarrow |Z_n| \geq 0.01 \frac{\sqrt{n}}{\sigma} \end{aligned}$$

Hence,

$$P(|\overline{X_n} - \mu| \geq 0.01) \leq 0.05 \Leftrightarrow P(|Z_n| \geq 0.01 \frac{\sqrt{n}}{\sigma}) \leq 0.05$$

As in Lecture 10, we do not know σ, but have the conservative estimate $\sigma \leq \frac{1}{2}$. With that upper bound, we can write

$$P(|Z_n| \geq 0.01 \frac{\sqrt{n}}{\sigma}) \leq P(|Z_n| \geq 0.02\sqrt{n})$$

So we look for a conservative estimate of n by looking for the n such that

$$P(|Z_n| \geq 0.02\sqrt{n}) \leq 0.05$$

is satisfied. The idea is now to see whether using an approximation of the central limit theorem, i.e. for n finite, gives us a finer estimate (i.e. smaller n) than we had by using Chebyshev's inequality and the law of large numbers in Lecture 10.

Let $n = 10000$, and assume, according to the central limit theorem but applied for that finite value of n, that Z_n has a normal distribution $N(0, 1)$. Then, by symmetry of that distribution,

$$P(|Z_{10000}| \geq 2) \leq 0.05 \Leftrightarrow 2P(Z_{10000} \geq 2) \leq 0.05$$

We want to see if this inequality is satisfied, which would confirm that according to the central limit theorem (used in our appoximate way), polling 10000 people is sufficient for the desired level of confidence and desired accuracy. To do so, we must turn the inequality into a form which involves the cumulative distribution Φ of the standard normal distribution. This is obtained by rewriting the inequality as

$$2(1 - P(Z_{10000} \leq 2)) \leq 0.05$$

To see whether this inequality holds, we turn to the tabulated values of Φ, and find

$$P(Z_n \leq 2) \approx 0.97725$$

Thus, our inequality becomes

$$2 \cdot 0.02275 \leq 0.05 \iff 0.0455 \leq 0.05$$

The inequality is indeed true, which tells us that phoning 10,000 people (and actually slightly fewer) would be sufficient to satisfy our accuracy and confidence criteria. Application of the central limit theorem thus gives a tighter bound than Chebyshev's inequality did.

We can find the value of n satisfying

$$P(|Z_n| \geq 0.02\sqrt{n}) \leq 0.05$$

by looking at the table for Φ. By symmetry of the standard normal distribution, we are looking for n such that

$$P(Z_n \geq 0.02\sqrt{n}) \leq 0.025 \iff 0.975 \leq P(Z_n \leq 0.02\sqrt{n})$$

According to the tables, this is satisfied when

$$0.02\sqrt{n} \approx 1.96 \iff n \approx 9604$$

- **Normal approximation of the binomial distribution**

The previous example involved a sum of Bernoulli random variables. We know that the result is a binomial random variable. The central limit theorem is often used to approximate a binomial distribution with a normal distribution. Indeed, while easy to write down, formulae involving the binomial distribution can be tedious to evaluate.

Let $(X_i)_{i=1,2,\ldots,n}$ be a sequence of i.i.d. Bernoulli random variables with parameter $p = \frac{1}{4}$.
We let

$$S_n = \sum_{i=1}^{n} X_i$$

We saw in Lecture 3 that S_n has a binomial distribution with parameters n and p. Furthermore,

$$E[S_n] = np = \frac{n}{4} \quad , \quad \mathrm{Var}(S_n) = np(1-p) = \frac{3n}{16}$$

Let $n = 14$. We want to find $P(S_n \leq 5)$.
The exact answer is

$$\left(\frac{3}{4}\right)^{14} + 14\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^{13} + \binom{14}{2}\left(\frac{1}{4}\right)^{2}\left(\frac{3}{4}\right)^{12} + \binom{14}{3}\left(\frac{1}{4}\right)^{3}\left(\frac{3}{4}\right)^{11}$$
$$+ \binom{14}{4}\left(\frac{1}{4}\right)^{4}\left(\frac{3}{4}\right)^{10} + \binom{14}{5}\left(\frac{1}{4}\right)^{5}\left(\frac{3}{4}\right)^{9} \approx 0.8883$$

Let us see how close to that answer an (approximate) application of the central limit theorem can take us. We consider

$$Z_n = \frac{S_n - \frac{14}{4}}{\sqrt{\frac{3 \cdot 14}{16}}} \approx \frac{S_n - 3.5}{1.6202}$$

which we approximate to be normally distributed with mean 0 and variance 1.

$$S_n \leq 5 \iff S_n - 3.5 \leq 1.5 \iff Z_n \leq 0.9258$$

We now use the tabulated values of Φ to find

$$P(Z_n \leq 0.9258) = \Phi(0.9258) \approx 0.823$$

We see that the approximation is decent, but not particularly impressive.
Here is a trick to improve that approximation. Since S_n only takes integer values,

$$P(S_n \leq 5) = P(S_n < 6)$$

A good approach is therefore to consider something in the middle, such as:

$$P(S_n \leq 5.5)$$

Repeating the same steps as before, this becomes

$$P(Z_n \leq 1.2344) \approx 0.891$$

The approximation is significantly better.

For a reliable and accurate application of the central limit theorem for that purpose, it would of course be even better if we considered a situation in which n is larger, since the central limit theorem becomes exact only in the limit $n \to +\infty$.