BIE-PST – Probability and Statistics

Lecture 8: Limit Theorems

Winter semester 2023/2024

Lecturer:
Francesco Dolce

Department of Applied Mathematics
Faculty of Information Technology
Czech Technical University in Prague

Table of contents

8 Limit theorems

8.1 Motivation

So far we have studied *individual* random *variables and vectors.*

Now we concentrate on the behavior of *sequences* of random variables, which arise from repeated experiments.

In particular, we are interested in the (arithmetic) *mean*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

and the *sum*

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

where $X_1, \ldots, X_n$ are *independent* random *variables* with an *identical distribution.*

Notation: *i.i.d. – independent and identically distributed.*

Limit theorems describe the behavior of $\bar{X}_n$ or S_n in limit for $n \to \infty$.

8.2 Basic inequalities

First, we obtain inequalities concerning tail probabilities:

Theorem 8.1 (– Markov's inequality)**.** *Let X be a random variable with a finite expectation. Then it holds that*

$$\mathrm{P}(|X| \geq a) \leq \frac{\mathrm{E}\,|X|}{a} \qquad \textit{for all } a > 0.$$

Proof. Denote the event $A = \{|X| \geq a\}$. Then it holds that $|X| \geq a\mathbb{1}_A$, where $\mathbb{1}_A$ is the indicator of the event A.

By taking expectation on both sides of the inequality we have

$$\mathrm{E}\,|X| \geq a\,\mathrm{E}(\mathbb{1}_A) = a\,\mathrm{P}(A) = a\,\mathrm{P}(|X| \geq a).$$

After dividing by a we obtain the inequality. □

Example 8.2 (– waiting for a bus)**.** Suppose that the time T which we spend waiting for a bus is exponentialy distributed with the expectation of 3 minutes. Find an upper bound for the probability that we need to wait for more than 10 minutes. Compare the estimate with the exact probability.

Because the waiting time T is non-negative and therefore $T = |T|$, using the Markov's inequality we obtain that

$$\mathrm{P}(T \geq 10) = \mathrm{P}(|T| \geq 10) \leq \frac{\mathrm{E}\,|T|}{10} = \frac{3}{10} = 0.3.$$

The expectation of the exponentially distributed waiting time is $\mathrm{E}\,T = 1/\lambda = 3$, thus the parameter λ is equal to $1/3$. The exact probability is then

$$\mathrm{P}(T \geq 10) = \int_{10}^{\infty} \lambda e^{-\lambda t}\,\mathrm{d}t = \left[-e^{-\lambda t}\right]_{10}^{\infty} = e^{-\frac{1}{3}\cdot 10} \doteq 0.036.$$

We see that the Markov's inequality provides a fast way to obtain an upper bound of the tail probability.

The Chebyshev's inequality follows from the Markov's inequality:

Theorem 8.3 (– Chebyshev's inequality)**.** *Let X be a random variable with a finite expectation and a finite variance. Then it holds that*

$$\mathrm{P}(|X - \mathrm{E}\,X| \geq \varepsilon) \leq \frac{\operatorname{var} X}{\varepsilon^2} \qquad \text{for all } \varepsilon > 0.$$

Proof. Can be obtained directly, similarly to Markov's inequality (for $(X - \mathrm{E}\,X)^2$), or by inserting $(X - \mathrm{E}\,X)^2$ instead of X and ε^2 instead of a into the Markov's inequality. We obtain

$$\mathrm{P}(|(X - \mathrm{E}\,X)^2| \geq \varepsilon^2) \leq \frac{\mathrm{E}\,|(X - \mathrm{E}\,X)^2|}{\varepsilon^2}.$$

Since $|(X - \mathrm{E}\,X)^2| = (X - \mathrm{E}\,X)^2 = |X - \mathrm{E}\,X|^2$ and a quadratic function is increasing for positive arguments, it holds that

$$(X - \mathrm{E}\,X)^2 \geq \varepsilon^2 \Leftrightarrow |X - \mathrm{E}\,X| \geq \varepsilon.$$

Finally we obtain

$$\mathrm{P}(|X - \mathrm{E}\,X| \geq \varepsilon) \leq \frac{\operatorname{var} X}{\varepsilon^2}.$$

□

Example 8.4 (– waiting for a bus)**.** Suppose that the time T which we spend waiting for a bus is exponentially distributed with the expectation of 3 minutes.

Find an upper bound for the probability that we need to wait for more than 10 minutes using the Chebyshev's inequality. Compare the estimate with the exact probability and with the bound obtained from the Markov's inequality.

Because $T \sim \mathrm{Exp}(\lambda)$ with $\lambda = 1/3$, we get $\mathrm{E}\,T = 1/\lambda = 3$ and $\operatorname{var} T = 1/\lambda^2 = 9$. Using the Chebyshev's inequality we obtain

$$\mathrm{P}(T \geq 10) = \mathrm{P}(T - \mathrm{E}\,T \geq 10 - 3) \leq \mathrm{P}(|T - \mathrm{E}\,T| \geq 7) \leq \frac{\operatorname{var} T}{7^2} = \frac{9}{49} \doteq 0.184.$$

The Markov's inequality provided a bound of $\mathrm{P}(T \geq 10) \leq 0.3$, so the Chebyshev's inequality provides a somewhat closer approximation of the exact probability $\mathrm{P}(T \geq 10) = 0.036$.

Example 8.5 (– waiting for a bus and a tram)**.** Suppose that during our way home, we need to wait for the bus and then for the tram. The time T_1 spent waiting for the bus is exponentially distributed with the expectation of 3 minutes, time T_2 spent waiting for the tram is exponentially distributed with the expectation of 2 minutes. The times are independent.

Find an upper bound for the probability that the total time we spend waiting, $T = T_1 + T_2$ will be more than 15 minutes. Use the Markov's and Chebyshev's inequalities and compare the estimate with the exact probability.

First we find the expectations and variances of T_1, T_2 and T.

$$T_1 \sim \mathrm{Exp}(\lambda), \quad \mathrm{E}\, T_1 = 1/\lambda = 3, \quad \lambda = 1/3, \quad \operatorname{var} T_1 = 1/\lambda^2 = 9.$$

$$T_2 \sim \mathrm{Exp}(\mu), \quad \mathrm{E}\, T_2 = 1/\mu = 2, \quad \mu = 1/2, \quad \operatorname{var} T_2 = 1/\mu^2 = 4.$$

Using the linearity of the expectation and independence of the waiting times we get:

$$\mathrm{E}\, T = \mathrm{E}(T_1 + T_2) \overset{\text{linearity}}{=} \mathrm{E}\, T_1 + \mathrm{E}\, T_2 = 3 + 2 = 5.$$

$$\operatorname{var} T = \operatorname{var}(T_1 + T_2) \overset{\text{independence}}{=} \operatorname{var} T_1 + \operatorname{var} T_2 = 9 + 4 = 13.$$

Using the Markov's inequality we obtain

$$\mathrm{P}(T \geq 15) = \mathrm{P}(|T| \geq 15) \leq \frac{\mathrm{E}\,|T|}{15} = \frac{5}{15} \doteq 0.333.$$

Using the Chebyshev's inequality we obtain

$$\mathrm{P}(T \geq 15) = \mathrm{P}(T - \mathrm{E}\, T \geq 15 - 5) \leq \mathrm{P}(|T - \mathrm{E}\, T| \geq 10) \leq \frac{\operatorname{var} T}{10^2} = \frac{13}{100} = 0.13.$$

The distribution of the sum is considerably more difficult to obtain than when dealing with just one variable. Using convolution we get:

$$\mathrm{P}(T \geq 15) = \int_{15}^{\infty} \int_{0}^{t} \lambda e^{-\lambda u} \mu e^{-\mu \cdot (t-u)} \,\mathrm{d}u \,\mathrm{d}t = \cdots = \frac{\mu e^{-\lambda \cdot 15} - \lambda e^{-\mu \cdot 15}}{\mu - \lambda} \doteq 0.019.$$

The upper bounds obtained using the inequalities seem somewhat imprecise, but they are easy to compute, using only expectations and variances.

8.3 Laws of large numbers

8.3.1 Weak law of large numbers

First we compute the expected value and variance of the mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

where $X_1, \ldots, X_n$ are *i.i.d.* random variables with $\mathrm{E}\, X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$.

- *Expected value*

$$\mathrm{E}\,\bar{X}_n = \mathrm{E}\,\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\,\mathrm{E}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\,X_i = \frac{n\mu}{n} = \mu.$$

- *Variance*

$$\operatorname{var}\bar{X}_n = \operatorname{var}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n^2}\operatorname{var}\sum_{i=1}^{n} X_i = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var} X_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

We used the linearity of the expectation in the first part and the behavior of the variance of a sum of independent random variables in the second part.

By inserting $\bar{X}_n$ into the Chebyshev's inequality we obtain the *weak law of large numbers*:

Theorem 8.6 (– weak law of large numbers)**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with finite expectation $\mathrm{E}\,X_i = \mu$ and finite variance σ^2. Then $\bar{X}_n$ converges to μ* in probability

$$\bar{X}_n \xrightarrow{P} \mu \qquad \textit{for} \quad n \to \infty.$$

This means that for all $\varepsilon > 0$ it holds that $\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| \geq \varepsilon) = 0$.

Proof. We use the Chebyshev's inequality for the arithmetic mean $\bar{X}_n$:

$$0 \leq \mathrm{P}(|\bar{X}_n - \mathrm{E}\,\bar{X}_n| \geq \varepsilon) = \mathrm{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\operatorname{var}\bar{X}_n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \qquad \text{for } n \to \infty.$$

The statement follows from the sandwich theorem. □

8.3.2 Strong law of large numbers

Theorem 8.7 (– strong law of large numbers (SLLN))**.** *Let $X_1, X_2, \dots$ be i.i.d. random variables with expected value $\mathrm{E}\,X_i = \mu$ (not necessarily finite). Then $\bar{X}_n$ converges to μ* almost surely *(with probability* 1*)*

$$\bar{X}_n \xrightarrow{a.s.} \mu \qquad \textit{for} \quad n \to \infty.$$

It means that the set where $X_n(\omega)$ converges as a numerical sequence has probability 1*:*

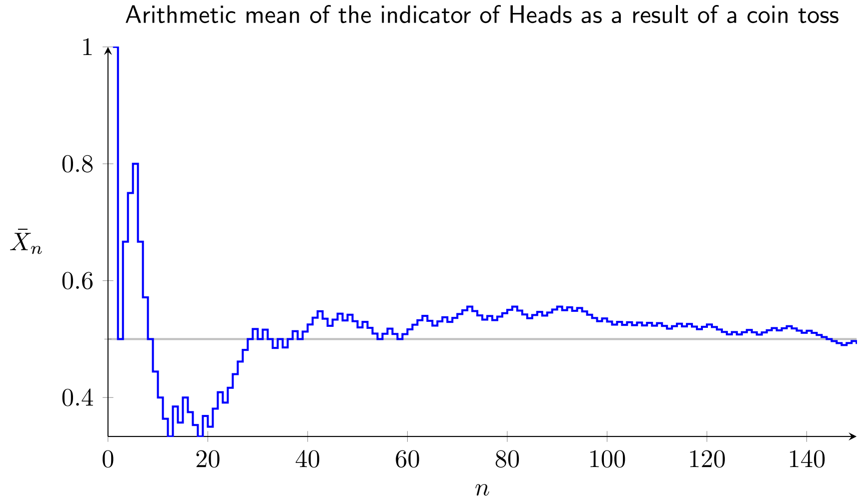
$$\mathrm{P}(\{\omega \in \Omega \colon \bar{X}_n(\omega) \to \mu \textit{ for } n \to \infty\}) = 1.$$

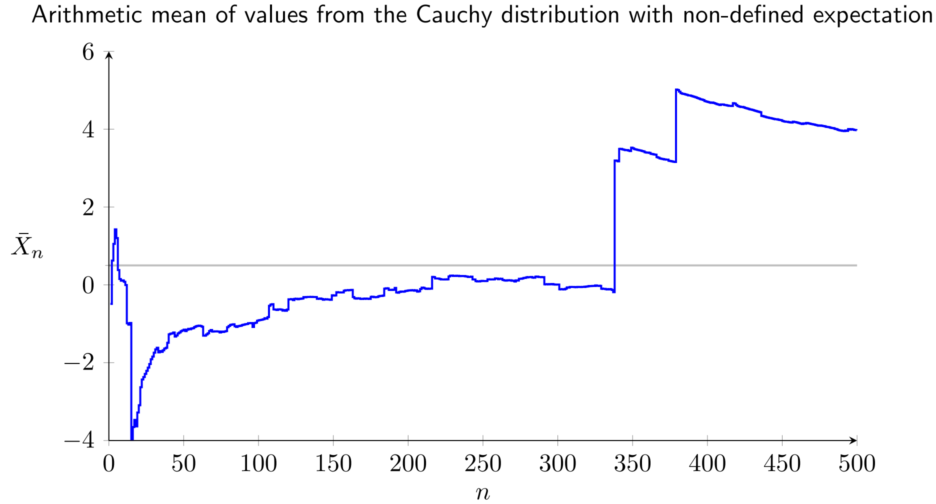
Proof. Considerably more difficult, see bibliography. □

In what sense is this law of large numbers "stronger"?

- It is enough to consider the existence of the expected value. Moreover, it can be infinite and the variance as well.

- Convergence almost surely *implies* convergence in probability.





Recall that for the arithmetic mean $\bar{X}_n$ of *i.i.d.* random variables with $\operatorname{E} X_i = \mu$ and $\operatorname{var} X_i = \sigma^2$ we have

$$\operatorname{E} \bar{X}_n = \mu, \qquad \operatorname{var} \bar{X}_n = \frac{\sigma^2}{n}.$$

Let us now find the characteristics of the sum:

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

- *Expected value*

$$\operatorname{E} S_n = \operatorname{E} \sum_{i=1}^{n} X_i \overset{\text{linearity}}{=} \sum_{i=1}^{n} \operatorname{E} X_i = \sum_{i=1}^{n} \mu = n\mu.$$

- *Variance*

$$\operatorname{var} S_n = \operatorname{var} \sum_{i=1}^{n} X_i \overset{\text{independence}}{=} \sum_{i=1}^{n} \operatorname{var} X_i = \sum_{i=1}^{n} \sigma^2 = n\sigma^2.$$

We can alternatively obtain this properties if we realize that $S_n = n \cdot \bar{X}_n$ and apply the expectation and variance.

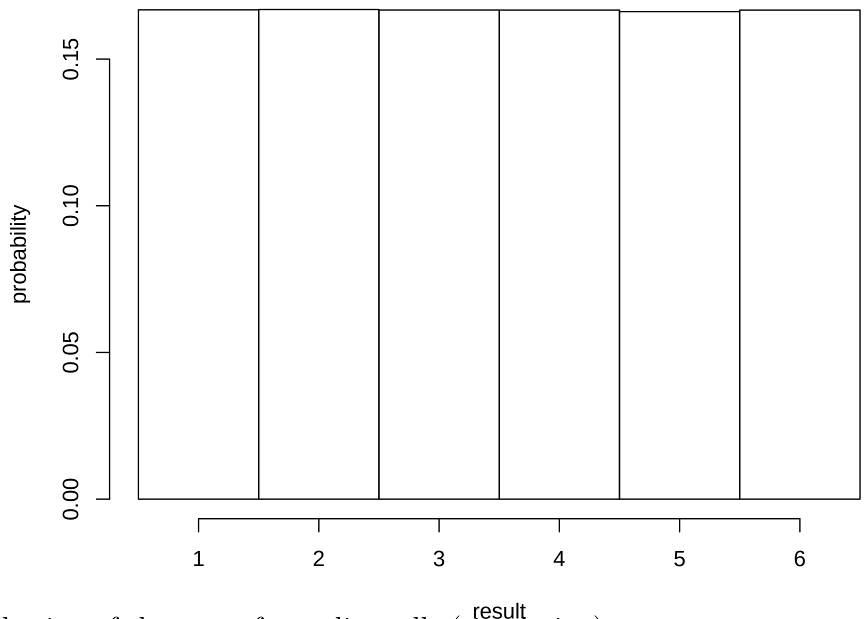
8.4 Central limit theorem

Laws of large numbers deal with convergence of the mean to the expected value. For large n, the *mean represents a reasonable approximation* of the expected value. In other words, the expectation is the ideal average of an infinite number of repeated experiments.

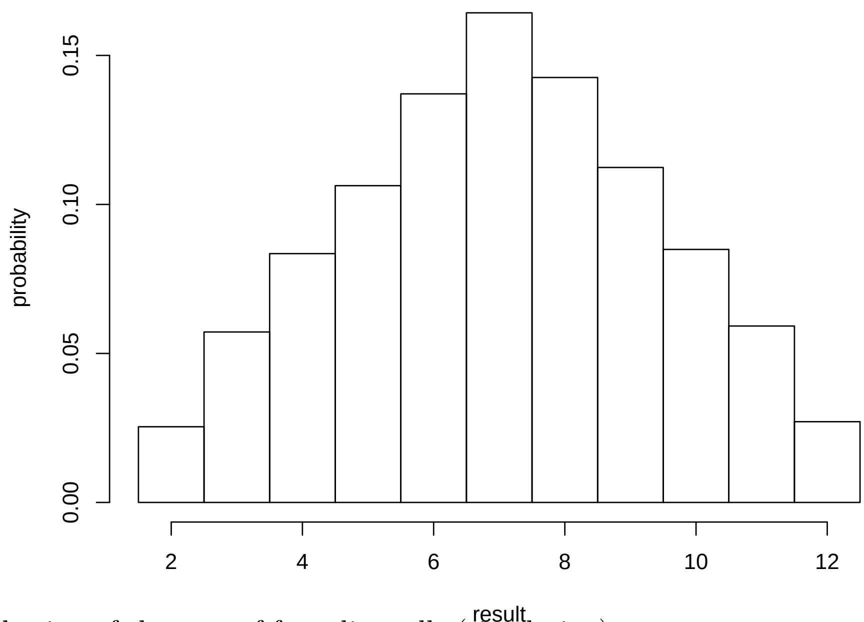
However, what is the distribution of the mean or the sum as a random variable?

Central limit theorem (CLT) says that under particular circumstances the distribution of the mean or a sum can be approximated by the *normal* distribution.

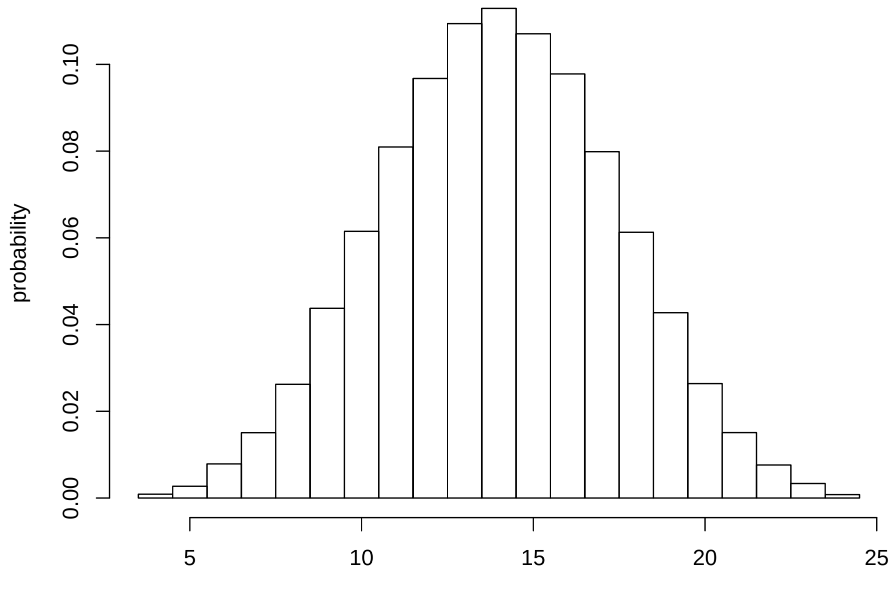
Distribution of one die roll (simulation).



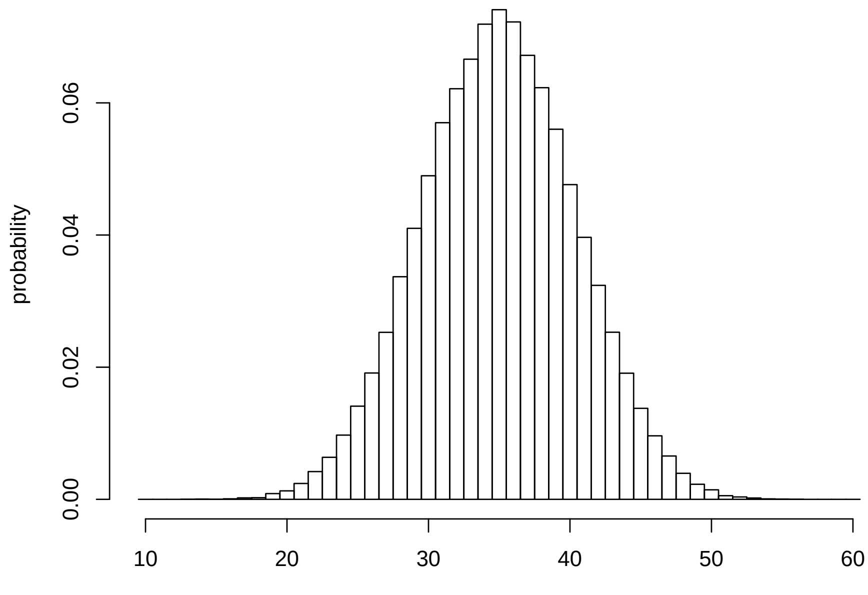
Distribution of the sum of two dice rolls (simulation).



Distribution of the sum of four dice rolls (simulation).



Distribution of the sum of ten dice rolls (simulation).



For understanding the statement of central limit theorem we need to define the convergence in distribution.

Definition 8.8. Let $X_1, X_2, \dots$ be a sequence of random variables with distribution functions $F_{X_1}, F_{X_2}, \dots$ and X be a random variable with a distribution function F_X. We say that variables X_i converge to X *in distribution*,

$$X_n \xrightarrow{\mathcal{D}} X \quad \text{or} \quad X_n \xrightarrow{\mathcal{L}} X \qquad \text{for } n \to \infty,$$

if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

in all continuity points of the distribution function F_X.

When X has a continuous distribution, for *large* n we can consider:

- $\mathrm{P}(X_n \leq x) = F_{X_n}(x) \approx F_X(x) = \mathrm{P}(X \leq x)$,
- $\mathrm{P}(X_n > x) = 1 - F_{X_n}(x) \approx 1 - F_X(x) = \mathrm{P}(X > x)$,
- $\mathrm{P}(a < X_n \leq b) = F_{X_n}(b) - F_{X_n}(a) \approx F_X(b) - F_X(a) = \mathrm{P}(a < X \leq b)$.

Theorem 8.9 (– Central limit theorem (CLT))**.** *Let* $X_1, X_2, \dots$ *be a sequence of i.i.d. random variables with finite expectations* $\mathrm{E}\, X_i = \mu$ *and finite variances* $\operatorname{var} X_i = \sigma^2 > 0$*. Then*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1) \qquad \text{for } \; n \to \infty.$$

Similarly

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1) \qquad \text{for } \; n \to \infty.$$

Proof. See bibliography. □

The symbol $\mathrm{N}(0,1)$ stands for a variable with the *standard normal distribution.*
Recall that

$$\mathrm{E}\, S_n = n \cdot \mu, \qquad \mathrm{E}\, \bar{X}_n = \mu,$$
$$\operatorname{var} S_n = n \cdot \sigma^2, \qquad \operatorname{var} \bar{X}_n = \sigma^2/n.$$

The central limit theorem states that if we take either the *standardised mean* or the *standardised sum*

$$Z_n = \frac{S_n - \mathrm{E}\, S_n}{\sqrt{\operatorname{var} S_n}} = \frac{\bar{X}_n - \mathrm{E}\, \bar{X}_n}{\sqrt{\operatorname{var} \bar{X}_n}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}},$$

the resulting variable *converges to the standard normal distribution.* For any $z \in \mathbb{R}$:

$$\mathrm{P}(Z_n \leq z) \xrightarrow{n\to\infty} \mathrm{P}(Z \leq z) = \Phi(z).$$

This allows us to effectively approximate the behavior of sums or means for large n. The theorem can be used regardless of the original distribution, even if it is unknown. However, the closer to the normal distribution, the more precise is the approximation.

CLT allows us to express probabilities of types $\mathrm{P}(\bar{X}_n < x)$, $\mathrm{P}(\bar{X}_n > x)$, etc. by means of the distribution function Φ of the *standard normal* distribution

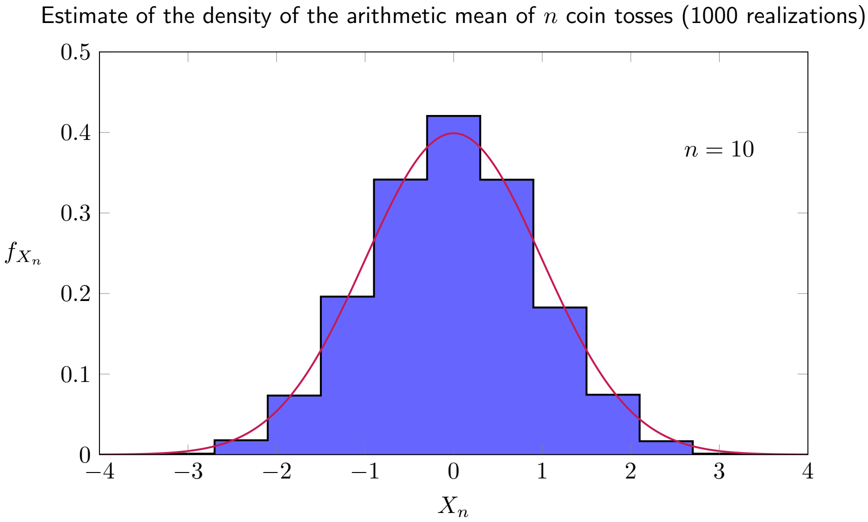
$$\Phi(z) = \mathrm{P}(Z \leq z) \quad \text{for} \quad Z \sim \mathrm{N}(0,1).$$

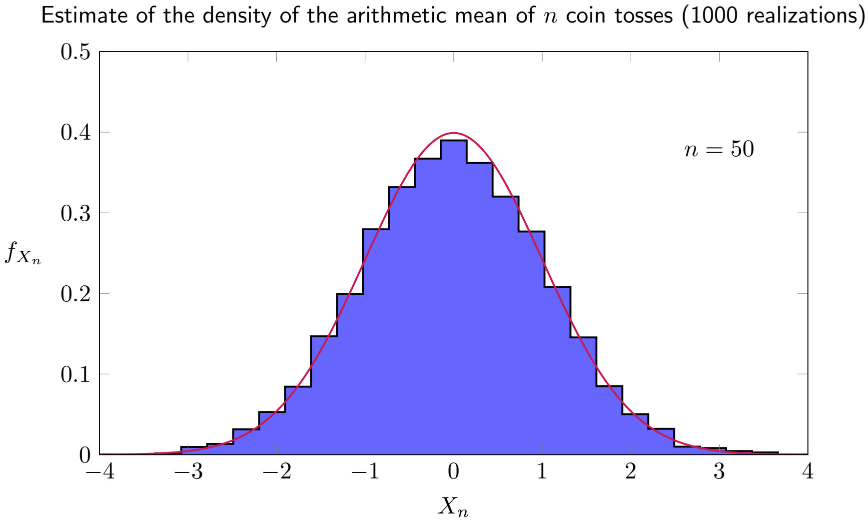
The advantage is that the *values* of Φ are *tabulated.*

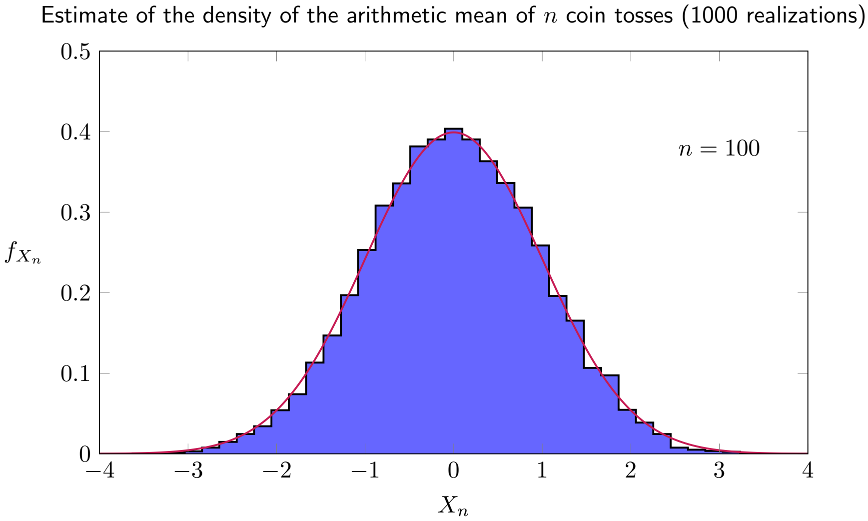
Another variants of the statement:

- $\dfrac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \mathrm{N}(0,1)$,
- $\bar{X}_n \overset{\text{approx.}}{\sim} \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ for large n,

- $S_n \overset{\text{approx.}}{\sim} \mathrm{N}(n\mu, n\sigma^2)$ for large n..







Example 8.10. *What is the probability that in 1000 independent tosses with a coin we get more than 525 times Heads?*

Let X_i be an indicator variable denoting, whether in the i-th toss Heads appears ($X_i = 1$) or not ($X_i = 0$). We want to calculate $\mathrm{P}(S_{1000} > 525)$. The number of successes (Heads) among n attempts (rolls) follows the binomial distribution $\mathrm{Binom}(n, p)$ with $n = 1000$ and $p = 1/2$. Computing the probability directly would be very demanding.

Instead of using the binomial distribution we use CLT. For tossing a coin it holds that $\mathrm{E}\,X_i = p = 1/2$ and $\mathrm{var}\,X_i = p(1-p) = 1/4$. For the sum it holds that $\mathrm{E}\,S_{1000} = np = 500$, and $\mathrm{var}\,S_{1000} = np(1-p) = 250$. We get

$$\begin{aligned}\mathrm{P}\left(\sum_{i=1}^{1000} X_i > 525\right) &= \mathrm{P}(S_{1000} - 500 > 525 - 500) = \mathrm{P}\left(\frac{S_{1000} - 500}{\sqrt{250}} > \frac{25}{\sqrt{250}}\right) = \\ &= 1 - \mathrm{P}\left(\frac{S_{1000} - 500}{\sqrt{250}} \le \frac{5}{\sqrt{10}}\right) \approx 1 - \Phi\left(\frac{5}{\sqrt{10}}\right) = 1 - \Phi(1.58) = 0.0571.\end{aligned}$$

Example 8.11 (– CLT vs. Markov's and Chebyshev's inequalities)**.** *Suppose that we operate a cargo lift with a maximum load of 600 kg. We need to lift 25 packages, each having an expected weight of 20 kilograms and a standard deviation of 8 kilograms. What is the probability that the lift will be overloaded? Use the Markov's and Chebyshev's inequalities and CLT.*

Let X_i be the weight of the i-th package. We have

$$\mathrm{E}\,X_i = \mu = 20 \qquad \text{and} \qquad \mathrm{var}\,X_i = \sigma^2 = 8^2 = 64.$$

The total weight of all $n = 25$ packages is $S_n = \sum_{i=1}^{n} X_i$, with

$$\mathrm{E}\,S_n = n\mu = 25 \cdot 20 = 500 \qquad \text{and} \qquad \mathrm{var}\,S_n = n\sigma^2 = 25 \cdot 64 = 1600.$$

The weights are surely non-negative, thus the Markov's inequality gives us:

$$\mathrm{P}(S_n \ge 600) = \mathrm{P}(|S_n| \ge 600) \le \frac{E|S_n|}{600} = \frac{500}{600} \doteq 0.83.$$

Using the Chebyshev's inequality we get

$$\mathrm{P}(S_n \ge 600) \le \mathrm{P}(|S_n - \mathrm{E}\,Sn| \ge 600 - 500) \le \frac{\mathrm{var}\,S_n}{100^2} = \frac{1600}{10000} = 0.16.$$

CLT gives us

$$\begin{aligned}\mathrm{P}(S_n > 600) &= 1 - \mathrm{P}(S_n \le 600) = 1 - \mathrm{P}\left(\frac{S_n - \mathrm{E}\,S_n}{\sqrt{\mathrm{var}\,S_n}} \le \frac{600 - 500}{\sqrt{1600}}\right) \\ &= 1 - \mathrm{P}\left(Z_n \le \frac{100}{40}\right) \approx 1 - \Phi(2.5) = 0.0062.\end{aligned}$$

We were able to use the inequalities and the central limit theorem to approximate the probability, even if we didn't know the distribution of the weights.