Session 2

Statistical Analysis (31007)

Yuri Balasanov

MSCA, University of Chicago

Disclaimer

The content of these lectures, any comments, opinions and materials are put together by the author especially for the course Statistical Analysis, they are sole responsibilities of the author, but not of the author's employers or clients.
The author cannot be held responsible for any material damage as a result of use of the materials presented in this document or in this course.
For any inquiries contact the author, Yuri Balasanov, at ybalasan@uchicago.edu or yuri.balasanov@research-soft.com

Outline of This Session

- Simulation of random variables
- Uniform distribution.
- Pseudo-random variables: generating and testing
- Discrete random variables:
  - Binomial distribution
  - Negative binomial distribution
  - Poisson distribution
- Continuous random variables
  - Exponential distribution
  - Gaussian distribution
- Simulation of linear model

Main Text:

Randall Pruim. 2011. Foundations and Applications of Statistics. An Introduction Using R. American Mathematical Society.

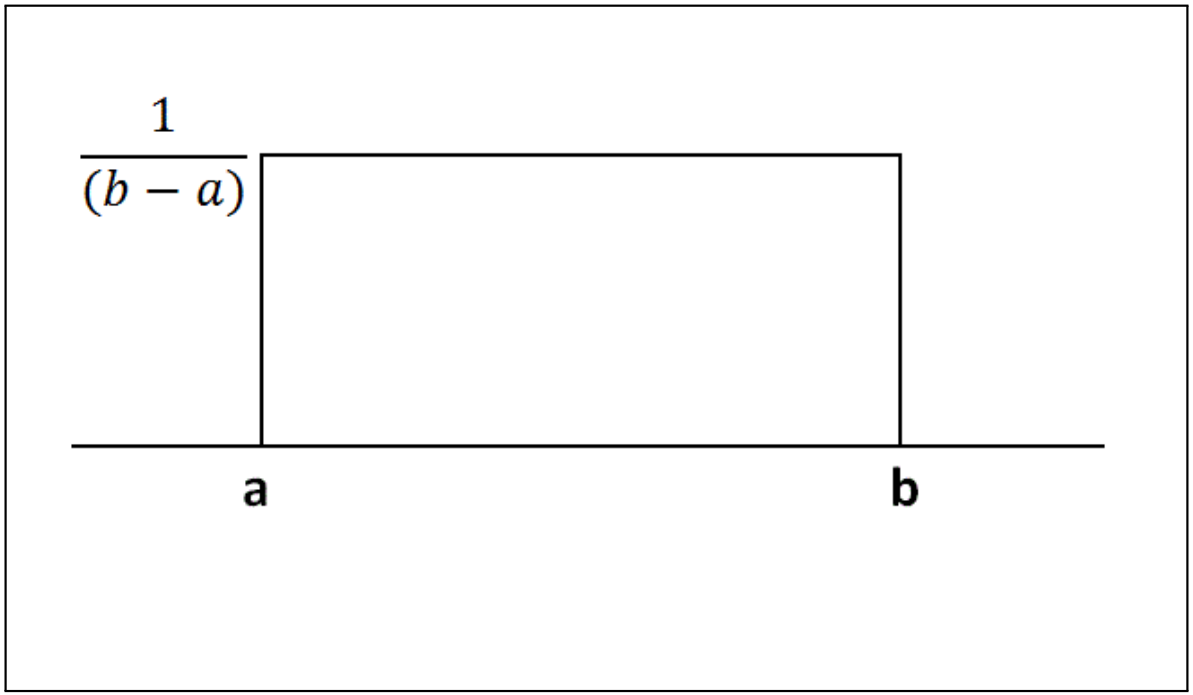
Steps for Simulating Random Variables

Simulation of random variables from any distribution is usually done in two steps:

1. First generate **random number**, i.e. realization of uniformly distributed random variable on $[0, 1]$
2. Then transform random number into a realization of random variable from the distribution of choice

Generating pseudo-random numbers of high quality is a challenging fundamental mathematical problem related to number theory and theory of complexity.
Random numbers play important role in such fields as numerical analysis (Monte Carlo method), analysis of complex systems, cryptography.
For additional information about generating random numbers see Harry Perros. Computer Simulation Techniques: The definitive introduction! 2009. http://www4.ncsu.edu/~hp//simulation.pdf

Uniform Distribution I

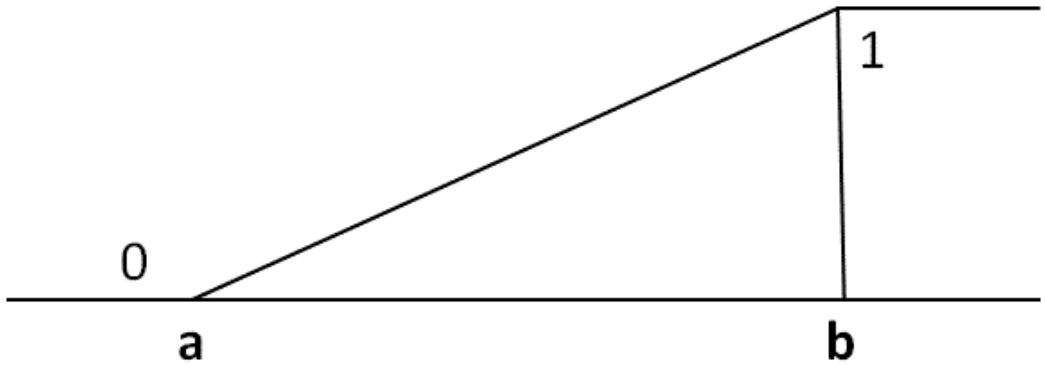


Definition

A continuous **uniform random variable** $X \sim Unif\,(a, b)$ on the interval $[a, b]$ is the variable with probability density function

$$f_X\,(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b]\,, \\ 0 & \text{otherwise.} \end{cases}$$

Uniform Distribution II



Cumulative distribution function of $Unif\,(a, b)$ is

$$F_X\,(x; a, b) = \begin{cases} 0, & x < a, \\ x, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

True and Pseudo-Random Numbers

There are two ways of generating random numbers:

1. Using a physical phenomenon as a source of randomness generates so called **true random numbers**. Examples of true random numbers are elapsed times between emissions of particles during radioactive decay, thermal noise from semiconductor, etc. True random numbers are especially useful in cryptography thanks to their completely unpredictable nature.
2. Using a mathematical algorithm produces pseudo-random numbers. Technically pseudo-random numbers are not random because they are calculated by some algorithm and are not unpredictable. Each algorithmic pseudo-random numbers generator can produce a sample of some finite length after which it repeats the cycle. But if the algorithm has *high complexity* (mathematical term) it generates a long enough sample of numbers that look like random. Their additional advantage is reproducibility.

How Pseudo-Random Numbers are Generated?

Pseudo-random number must be:

- Uniformly distributed
- Statistically independent
- Reproducible
- Non-repeating for some reasonable length

Historically the first generator was introduced by von Neuman: mid-square method. He squared the previous number and extracted the digits in the middle.
Since then a number of other generators have been developed: the congruential, the Tausworthe, the lagged Fibonacci algorithms to name a few.
For example, the lagged Fibonacci algorithm uses the sequence of Fibonacci numbers: $0, 1, 1, 2, 3, 5, 8, 13, 21, \ldots$ to generate pseudo-random numbers as $x_n = x_{n-j} + x_{n-k} \pmod{m}$, $0 < j < k$.

How Pseudo-Random Numbers are Tested?

Usually, the following tests are conducted:

- **Frequency (monobit) test.** Concatenate the generated random numbers in a string of bits, turn zeros into $-1's$; create $S_n = \frac{|X_1+X_2+...+X_n|}{\sqrt{2n}}$, where each $X_i = \pm 1$. Compute P-value as

$$P = \text{erfc}\left(S_n\right) = \frac{\Gamma\left(\frac{1}{2}, S_n^2\right)}{\sqrt{\pi}}.$$

  If P is small (≤ 0.01) the generator fails the test.
- **Serial test.** There are 2^k different combinations of k bits. Serial test determins if they all appear with the same frequency.
- **Autocorrelation test.** If a sequence of n bits is random it will be different from another sequence of n bits.
- **Runs test.** This is the test of independency
- **Chi-Square test for goodness of fit**

Discrete Random Variables I

Bernoulli Trials

Flipping a coin is a particular case of a more general random experiment of flipping unfair coin called Bernoulli trial.

Definition

A single Bernoulli trial is a random experiment with two elementary outcomes $\Omega = \{\text{"Success"},\text{"Failure"}\}$,

$$\begin{aligned} \mathbb{P}\{\text{"Success"}\} &= p, \\ \mathbb{P}\{\text{"Failure"}\} &= 1-p. \end{aligned}$$

Define random variable X_B associated with the Bernoulli trial as indicator of success:

$$X_B = \begin{cases} 1, & \text{if "Success"}, \\ 0, & \text{if "Failure"}. \end{cases}$$

A sequence of N independent Bernoulli trials with the same probability of success results in two important distributions: binomial and negative binomial.

Discrete Random Variables II

Binomial Distribution

Let the number of successes N_B in N Bernoulli trials be

$$N_B = \sum_{i=1}^{N} X_{B,i} \in \{0, 1, 2, \ldots, N\}.$$

Definition

Random variable $N_B \sim Binom(p, N)$ has binomial distribution; it has probability mass distribution $\mathbb{P}\{N_B = n; p, N\} = \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{N-n}$.

Fact

$\mathbb{E}[N_B] = Np, \mathbb{V}[N_B] = Np(1-p)$

Example

Clinical trials administered to 100 patients. If probability of success is 30% what is the probability of getting more than 35 positive results?

Discrete Random Variables III

Negative Binomial Distribution

Let N_{NB} be a number of "Fails" in Bernoulli trials with probability of success p that run until we see s "Successes", $N = s + N_{NB}$ is the total number of trials.

Definition

Random variable $N_{NB} \sim NBinom(p, s)$ has **negative binomial distribution**; it has probability mass distribution

$$\mathbb{P}\{N_{NB} = n; p, s\} = \begin{pmatrix} n+s-1 \\ n \end{pmatrix} p^s (1-p)^n.$$

Fact

$$\mathbb{E}[N_{NB}] = s\frac{1-p}{p}, \mathbb{V}[N_{NB}] = s\frac{(1-p)}{p^2}$$

Discrete Random Variables IV

Negative Binomial Distribution

Definition

Random variable $N_G \sim NBinom\,(p, 1)$ has **geometric distribution**.

Example

Fund-raising drive continues until $s = 40$ people subscribe to the program. Probability that a person subscribes after being contacted is $p = 0.53$. What is the probability that at least 35 people will reject the offer before the necessary subscriptions are obtained?

- See the table of functions on page 55.
- Pay attention to the footnote on page 55: there are two common ways of defining the negative binomial distributions: as number of fails before s successes achieved, or as total number of trials before s successes achieved. In R, in our textbook and in these lectures it is the number of fails.

Discrete Random Variables V

Poisson Distribution

Imagine binomial trials with the same small probability of success. If we repeat the trials every unit of time, success is going to be a rare event. The distribution becomes close to Poisson distribution.
Poisson distribution counts events in fixed time period or space interval.

Definition

Random Variable $N_P \sim Pois(\lambda)$ has Poisson distribution if it has mass distribution function $\mathbb{P}\{N_P = n; \lambda\} = e^{-\lambda}\frac{\lambda^n}{n!}$.

Fact

$\mathbb{E}[N_P] = \mathbb{V}[N_P] = \lambda$.

Example

(2.7.1, page 82) Customers arrive on average 6 times per hour. How unusual it is to have 10 or more customers in 1 hour?

Obtaining Realizations of Discrete Random Variables

Simulation of Binomial and Negative Binomial Distributions

There are several ways of simulating each Bernoulli trial:

- Use the N generated random numbers ε_i from $[0, 1]$. The outcome i is "Success" if $\varepsilon_i \in [0, p]$ and failure if $\varepsilon_i \in (p, 1]$. This can be done either directly or with the help of function *runif.*
- Use function *rbinom*
- Use function *sample(0:1,...)*

1. **Binomial Distribution**. Use the sequence of variables $X_{B,i}$ with values 0 or 1 to create a trajectory (using cumsum). Alternatively, use function *rbinom(n,size,prob)*.
2. **Negative Binomial Distribution**. Use function rnbinom(n,size,prob)
3. **Poisson Distribution**. Use function *rpois(n, lambda)*

See the lists of functions on page 54, 55.

Definition

Exponential random variable $X_E \sim Exp(\lambda)$ is a continuous random variable with probability density and cumulative functions

$$\begin{aligned} f_{X_E}(x;\lambda) &= \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \end{cases} \\ F_{X_E}(x) &= \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \end{cases} \\ f_{X_E}(x;\lambda) &= F_X(x) = 0, x < 0. \end{aligned}$$

Fact

$\mathbb{E}[X_E] = \frac{1}{\lambda}; \mathbb{V}[X_E] = \frac{1}{\lambda^2}$

The most important application of exponential distribution comes from its connection with Poisson distribution.

If $N_P \sim Pois(\lambda)$, then time intervals between the events are $X_E \sim Exp(\lambda)$. This helps understanding how the events counted by N_P are distributed in time/space.

Fact

For $X_E \sim Exp(\lambda)$ the following is true:

$$\mathbb{P}\{X_E > b | X_E > a\} = \mathbb{P}\{X_E > b - a\}, b > a > 0.$$

Example

You come to a bus stop at a random time. Arrival of busses is a Poisson process with intensity 0.1. A passenger on the bus stop has been waiting for 5 minutes already. What is your expected waiting time?

Continuous Random Variables III

Gaussian Distribution

Definition

A continuous random variable $X_N \sim Norm(\mu, \sigma)$ has **normal (Gaussian) distribution** if its probability density function is (note typo on page 136):

$$f_{X_N}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Fact

$\mathbb{E}[X_N] = \mu; \mathbb{V}[X_E] = \sigma^2$

Definition

A random variable $Z \sim Norm(0, 1)$ is called **standard normal (Gaussian)**.

Continuous Random Variables IV

Gaussian Distribution

Probability density and probability distribution functions of Z are:

$$\begin{aligned} f_Z(z;0,1) &= \phi(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \\ F_Z(z;0,1) &= \Phi(z) \end{aligned}$$

Distribution function of $X_N \sim Norm(\mu,\sigma)$ is $\Phi\left(\frac{x-\mu}{\sigma}\right)$ and $X_N = \mu + \sigma Z$.

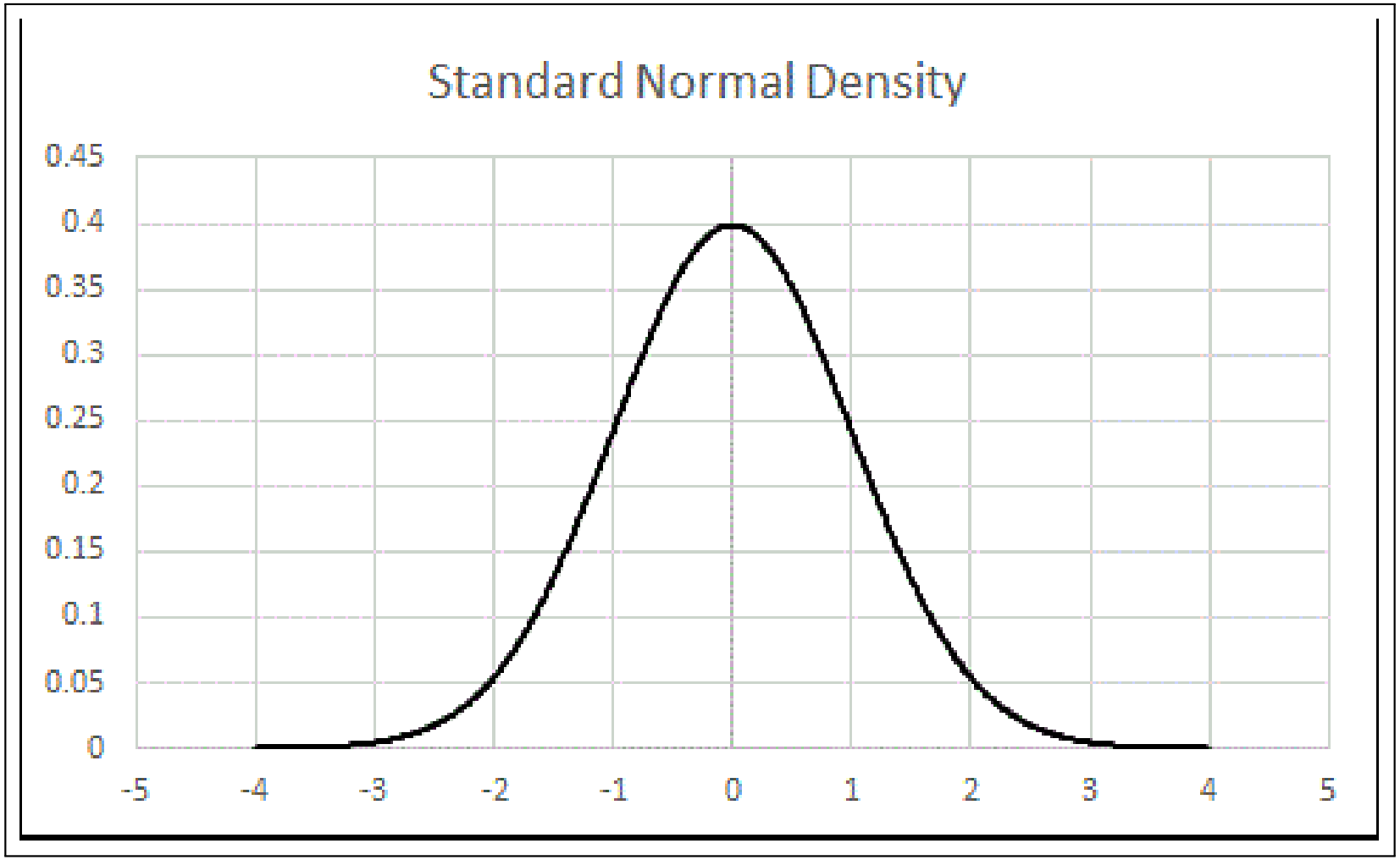
Definition

For any value x expression $\frac{x-\mu}{\sigma}$ is called **standardized score or z-score**.

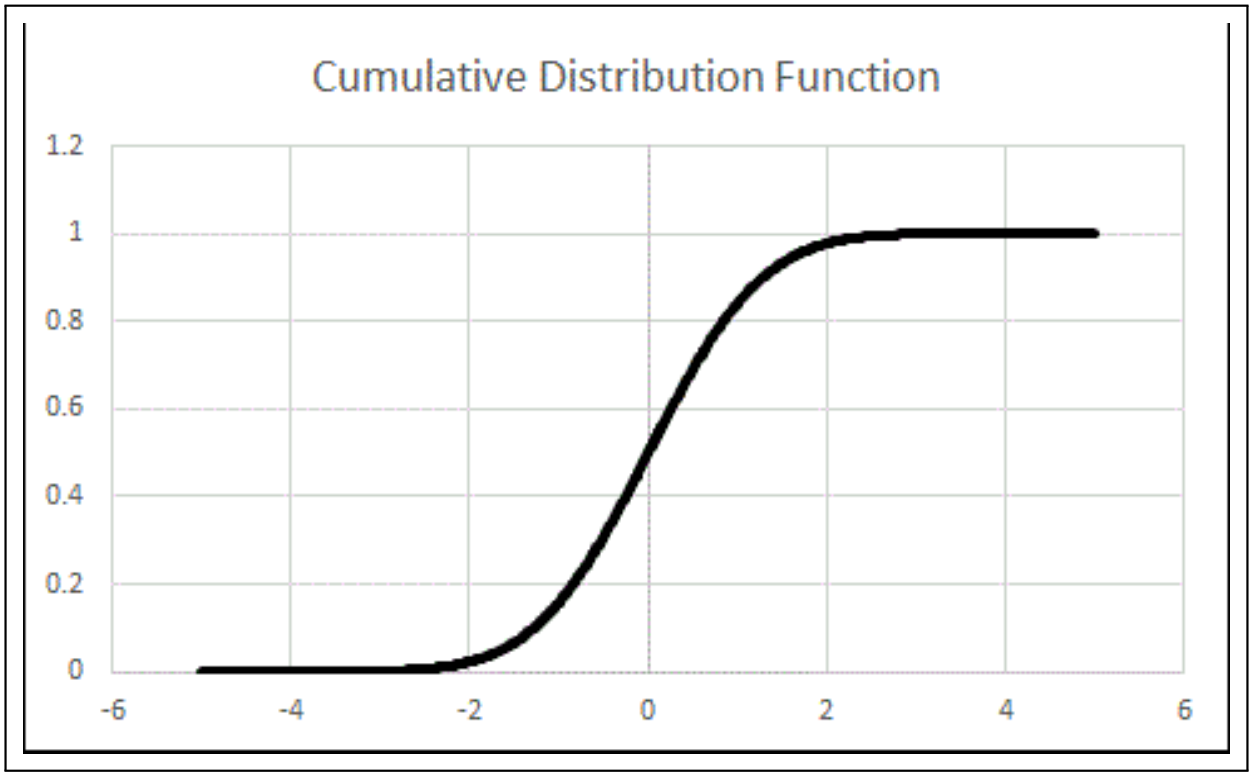
Gaussian distribution is the most remarkable among all distributions, it has many important features. Among them:

- It is the limit distribution for Central Limit Theorem (CLT)
- It is completely defined by its two moments μ and σ
- For Gaussian variables X and Y independence is equivalent to zero correlation

Gaussian Probability Density



Quantile of a Distribution I



Let random variable X have strictly increasing cumulative distribution function $F_X(x) = \mathbb{P}\{X \leq x\}$.

Definition

The **p-th quantile** of the distribution F_X is such value x_p that $F_X(x_p) = \mathbb{P}\{X \leq x_p\} = p$.

Quantile of a Distribution II

P-th quantile x_p is the value of the inverse probability distribution function: $x_p = F_X^{-1}(p)$: $F_X(x_p) = F_X\left(F_X^{-1}(p)\right) = p$.

Example

Financial firms need to define the amount of risk capital they need to survive potential extreme losses. Common measure is Value at Risk (VaR). which is a low level quantile of the loss distribution. For example, $x_{0.01}$.

Example

Let $F_X(x;\lambda)$ be an exponential distribution for which

$$\begin{aligned} f_X(x;\lambda) &= \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \end{cases} \\ F_X(x) &= \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \end{cases} \\ f_X(x;\lambda) &= F_X(x) = 0, x < 0. \end{aligned}$$

Find p-th quantile function as $x_p = -\frac{\ln(1-p)}{\lambda}$.

Simulation of Continuous Random Variables

Fact

Cumulative distribution function $F_X(X)$ *with the random variable* X *as an argument is a random variable with uniform distribution on* $[0,1]$.

Fact

Let U *be a uniform random variable on* $[0,1]$. *Consider random variable* $X = F^{-1}(U)$, *where* F *is a distribution function. Then the distribution function of* X *is* F.

Example

Simulate $X_E \sim Exp(\lambda)$ by generating U defining $X_E = -\frac{\ln(1-U)}{\lambda}$.

In R we can use function *rexp(n, rate = 1)* for generating exponentially distributed variables and *rnorm(n, mean = 0, sd = 1)* for generating Gaussian random variables.