6 The law of large numbers and the central limit theorem

In this section we study two of the most important results in probability and statistics: the law of large numbers and the central limit theorem.

6.1 Markov and Chebyshev's inequalities

Lemma 6.1. Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be discrete or continuous random variables with well-defined expectations. Suppose that $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. Then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Proof. We focus on the discrete case. Consider the random variable $Z = Y - X$. For all $\omega \in \Omega$ we have $Z(\omega) = Y(\omega) - X(\omega) \geq 0$ by the hypothesis. Thus $S_Z = \{Z(\omega) : \omega \in \Omega\} \subseteq [0, \infty)$, and so

$$\mathbb{E}[Y] - \mathbb{E}[X] = \mathbb{E}[Z] = \sum_{z \in S_Z} z \cdot \mathbb{P}(Z = z) \geq 0.$$

The first equality uses linearity of expectation, the second is by definition of $\mathbb{E}[Z]$, the final inequality holds since $\mathbb{P}(Z = z) \geq 0$ and $z \geq 0$ for all $z \in S_Z$. It follows that $\mathbb{E}[X] \leq \mathbb{E}[Y]$. ☐

Definition 6.2. Let $A \subseteq \Omega$ be an event. The *indicator variable of the event* A is the random variable $\mathbf{1}_A : \Omega \to \mathbb{R}$ defined for all $\omega \in \Omega$ by

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

Remark 6.3. Note that $\mathbf{1}_A \sim \mathrm{Ber}_p$ with $p = \mathbb{P}(A)$, giving $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$.

We are now ready to prove Markov's inequality, a fundamental result. It is very intuitive – a non-negative random variable rarely takes values that are much larger than its expectation [1].

Theorem 6.4 (Markov's inequality)**.** *Let $X : \Omega \to \mathbb{R}$ be a non-negative random variable with well-defined expectation. Then, given any $t > 0$, we have*

$$\mathbb{P}\big(X \geq t\big) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Let $A \subseteq \Omega$ denote the event $A := \{\omega \in \Omega : X(\omega) \geq t\}$. We claim that $t \cdot \mathbf{1}_A(\omega) \leq X(\omega)$ for all $\omega \in \Omega$. To see this, we split according to whether $\omega \in A$ or $\omega \notin A$.

- If $\omega \in A$ then $t \cdot \mathbf{1}_A(\omega) = t \leq X(\omega)$, by definition of A.
- If $\omega \notin A$ then $t \cdot \mathbf{1}_A(\omega) = 0 \leq X(\omega)$. (We used X is *non-negative* here, as this gave $X(\omega) \geq 0$.)

Thus, by Lemma 6.1 we have $\mathbb{E}[t \cdot \mathbf{1}_A] \leq \mathbb{E}[X]$, and so

$$t \cdot \mathbb{P}(X \geq t) = t \cdot \mathbb{P}(A) = t \cdot \mathbb{E}\big[\mathbf{1}_A\big] = \mathbb{E}\big[t \cdot \mathbf{1}_A\big] \leq \mathbb{E}[X].$$

Dividing the left and right-hand sides by t gives the theorem. ☐

[1] A non-negative random variable is just a random variable which never takes negative values.

Remark 6.5.

- Markov's inequality is used *very* often in mathematics. Here are two reasons:
  - It is very flexible as it assumes almost nothing about X; just that it is non-negative [2].
  - It provides a connection between the expectation $\mathbb{E}[X]$ and probabilities associated to X.
- Markov's inequality is only really useful if $t > \mathbb{E}[X]$, as otherwise $\mathbb{P}(X \geq t) \leq 1$ is better.

Example 6.6. Let $X \sim \mathrm{bin}_{100,0.1}$. Then $\mathbb{E}[X] = 100 \cdot 0.1 = 10$. By Markov's inequality, we have $\mathbb{P}(X \geq 50) \leq 10/50 = 0.2$.

Example 6.7. Let X be a non-negative random variable with $\mathbb{P}(X \geq 10) = 0.3$. Then, by Markov's inequality, $\mathbb{E}[X] \geq 0.3 \cdot 10 = 3$.

Example 6.8. On a social network, an average user has 300 friends. Equivalently, if we select a random person on the network and let X equal the number of their friends then $\mathbb{E}[X] = 300$. Markov's inequality then gives $\mathbb{P}(X \geq 900) \leq 1/3$, and so at most a third of the users have 900 friends or more.

One of the most important applications of Markov's inequality is to prove Chebyshev's inequality, which shows that a random variable with small variance is typically close to its expectation.

Theorem 6.9 (Chebyshev's inequality)**.** *Let X be a random variable with well-defined expectation and variance. Then, for all $\varepsilon > 0$, we have*

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq \varepsilon\Big) \leq \frac{\mathbb{V}\mathrm{ar}(X)}{\varepsilon^2}.$$

Proof. Note that if $\omega \in \Omega$ then $|X(\omega) - \mathbb{E}[X]| \geq \varepsilon$ if and only if $|X(\omega) - \mathbb{E}[X]|^2 \geq \varepsilon^2$. This gives

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geq \varepsilon\big) = \mathbb{P}\big(|X - \mathbb{E}[X]|^2 \geq \varepsilon^2\big).$$

Now apply Markov's inequality to the non-negative random variable $\big(X - \mathbb{E}[X]\big)^2$, with $t = \varepsilon^2$, to get

$$\mathbb{P}\big(|X - \mathbb{E}[X]| \geq \varepsilon\big) = \mathbb{P}\big(|X - \mathbb{E}[X]|^2 \geq \varepsilon^2\big) \leq \frac{\mathbb{E}\big[\big(X - \mathbb{E}[X]\big)^2\big]}{\varepsilon^2} = \frac{\mathbb{V}\mathrm{ar}(X)}{\varepsilon^2}.$$

The last equality here holds by definition of variance. □

Remark 6.10. Since $\mathbb{V}\mathrm{ar}(X) = \sigma_X^2$, if we take $\varepsilon = \alpha\sigma_X$ with $\alpha > 0$, Chebyshev's inequality gives

$$\mathbb{P}\Big(|X - \mathbb{E}[X]| \geq \alpha\sigma_X\Big) \leq \alpha^{-2}.$$

It follows that $|X - \mathbb{E}[X]|$ is typically not much larger than σ_X (see Remark 5.32).

Example 6.11. Let $X \sim \mathrm{bin}_{50,0.4}$. Then $\mathbb{E}[X] = 20$ and $\mathbb{V}\mathrm{ar}(X) = 12$ so $\mathbb{P}(|X - 20| \geq 5) \leq 12/25$.

[2] Check that the theorem is false if X can take negative values.

Example 6.12 (Binomial distribution)**.** Let $X \sim \text{bin}_{n,p}$ for $n \in \mathbb{N}$, $p \in (0,1)$. By Chebyshev's inequality, for all $\varepsilon > 0$, we have

$$\mathbb{P}\big(|X - np| \geq \varepsilon n\big) \leq \frac{\mathbb{V}\text{ar}(X)}{(\varepsilon n)^2} = \frac{np(1-p)}{(\varepsilon n)^2} = \frac{p(1-p)}{\varepsilon^2 n}.$$

Hence, $\lim_{n\to\infty} \mathbb{P}\big(|X - np| \geq \varepsilon n\big) = 0$.

Example 6.13. We toss a coin which appears as tails with some unknown probability $p \in (0,1)$ a total number of n times. Let $\hat{S}_n$ be the number of times that tails appears. Then, $\hat{S}_n \sim \text{bin}_{n,p}$. How large must n be to guarantee that $\hat{p}_n = \hat{S}_n/n$ satisfies $|\hat{p}_n - p_n| \leq 0.01$ with probability at least 0.95?

This question is ideal for applying Chebyshev's inequality. First note that as $\hat{S}_n \sim \text{bin}_{n,p}$ we have $\mathbb{V}\text{ar}(\hat{S}_n) = np(1-p)$. Secondly, by Proposition 5.39 (ii) we have $\mathbb{V}\text{ar}(\hat{S}_n/n) = n^{-2}\mathbb{V}\text{ar}(\hat{S}_n)$. Therefore

$$\mathbb{V}\text{ar}\left(\frac{\hat{S}_n}{n}\right) = \frac{\mathbb{V}\text{ar}(\hat{S}_n)}{n^2} = \frac{np(1-p)}{n^2} \leq \frac{1}{4n}$$

where the last inequality holds since $p(1-p) \leq 1/4$ for $p \in [0,1]$. Thus

$$\mathbb{P}\big(|\hat{p}_n - p| \geq 0.01\big) = \mathbb{P}\big(\big|\hat{S}_n - np\big| \geq 0.01n\big) \leq \frac{1}{4n \cdot (0.01)^2} = \frac{2500}{n},$$

which is smaller than 0.05 for all $n \geq 50,000$.

Example 6.14 (Mean and variance estimation)**.** We return to the estimators for mean μ and variance σ^2 of an unknown distribution using independent random samples $\hat{X}_1, \ldots, \hat{X}_n$ discussed in Section 5.6. We introduced the sample average $\hat{\mu}_n = n^{-1}(\hat{X}_1 + \cdots + \hat{X}_n)$ and the estimator for sample fluctuations $\hat{\sigma}_n^2 := n^{-1}\sum_{i=1}^n (\hat{X}_i - \hat{\mu}_n)^2$. We proved that $\mathbb{E}\left[\hat{\mu}_n\right] = \mu, \mathbb{E}\left[\hat{\sigma}_n^2\right] = \left(1 - \frac{1}{n}\right)\sigma^2$ and both variances of $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ converge to zero as $n \to \infty$. By Chebyshev's inequality, for $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \varepsilon\right) = 0$$

and

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\hat{\sigma}_n^2 - \left(1 - \frac{1}{n}\right)\sigma^2\right| \geq \varepsilon\right) = 0 \implies \lim_{n\to\infty} \mathbb{P}\big(\big|\hat{\sigma}_n^2 - \sigma^2\big| \geq \varepsilon\big) = 0.$$

Estimators with these properties are called *consistent*.

6.2 The law of large numbers

An infinite collection of random variables $X_1, X_2, \ldots$ are *independent and identically distributed* [3] if:

- $X_1, \ldots, X_n$ are independent for all $n \in \mathbb{N}$ (as in Definition 4.24), and
- all X_i follow the same distribution, that is $F_{X_i}(t) = F_{X_j}(t)$ for all $i, j \in \mathbb{N}$ and $t \in \mathbb{R}$.

[3] It is very common to write i.i.d. to abbreviate this condition.

Theorem 6.15 (Law of large numbers)**.** *Suppose that $X_1, X_2, \dots$ are independent and identically distributed random variables with well-defined expectations and variances, where $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\mathrm{ar}(X_i) \le c$ for all $i \in \mathbb{N}$. Then, setting $S_n = X_1 + \cdots + X_n$, for any $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\frac{S_n}{n} - \mu\Big| \ge \varepsilon\Big) = 0.$$

Proof. As $X_1, \dots, X_n$ are independent by Theorem 5.50 we have

$$\mathbb{V}\mathrm{ar}(S_n) = \mathbb{V}\mathrm{ar}(X_1) + \cdots + \mathbb{V}\mathrm{ar}(X_n) \le nc,$$

using that $\mathbb{V}\mathrm{ar}(X_i) \le c$ for $i \in \{1, \dots, n\}$. Applying Chebyshev's inequality to S_n gives

$$\mathbb{P}\left(\Big|\frac{S_n}{n} - \mu\Big| \ge \varepsilon\right) = \mathbb{P}\left(\big|S_n - n\mu\big| \ge \varepsilon n\right) \le \frac{\mathbb{V}\mathrm{ar}\,(S_n)}{(\varepsilon n)^2} \le \frac{cn}{\varepsilon^2 n^2} = \frac{c}{n\varepsilon^2}.$$

The right hand side tends to zero as $n \to \infty$. □

Remark 6.16. We have thought of the expectation $\mathbb{E}[X]$ of a random variable X as a sort of average, or typical value. The law of large numbers gives strong support to this view: if we take many independent samples according to X (e.g. n repeated dice rolls) and average the results, the result is extremely likely to be close to $\mathbb{E}[X]$.

Monte Carlo simulation. Here is a useful way to use randomness (and the law of large numbers) to approximate quantities. We are given a set S and and random variable Y which takes values in S and would like to estimate $\mathbb{P}(Y \in A)$ for some set $A \subseteq S$.

A natural way to try to do this is to take independent random variables $Y_1, \dots, Y_n$ which all follow the same distribution as Y and to use the ratio $|\{1 \le i \le n : Y_i \in A\}|/n$ to estimate $\mathbb{P}(Y_1 \in A)$. But does this work (most of the time)?

Yes, it does. Apply Theorem 6.15 to the random variables $X_i = \mathbf{1}_{\{Y_i \in A\}}$, for $i = 1, \dots, n$. Then for any $\varepsilon > 0$, we find

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{|\{1 \le i \le n : Y_i \in A\}|}{n} - \mathbb{P}\,(Y_1 \in A)\right| > \varepsilon\right) = 0, \tag{1}$$

This gives rise to the Monte Carlo method of approximation.

Example 6.17 (Integral approximation)**.** Given $f : [a, b] \to [0, \infty)$, we would like to find $\int_a^b f(x)dx$. This is generally difficult as many functions do not have an anti-derivative. However, if we are happy to *estimate* the integral then Monte Carlo simulation offers an solution.

Suppose that f is bounded by M on $[a, b]$. Let $Y_1, Y_2, \dots, Y_n$ be independent random variables with the uniform distribution on $[a, b] \times [0, M] = \{(x, y) : a \le x \le b, 0 \le y \le M\}$. As in the previous example, we have $\mathbb{P}\,(Y_1 \in A) = \mathrm{area}(A)/(M(b - a))$. In particular, letting B denote the set of points between the x-axis and f, we have

$$B = \big\{(x, y) : a \le x \le b, 0 \le y \le f(x)\big\}.$$

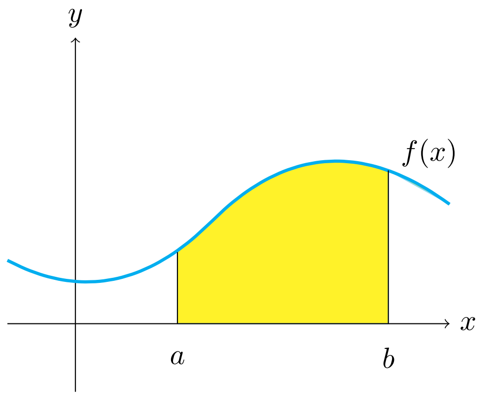

Figure 1: The area of the yellow region between the curve $f(x)$ and the x-axis represents $\int_a^b f(x)dx$.

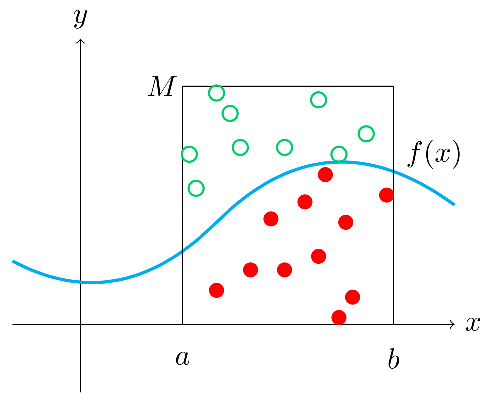

Figure 2: 11 of the 20 points lie between $f(x)$ and the x-axis, so we approximate the integral by $\frac{11}{20} \cdot M(b-a)$.

In particular, the area$(B) = \int_a^b f(x)dx$. It follows that $\mathbb{P}(Y_1 \in B) = (M(b-a))^{-1} \cdot \int_a^b f(x)dx$. As justified in (1), with probability close to 1 this gives

$$\frac{\big|\{1 \le i \le n : Y_i \in A\}\big|}{n} \approx \mathbb{P}(Y_1 \in B) = \frac{\int_a^b f(x)dx}{M \cdot (b-a)} \implies \int_a^b f(x)dx \approx M(b-a) \cdot \frac{|\{1 \le i \le n : Y_i \in B\}|}{n}.$$

Example 6.18 (Finding π)**.** A nice application of Monte Carlo simulation is in approximating π. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with the uniform distribution on $[-1,1]^2 = \{(x,y) : -1 \le x, y \le 1\}$. In other words, the two components of each Y_i are independent random variables with the continuous uniform distribution on $[-1,1]$. For a set $A \subseteq [-1,1]^2$, we have $\mathbb{P}(Y_1 \in A) = \text{area}(A)/4$, where area$(A)$ denotes the area of the set A [4]. As $\mathbb{P}(|Y_1| \le 1) = \mathbb{P}\left(Y_1 \in \{(x,y) \in [-1,1]^2 : x^2 + y^2 \le 1\}\right) = \pi/4$, the law of large numbers gives

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{|\{1 \le i \le n : |Y_i| \le 1\}|}{n} - \frac{\pi}{4}\right| > \varepsilon\right) = 0.$$

Thus $4 \cdot |\{1 \le i \le n : |Y_i| \le 1\}|/n$ is extremely likely to be a good approximation of π if n is large.

6.3 De Moivre-Laplace and the central limit theorem

Chebyshev's inequality shows that a random variable X with $X \sim \text{bin}_{n,p}$ typically differs from $\mathbb{E}(X) = np$ by about $\sigma_X = \sqrt{\mathbb{V}\text{ar}(X)} = \sqrt{np(1-p)}$. In this subsection we study the De Moivre–Laplace theorem which gives a finer description of the behaviour of X, showing that X is well–approximated by a normal distribution. The central limit theorem, a later generalisation of De Moivre–Laplace, is widely considered as the most important result in probability theory and statistics.

[4] The definition of area is clear for rectangles, circles and other familiar objects. For more general sets, such a definition is content of second and third year modules.

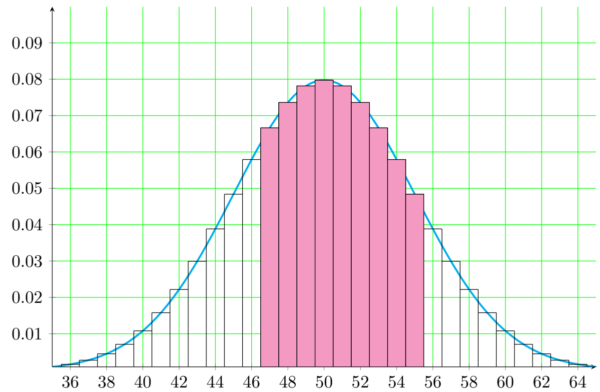

Figure 3: Mass function of $X \sim \mathrm{bin}_{100,0.5}$. The smooth curve is $x \mapsto \sqrt{1/(50\pi)}\exp\big(-(x-50)^2/50\big)$. The area of the shaded region equals to $\mathbb{P}\,(47 \le X \le 55)$ and is approximated by the integral of the smooth curve from 46 to 55, or, more precisely, by the integral from 46.5 to 55.5.

Theorem 6.19 (De Moivre–Laplace)**.** *Let $p \in (0,1)$. Given $n \in \mathbb{N}$ let $X_n \sim \mathrm{bin}_{n,p}$. Then for any $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_n - np}{\sqrt{np(1-p)}} \le t\right) = \mathbb{P}\big(\mathcal{N} \le t\big) = \Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}dx,$$

where $\mathcal{N} \sim N(0,1)$ follows the standard normal distribution.

Proof. Omitted. □

The De Moivre–Laplace theorem is very useful as it allows us to estimate $\mathbb{P}(X_n \le k)$ for $X_n \sim \mathrm{bin}_{n,p}$ by a calculation for the standard normal distribution. Given such k, let

$$x(n,k) := \frac{k - np}{\sqrt{np(1-p)}}.$$

Then letting $\mathcal{N}$ denote a standard normal distribution, Theorem 6.19 gives [5]

$$\mathbb{P}(X_n \le k) = \mathbb{P}\bigg(\frac{X - np}{\sqrt{np(1-p)}} \le x(n,k)\bigg) \approx \Phi\big(x(n,k)\big) = \mathbb{P}\Big(np + \sqrt{np(1-p)}\cdot\mathcal{N} \le k\Big). \tag{2}$$

We note that if $|x(n,k)|$ is not too large, the so-called midpoint rule [6] is a little more accurate

$$\mathbb{P}\,(X_n \le k) \approx \Phi\left(x(n,k) + \frac{1}{2\sqrt{np(1-p)}}\right) = \mathbb{P}\left(np + \sqrt{np(1-p)}\cdot\mathcal{N} \le k + 1/2\right). \tag{3}$$

This is called the *continuity correction* and is supported by Figure 3 and the following example.

Example 6.20 (Dice)**.** We roll a fair dice 600 times, and let X denote the number of times '6' appears. We are interested in the event $\{90 \le X \le 100\}$. Exact computations reveal that

$$\mathbb{P}\,(90 \le X \le 100) = \mathrm{bin}_{600,1/6}(90) + \mathrm{bin}_{600,1/6}(91) + \cdots + \mathrm{bin}_{600,1/6}(100) = 0.4024\ldots$$

[5]As rule of thumb, one often considers the criterion $np(1-p) \ge 9$ for the applicability of this approximation.

[6]This is a result from numerical integration, which we will not discuss further.

As $\sqrt{np(1-p)} = \sqrt{600 \cdot \frac{1}{6} \cdot \frac{5}{6}} = 9.1287\ldots$, (2) gives

$$\mathbb{P}\left(90 \le X \le 100\right) = \mathbb{P}\left(X \le 100\right) - \mathbb{P}\left(X \le 89\right) \approx \Phi(0) - \Phi\left(-\frac{11}{9.1287\ldots}\right) = 0.3858\ldots$$

Similarly, the estimate (3) involving the continuity correction yields the estimate

$$\mathbb{P}\left(90 \le X \le 100\right) = \mathbb{P}\left(X \le 100\right) - \mathbb{P}\left(X \le 89\right) \approx \Phi(0.0547\ldots) - \Phi(-1.1502\ldots) = 0.3968\ldots.$$

Example 6.21 (Greedy manager)**.** A hotel has 250 rooms. Statistically, the probability of a no-show is 0.15 per room and night. Looking to take advantage here, the hotel manager decides to overbook and tolerate a probability of 0.03 to exceed capacities. How many rooms can the manager offer?

Let n be the number of bookings and X be the number of guests showing up. Then, $X \sim \mathrm{bin}_{n,0.85}$. Overbooking happens if and only if $X \ge 251$, and this event shall have probability at most 3%. Using the De Moivre-Laplace theorem with continuity correction, that is (3), we obtain

$$\mathbb{P}\left(X \ge 251\right) = 1 - \mathbb{P}\left(X \le 250\right) \approx 1 - \Phi\left(x(n, 250) + \frac{1}{2\sqrt{0.1275n}}\right).$$

From Table 2, we can read off that the right hand side is smaller than 0.03 if (and only if)

$$x(n, 250) + \frac{1}{2\sqrt{0.1275n}} \ge 1.89$$

As $x(n, 250) = (250 - 0.85n)/\sqrt{0.1275n}$, this corresponds to $n \le 281$. Thus, the hotel manager can offer at most 281 beds.

Example 6.22 (Bribes)**.** A city has $N = 1,000,000$ residents, and two candidates, A and B, run for mayor. Every citizen flips a fair coin to come to reach a decision and so the two candidates are equally likely to win. (A tie is possible, but very unlikely).

Now suppose candidate A bribes $1,000$ voters. Does this increase the candidate's chances significantly? Assume that the remaining $999,000$ people still flip fair coins and denote by X the total number of votes candidate A receives from <u>these</u> people. Then X follows the binomial distribution with parameters $999,000$ and $1/2$. By (2), we have

$$\begin{aligned}\mathbb{P}\left(A \text{ wins}\right) = \mathbb{P}\left(X \ge 499,001\right) = 1 - \mathbb{P}\left(X \le 499,000\right) &= 1 - \mathbb{P}\left(\frac{X - 499500}{499.7499\ldots} \le -\frac{500}{499.7499\ldots}\right) \\ &\approx \Phi(1) = 0.841\ldots\end{aligned}$$

Thus, only $1,000$ bribes are enough to boost the candidate's chances significantly!

The central limit theorem generalises the De Moivre-Laplace theorem to <u>arbitrary</u> distributions with finite variance.

Theorem 6.23 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with well-defined expectations and variances, where $\mu := \mathbb{E}\left[X_i\right]$ and $\sigma^2 := \mathbb{V}\mathrm{ar}(X_i) > 0$. Then, setting $S_n = X_1 + \cdots + X_n$, for all $t \in \mathbb{R}$ we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\cdot\mu}{\sigma\sqrt{n}} \le t\right) = \mathbb{P}\big(\mathcal{N} \le t\big) = \Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\,dx, \tag{4}$$

where $\mathcal{N} \sim N(0,1)$ follows the standard normal distribution.

Proof. Omitted, as all proofs are quite involved. See Statistics or Fourier Analysis in later years. $\square$

Remark 6.24.

- The central limit theorem establishes the normal distribution as a key probability distribution. It is remarkable that the same limit appears in (4), regardless of the distribution you start with (i.e. the distribution of the random variables X_i).
- The theorem also explains why the normal distribution appears so often in the real-world: if a random quantity is determined as a sum of many (roughly) independent contributions then it can be approximated by a random variable following the normal distribution with suitable expectation and variance [7].

Example 6.25. Your friend rolls a fair dice $N = 10,000$ times and claims to have obtained a total sum of $35,854$. Should you believe this?

Let $X_1, \ldots, X_N$ denote the results of the individual dice rolls and let S denote the total sum. Then $X_1, \ldots, X_N$ all follow the uniform distribution on $\{1, \ldots, 6\}$ and these random variables are independent. We have $\mathbb{E}[X_1] = 3.5$, and $\mathbb{V}\mathrm{ar}(X_1) = \frac{35}{12}$, as computed in Section 5. By linearity of expectation, $\mathbb{E}[S] = \sum_{i=1}^{10000} \mathbb{E}[X_i] = 35,000$ and your friend claims a deviation of at least 854, which has probability

$$\mathbb{P}\Big(|S - \mathbb{E}[S]| \geq 854\Big) = \mathbb{P}\left(\frac{|S - N \cdot \mathbb{E}[X_1]|}{\sqrt{N} \cdot \sigma_{X_1}} \geq \frac{854}{100 \cdot \sqrt{\frac{35}{12}}}\right) \approx 2 \cdot \left(1 - \Phi\Big(\sqrt{\frac{12}{35}} \cdot 8.54\Big)\right)$$
$$= 2 \cdot \big(1 - \Phi(5.0005...)\big) = 5.7 \cdot 10^{-7} \text{ [3dp]}.$$

Theorem 6.23 justifies the approximation. This is a tiny probability, so the claim is extremely unlikely.

Confidence intervals. Let $\hat{X}_1, \ldots, \hat{X}_n$ be independent random variables following the same distribution with mean μ and variance σ^2. We consider these random variables as observed data and would like to estimate the unknown mean μ. The natural guess is to use the same average

$$\hat{\mu}_n = \frac{\hat{X}_1 + \hat{X}_2 + \ldots + \hat{X}_n}{n}. \tag{5}$$

How certain can we be that $\hat{\mu}_n$ is close to μ?

The central limit theorem is very useful in this context. Given $\alpha \in (0,1)$, an *α-confidence interval I_α* is an interval which satisfies

$$\mathbb{P}\big(\mu \in I_\alpha\big) \geq \alpha. \tag{6}$$

Here it is the *interval I_α* that is randomly chosen (based on $\hat{X}_1, \ldots, \hat{X}_n$), and not μ (which is fixed, as μ is the mean of the distribution). Thus, in words, (6) says that *the random interval I_α covers μ with probability at least α.*

[7] A nice example here is human height, which is roughly the sum of the length of many bones.

Here, we only construct *approximate* confidence intervals. Let us first assume that the value of σ is known to us. Given the samples $\hat{X}_1, \ldots, \hat{X}_n$, take $\hat{\mu}_n$ as in (5) and set I_α to be

$$I_\alpha = \left[\hat{\mu}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right],$$

where we still need select the value of z_α. This approximately satisfies (6), as by Theorem 6.23

$$\begin{aligned}\mathbb{P}(\mu \in I_\alpha) = \mathbb{P}\left(\mu \in \left[\hat{\mu}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right]\right) &= \mathbb{P}\Bigg(-z_\alpha \le \frac{\sum_{i=1}^n \hat{X}_i - n \cdot \mu}{\sigma\sqrt{n}} \le z_\alpha\Bigg) \\ &\approx \Phi(z_\alpha) - \Phi(-z_\alpha) = 2\Phi(z_\alpha) - 1 = \alpha,\end{aligned}$$

provided we take $z_\alpha = \Phi^{-1}\left(\frac{\alpha+1}{2}\right)$. Table 1 gives the values of z_α which are often chosen in applications.

α	0.9	0.95	0.97	0.99
z_α	1.645	1.96	2.17	2.575

Table 1: Table for z_α for popular values of α.

It turns out that, if the unknown distribution is $N(\mu, \sigma^2)$, then I_α gives an *exact* confidence interval, that is, $\mathbb{P}(\mu \in I_\alpha) = \alpha$ independently of the value of n.

Typically σ is unknown in applications. Then, one can either use an upper bound for σ valid for all possible values of μ or needs to estimate σ from the data adding a second level of approximation.

Example 6.26 (Swiss babies)**.** Between 1901 and 2016, $n = 9,569,478$ babies were born in Switzerland, $4,907,770$ of them were boys. We assume that the number of boys born followed the binomial distribution with parameters n and p where p is unknown. In other words, we are in the situation of the previous example with Bernoulli random variables $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ with parameter p. The observed probability is $\hat{p}_n = \hat{\mu}_n = 0.512856\ldots$ [8] We do not know $\sigma = \sqrt{p(1-p)}$, but we can use the bound $\sigma \le 1/2$ valid for all p. [9] An approximate 95%-confidence interval is given by

$$I_{0.95} = \left[\hat{p}_n - \frac{1.96}{2\sqrt{n}}, \hat{p}_n + \frac{1.96}{2\sqrt{n}}\right] = [\hat{p}_n - 0.000316\ldots, \hat{p}_n + 0.000316\ldots] = [0.512539\ldots, 0.51317\ldots].$$

We caution that this does not mean that $I_{0.95}$ covers p with probability at least (roughly) 0.95. Indeed, there is no longer any randomness here since p is fixed (but unknown) and $I_{0.95}$ also fixed above. Thus $I_{0.95}$ either contains p or not. We can only say the following: if $p \notin I_{0.95}$, then the probability that observed interval $I_{0.95}$ took such an extreme value was at most (roughly) 0.05.

[8] The ratio is similar in the UK and in other western societies.

[9] As the true value of p should be close to 1/2, this bound is very precise.

Most important takeaways in this chapter. You should

- know the statements of Markov's inequality, Chebyshev's inequality, the law of large numbers and the De Moivre-Laplace theorem.
- be able to apply these results in standard situations,
- be familiar with the concept of confidence intervals, their interpretation and how to find them in simple examples.

Table 2: Table for $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx, t \geq 0$.

t	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999