6.3. Continuous Probability Distributions

Uniform distribution

X is called uniformly distributed on the interval $[a, b]$, denoted as $X \sim U(a, b)$ with $a < b$ (tasaisesti jakaunut välillä $[a, b]$, $X \sim \mathrm{Tas}(a, b)$), if the probability is the same within all subintervals of the same size. The density function of a uniform distribution in the interval $[a, b]$ is:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \le x \le b \\ 0, & \text{otherwise.} \end{cases}$$

The cumulative distribution function is:

$$F(x) = \int_a^x f(t)\,dt = \begin{cases} 0, & \text{for } x < a \\ \frac{x-a}{b-a}, & \text{for } a \le x \le b \\ 1, & \text{for } x > b \end{cases}$$

$$\mu = \mathsf{E}\,[X] = \frac{1}{2}(a+b)$$

$$\sigma^2 = \mathsf{Var}\,[X] = \frac{1}{12}(b-a)^2$$

The Normal Distribution (Normalijakauma)

The normal distribution is the most important continuous probability distribution in statistics, because the central limit theorem (to be discussed shortly) states that many random variables may be approximated as normally distributed random variables.

X is said to be normally distributed with parameters μ and σ^2 (normaalijakautuneeksi parametrein μ ja σ^2), if its density is of the form

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

It is denoted $X \sim N(\mu, \sigma^2)$.

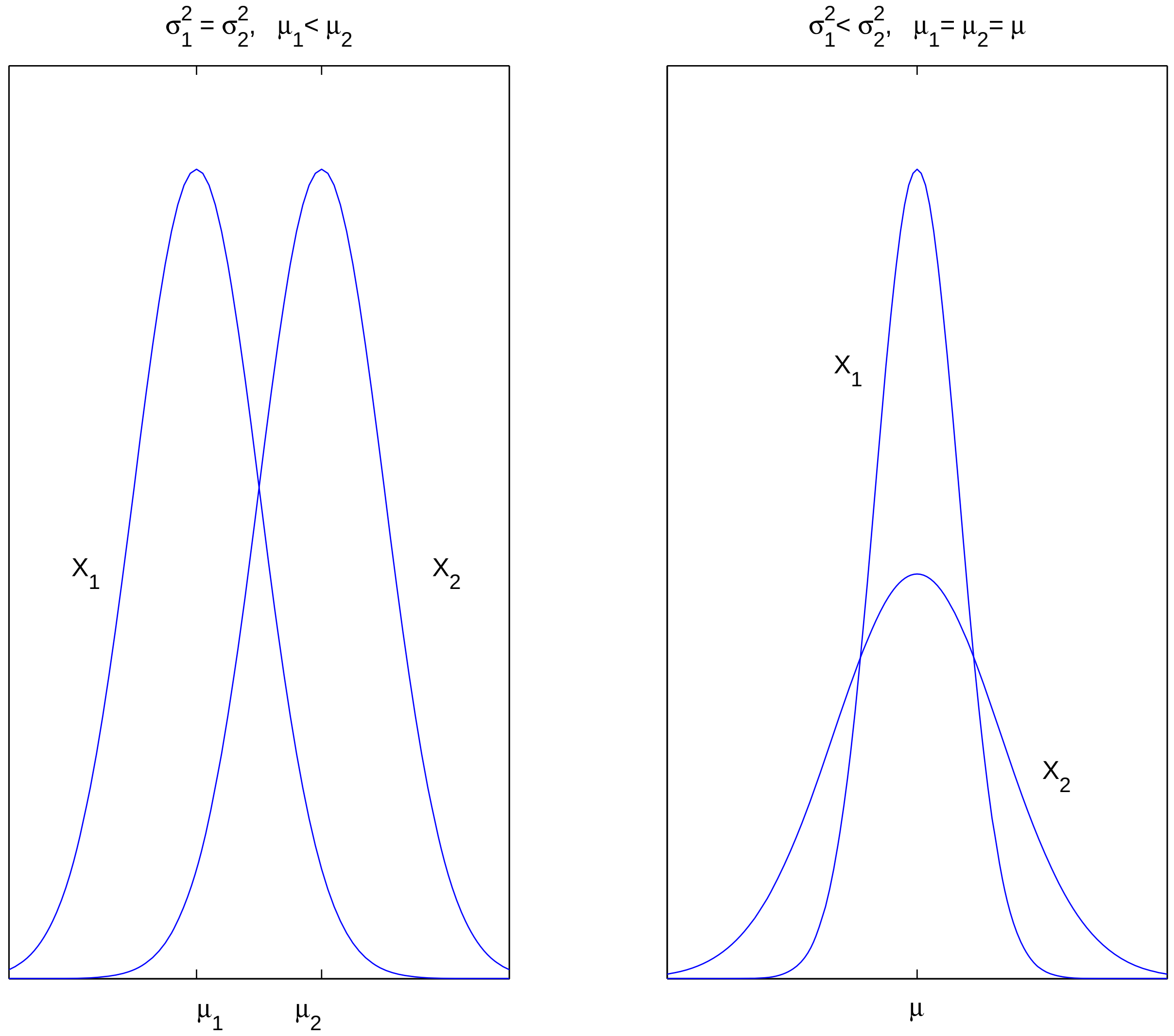
The expected value of the normal distribution is

$$\mathsf{E}\,[X] = \mu$$

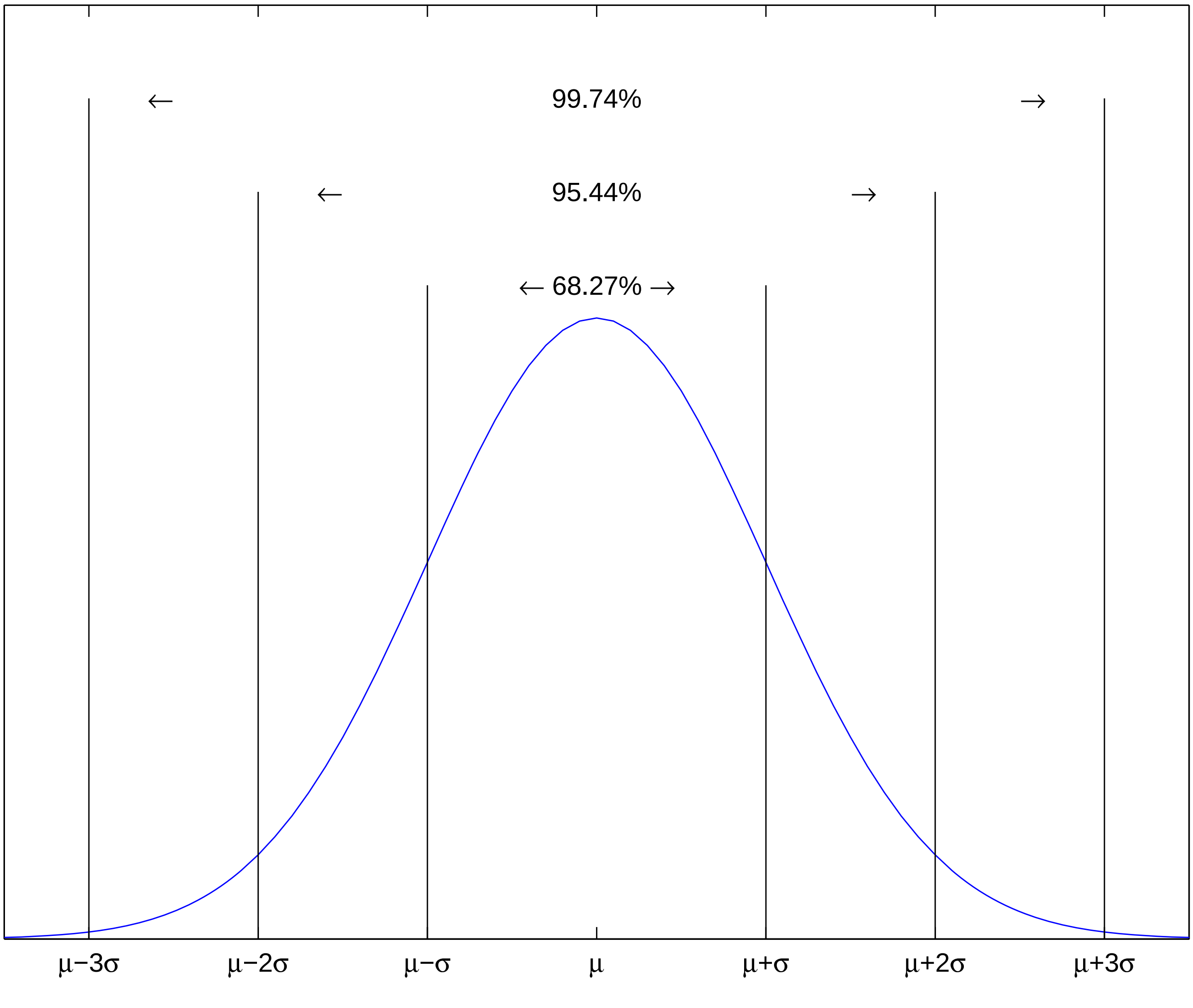
and its variance

$$\mathsf{Var}\,[X] = \sigma^2.$$

The normal distribution is a symmetric unimodal distribution, which implies that its expected value μ is also its mode and its median. Its density function approaches the x-axis at both sides of μ. The plots below illustrate the impact of changing the values of the expected value μ and the variance σ^2:



Evaluating the integral under the density function of the normal distribution provides the justification for the empirical rule according to which for symmetric unimodal distributions there are about 68% of the observations within one standard deviation around the mean, 95% within two standard deviations around the mean, and 99% within three standard deviations around the mean:



Sums of Normal Random Variables

Sums of normal random variables are also normally distributed themselves.

Recall from our discussion of the mean and the variance that when $X_1, X_2, \ldots, X_n$ are *independent* random variables with expected values $\mu_1, \mu_2, \ldots, \mu_n$ and respective variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then their weighted sum $\underline{Y := \sum_{i=1}^n a_i X_i}$ will have an expected value $\underline{\mu = \sum_{i=1}^n a_i \mu_i}$ and variance $\underline{\sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2}$. Hence, for (weighted) sums of *independent* normally distributed random variables:

$$X_i \sim N(\mu_i, \sigma_i^2) \quad \Rightarrow \quad Y \sim N(\mu, \sigma^2)$$

with notation as defined above, in particular

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right), \quad \bar{X} \sim N\left(\frac{\sum \mu_i}{n}, \frac{\sum \sigma_i^2}{n^2}\right).$$

Setting $X_1 = X/\sigma_X$ and $X_2 = -\mu_X/\sigma_X$ yields

$$Z_X := \frac{X - \mu_X}{\sigma_X} \sim N(0,1), \qquad \text{because}$$

$$E\left(\frac{X}{\sigma_X} - \frac{\mu_X}{\sigma_X}\right) = \frac{E(X)}{\sigma_X} - \frac{\mu_X}{\sigma_X} = 0, \quad V\left(\frac{X}{\sigma_X} - \frac{\mu_X}{\sigma_X}\right) = \frac{V(X)}{\sigma_X^2} = 1.$$

The Standard Normal Distribution
(Standardoitu normaalijakauma)

Defining

$$Z = \frac{X - \mu}{\sigma}$$

yields the standard normal distribution with zero mean and unit variance, denoted by $Z \sim N(0, 1)$. Its density function f is usually denoted by $\phi(z)$ and the cumulative distribution function F by $\Phi(z)$. That is

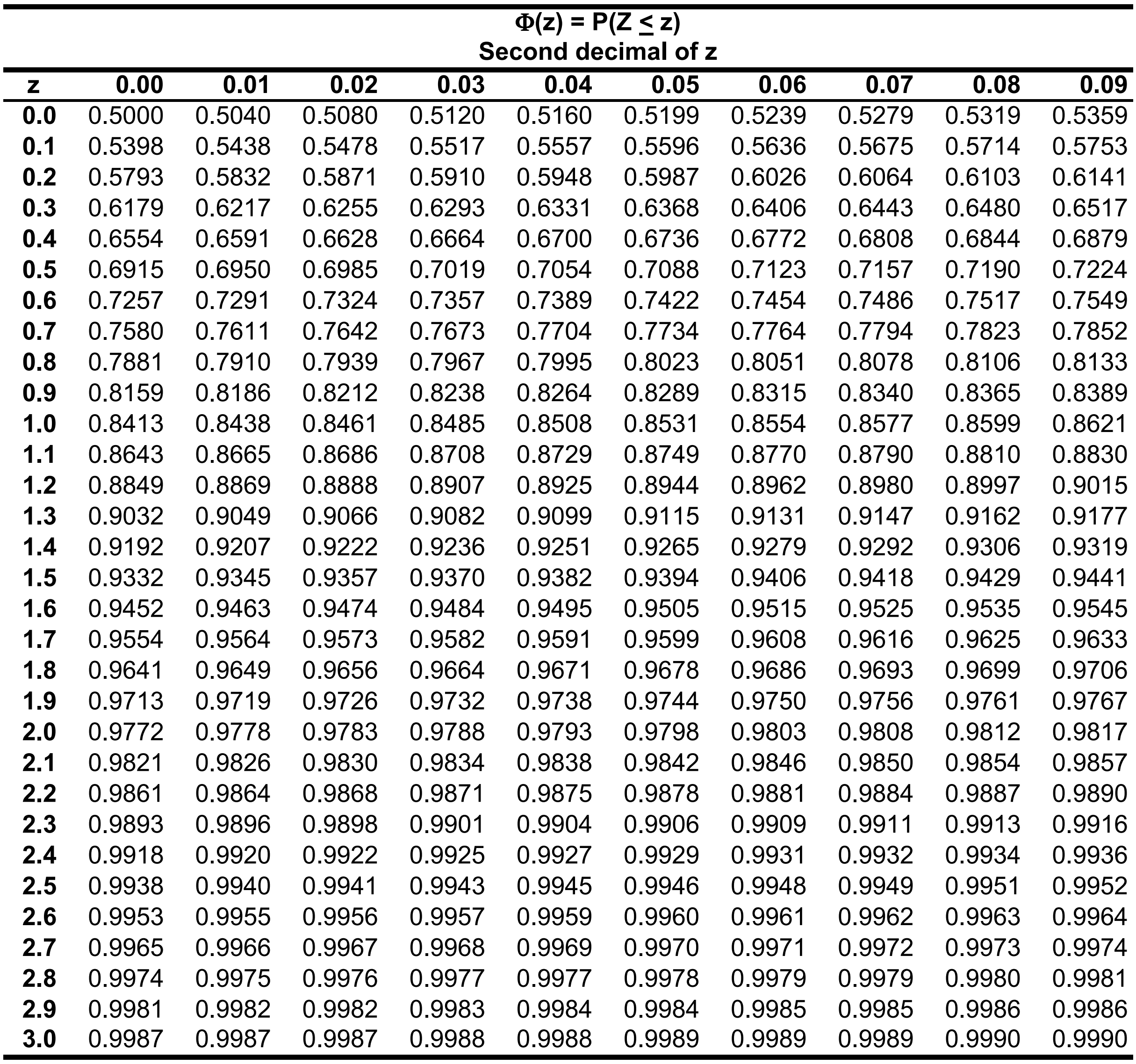
$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty$$

and

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \phi(t)\, dt.$$

Values of $\Phi(z)$ are tabulated (see next page) and implemented in Excel as the function NORMSDIST. For the reverse problem of finding z for given values of $\Phi(z)$ one may apply the excel function NORMSINV.

Example. $P(Z \leq 1.34) = \Phi(1.34) = 0.9099$.



Φ(z) = P(Z ≤ z)

Second decimal of z

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Φ is usually tabulated only for positive values of z, because due to the symmetry of ϕ around zero:

$$\Phi(-z) = P(Z \leq -z) = P(Z > z) = 1 - P(Z \leq z) = 1 - \Phi(z).$$

Example. $Z \sim N(0,1)$, then:

$$P(Z \leq -1) = \Phi(-1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587.$$

With the tabulated values we can calculate probabilities of any $X \sim N(\mu, \sigma^2)$ with given parameter values of μ and σ^2. The cumulative distribution function of X has the following relation to Φ:

$$\begin{aligned} F(x) = P(X \leq x) &= P\left(\tfrac{X-\mu}{\sigma} \leq \tfrac{x-\mu}{\sigma}\right) \\ &= P\left(Z \leq \tfrac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\tfrac{x-\mu}{\sigma}\right). \end{aligned}$$

That is, if $X \sim N(\mu, \sigma^2)$ then

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \Phi(z).$$

<u>Example.</u> $X \sim N(3,4) = N(3,2^2)$

a) $P(X \leq 0) = P\left(\dfrac{X-3}{2} \leq \dfrac{0-3}{2}\right) = P(Z \leq -1.5)$
$= \Phi(-1.5) = 1 - \Phi(1.5) = 1 - 0.9332$
$= 0.0668.$

b) $P(X > 1) = 1 - P(X \leq 1) = 1 - P(Z \leq (1-3)/2)$
$= 1 - \Phi(-1) = 1 - [1 - \Phi(1)] = \Phi(1)$
$= 0.8413.$

c) $P(X \leq a) = 0.8$. What is a?

$$P(X \leq a) = P\left(Z \leq \frac{a-3}{2}\right) = \Phi\left(\frac{a-3}{2}\right) = 0.8$$

table: $0.8 \approx \Phi(0.84) \ \Rightarrow \ \dfrac{a-3}{2} = 0.84 \ \Leftrightarrow \ a = 4.68.$

<u>Excel:</u> NORMSINV(0.8)=0.8416≈0.84.

Example. A factory produces lamps, the lifetimes of which follow the normal distribution. The average lifetime of a lamp is 800 hours with a standard deviation of 40 hours. What is the probability that a randomly selected lamp will last for at least 700 hours but not more than 850 hours?

Denote $X =$ lamps lifetime $\sim N(800, 40^2)$.

Then:

$$
\begin{aligned}
& P\,(700 \leq X \leq 850) \\
=& P\left(\frac{700-800}{40} \leq Z \leq \frac{850-800}{40}\right) \\
=& P(-2.5 \leq Z \leq 1.25) \\
=& P(Z \leq 1.25) - P(Z \leq -2.5) \\
=& \Phi(1.25) - \Phi(-2.5) \\
=& 0.8944 - \underbrace{(1-0.9938)}_{0.0062} \\
=& 0.8882.
\end{aligned}
$$

The Central Limit Theorem
(Keskeinen raja-arvolause)

Let $X_1, X_2, \ldots, X_n$ be independent random variables (not necessarily normal, may come from different distributions) with $\mathsf{E}[X_i] = \mu$ and $\mathsf{Var}[X_i] = \sigma^2$ (both finite), $i = 1, \ldots, n$. Let

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

Then by independence:

$$\mathsf{E}[\bar{X}_n] = \mu, \text{ and } \mathsf{Var}[\bar{X}_n] = \frac{\sigma^2}{n}.$$

Furthermore as $n \to \infty$:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to Z \sim N(0,1).$$

We say that Z_n is asymptotically $N(0,1)$ distributed, denoted as $Z_n \overset{as.}{\sim} N(0,1)$, which implies both

$$\bar{X}_n \overset{as.}{\sim} N(\mu, \sigma^2/n), \text{ and } \sum_{i=1}^{n} X_i \overset{as.}{\sim} N(n\mu, n\sigma^2).$$

Note. For most practical applications the normal approximation works already for $n \geq 30$.

Normal Distribution as an Approximation of the Binomial and Poisson Distribution

Recall that the Binomial Distribution Bin(n, p) describes the sum of n independent Bernoulli random variables with success probability p, expected value p, and variance pq ($q = 1 - p$):

$$X_i \sim \text{Ber}(p) \quad \Rightarrow \quad \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p).$$

Applying the central limit theorem instead with $\mu = p$ and $\sigma^2 = pq$ yields

$$\sum_{i=1}^{n} X_i \overset{as.}{\sim} N(np, npq),$$

that is,

$$\text{Bin}(n, p) \overset{n\to\infty}{\longrightarrow} N(np, npq).$$

This approximation works reasonable well when both $np > 5$ and $nq > 5$ are satisfied.

<u>Note.</u> We may also approximate Poisson distributed random variables as normally distributed for large λ as Poi$(\lambda) \approx N(\lambda, \lambda)$.

Example.

Consider tossing a coin 64 times. What is the probability of throwing up to 25 heads?

$$X := \#\text{ heads up} \sim \text{Bin}(64, \frac{1}{2}) \overset{as.}{\sim} N(\overbrace{32}^{np}, \overbrace{16}^{npq}).$$

Exact calculation:

$$P(X \le 25) = \sum_{x=0}^{25} \binom{64}{x} 0.5^x 0.5^{64-x} = 0.0517$$

Normal approximation:

$$\begin{aligned} P(X \le 25) &= P\left(\frac{X-32}{4} \le \frac{25-32}{4}\right) \\ &= P(Z \le -1.75) = \Phi(-1.75) \\ &= 1 - \Phi(1.75) = 1 - 0.9599 \\ &= 0.0401. \end{aligned}$$

The approximation improves by applying the continuity correction (jatkuvuuskorjaus), that is, subtracting 0.5 from any left limit and adding 0.5 to any right limit for events of the form $\{a \le X \le b\}$:

$$P(X \le 25.5) = 1 - \Phi(1.625) = 1 - 0.9479 = 0.0521.$$

Exponential distribution (Eksponenttijakauma)

The exponential distribution is applied among others for modelling the randomly varying duration of events or the waiting time between two random events. It is the continuous limit of the geometric distibution.

X is exponentially distributed with parameter θ (eksponentiaalisesti jakautunut parametrilla θ), denoted as $X \sim \text{Exp}(\theta)$, if its density function is

$$f(x) = \theta e^{-\theta x}, \quad x \geq 0,\ \theta > 0.$$

The expected value and variance are

$$E(X) = \frac{1}{\theta} \quad \text{and} \quad V(X) = \frac{1}{\theta^2}.$$

The cumulative distribution function is

$$F(x) = \int_0^x f(t)dt = 1 - e^{-\theta x}, \quad x \geq 0.$$

Example. The duration of calls arriving at a customers service are exponentially distributed. The average duration of a call is 4 minutes. What is the probability that the next call will last less than 2 minutes?

$X =$ duration of call $\sim$ Exp(1/4)
$\Rightarrow \; P(X \leq 2) = F(2) = 1 - e^{-\frac{1}{4}\cdot 2} \approx 0.39$.

The exponential distribution is the only continuous distribution which has the so called memoryless property (unohtavaisuusominaisuus):

$$P(X > x + x_0 | X > x_0) = P(X > x).$$

The geometric distribution is also memoryless, but it is discrete (recall that we could interpret the number of lottery games needed for a lottery winning as a waiting time).

Example. (continued)
If a call to the customers service has lasted for 1 minute already, the probability that it will last for at least another 2 minutes is

$$P(X > 2 + 1 | X > 1) = P(X > 2) \approx 1 - 0.39 = 0.61.$$