Stochastic Biomodelling

Andrzej Mizera (andrzej.mizera@uni.lu)
Faculty of Science, Technology and Communication
University of Luxembourg

Course content

1. Introduction
   - Stochasticity in biological processes
   - Deterministic vs stochastic biomodelling
2. Prerequisite:
   - Crash course on probability theory
3. Stochastic modelling of chemical kinetics: the chemical master equation (CME)
4. Stochastic simulation of the CME – Gillespie's direct method algorithm
5. Practicals:
   - Implementing the Gillespie's algorithm in MATLAB and investigating its characteristics on various biochemical systems
   - Comparing the obtained simulation results with the solutions in the deterministic formulation

Some words on what is the role of mathematical modelling in Systems Biology

Stochastic Biomodelling

Prerequisite: Crash course on probability theory

Sample space

- **Experiments**: situations where the well defined outcomes occur randomly
- **Sample space**: the set of all possible outcomes of an experiment
- The sample space is denoted by **Ω** and its element (a possible outcome) is **ω**.
- **Exactly one outcome occurs in each experiment!**

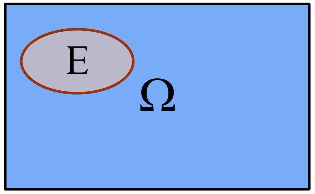
- Examples:

	ω	Ω
Toss a coin	ω = Head, Tail, (edge?)	Ω = {H, T}
Roll a die	ω = 1, 2, 3, 4, 5, 6	Ω = {1, 2, 3, 4, 5, 6}
Sex of a newborn baby	ω = Male, Female	Ω = {M, F}

Events

- A collection of possible outcomes of an experiment, or a subset of the sample space Ω, is called an **event** E.

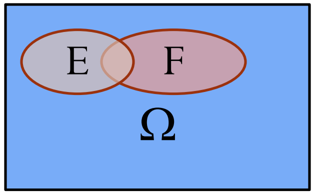
$$E \subseteq \Omega$$



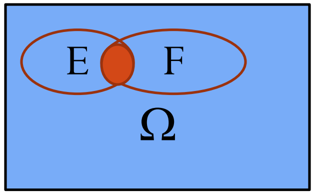
- Events may be classified into four types:
  - **null event**, denoted $\emptyset$, is the empty subset of the sample space Ω
  - **atomic event** is a subset consisting of a single element of Ω.
  - **compound event** is a subset consisting of more than one element of Ω.
  - the **sample space** Ω is also an event.

Set operations on events

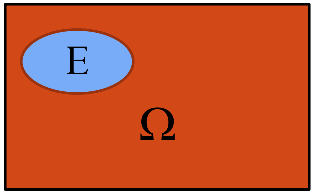
- The **union** of two events E and F (denoted $E \cup F$) is an event that at least one of E and F occurs.



- The **intersection** of two events E and F (denoted $E \cap F$) is an event that both E and F occur.

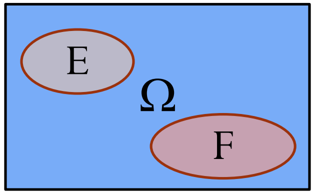


- The complement of an event E (E^c or $\bar{E}$) is the event that E does not occur. It consists of all those elements of Ω that are not in E.



Disjoint events

- Two events E and F are disjoint or mutually exclusive if they cannot both occur.
  - Their intersection is empty: **$E \cap F = \emptyset$**

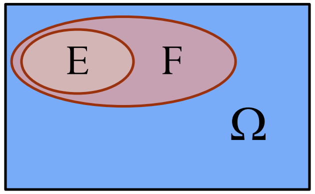


- Observe that:
  - **$E \cap E^c = \emptyset$**
  - **$E \cup E^c = \Omega$**

True events

- The event E is **true** if the outcome of an experiment (ω) is contained in the event E, i.e. if $\omega \in E$.

- The event E **implies** the event F ($E \Rightarrow F$) if the truth of F automatically follows from the truth of E.
  - If E is a subset of F then the occurrence of E necessarily implies occurrence of F:



Probability axioms

- The real valued function P(·) is a probability measure if it acts on subsets of Ω and obeys the following axioms:
  - P(Ω) = 1
  - If $E \subseteq \Omega$, then $0 \leq P(E) \leq 1$
  - If E and F are disjoint, then $P(E \cup F) = P(E) + P(F)$
    - If $E_1, E_2, \ldots, E_n$ are mutually disjoint, then $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

Probability properties

- $P(\Omega) = 1$
- $P(E^c) = 1 - P(E)$
- $P(\emptyset) = 0$
- If $E \subseteq F$, then $P(E) \leq P(F)$
- Addition law: $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
  - For disjoint E and F: $P(E \cup F) = P(E) + P(F)$

Conditional probability & independence

- The **conditional probability** (denoted by P(E | F)) describes the probability of an event (E) given that another event (F) has already occurred:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}, \quad \text{for } P(F) > 0$$

  - Multiplication rule: $P(E \cap F) = P(E) \cdot P(F|E)$

- Two events E and F are **independent** if and only if their joint probability equals the product of their probabilities:

$$P(E \cap F) = P(E) \cdot P(F)$$

  - This implies:

$$P(E) = \frac{P(E) \cdot P(F)}{P(F)} = \frac{P(E \cap F)}{P(F)} = P(E|F)$$

Independence

Example: A math teacher gave the class two tests. 25% of the class passed both tests and 42% of the class passed the first test. What percent of the students who passed the first test have also passed the second test?

Solution:

In other words, we need to find the probability that the second test was passed given that the first test was passed.

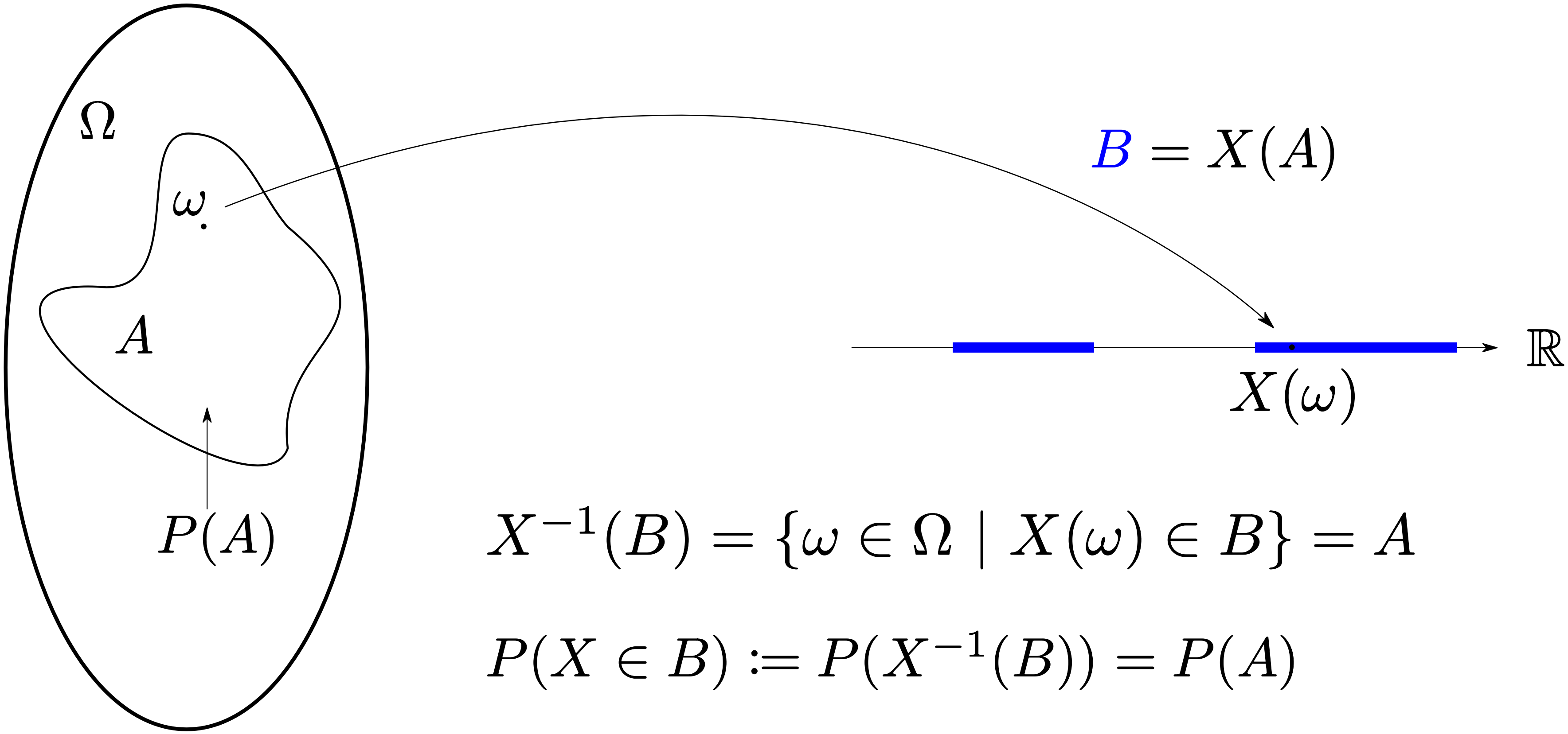
$$\mathbf{P(Second \mid First) = P(First \cap Second) \;/\; P(First) = 0.25/0.42 = 0.59}$$

Random variables

- The elements ω of the sample space Ω can be rather abstract objects.
- In practice, one often deals with simple numbers (integers, reals): one wants to 'quantify' the outcomes of random experiments and operate on them in mathematical terms (add, multiply, apply mathematical functions to them).
- The aim is thus to associate real numbers with the elements of the sample space. This idea gives rise to the concept of a **random variable**.
- For the purpose of this course a random variable X is defined to be a real-valued function (map) $X : \Omega \to \mathbb{R}$ that assigns to each atomic event $\omega \in \Omega$ a real number $X(\omega)$.

  Remark: This is the 'school' definition of a random variable. In fact, it is not complete. The complete definition contains additionally a requirement of X to be a ***measurable map***. The complete definition requires some basic knowledge of ***measure theory***. Moreover, in general, the values of a random variable need not to be elements of $\mathbb{R}$, but can belong to some other set (e.g., $\mathbb{R}^n$).

Random variables



$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\} = A$$

$$P(X \in B) := P(X^{-1}(B)) = P(A)$$

Random variables

- We distinguish between **discrete** and **continuous** random variables.
    - Discrete random variable: taking any of a specified finite or countable list of values
    - Continuous random variable: taking a value from a continuous range of values

Discrete random variable

Example: Tossing a fair coin twice

- $\Omega = \{HH, HT, TH, TT\}$
- There may be **several** random variables associated with this experiment.
  - Let random variable X be the number of tails in the outcome: $X(HH) = 0; X(HT) = 1; X(TH) = 1; X(TT) = 2$.

The **observed value** of a random variable is the number corresponding to the actual outcome, i.e., if the outcome of the experiment is $\omega \in \Omega$, then $X(\omega) \in \mathbb{R}$ is the observed value.

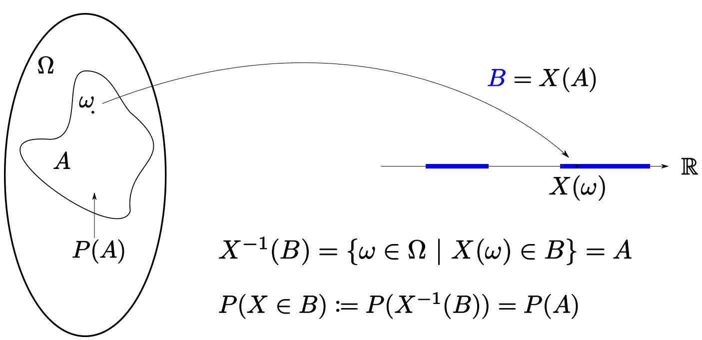
- The observed value is denoted as x.
- Example: the set of possible observed values for X is $S_X = \{X(\omega) \mid \omega \in \Omega\}$, so in the example above it is: $S_X = \{0, 1, 2\}$.

Notice that the values in S_X need not to be equally likely!

Discrete probability distribution

- **Probability distribution** assigns a probability to each (measurable) subset of the possible outcomes of a random experiment.
- A discrete probability distribution is a probability distribution characterized by a **probability mass function** (**PMF**), i.e., a function that gives the probability that a discrete random variable is exactly equal to some value.

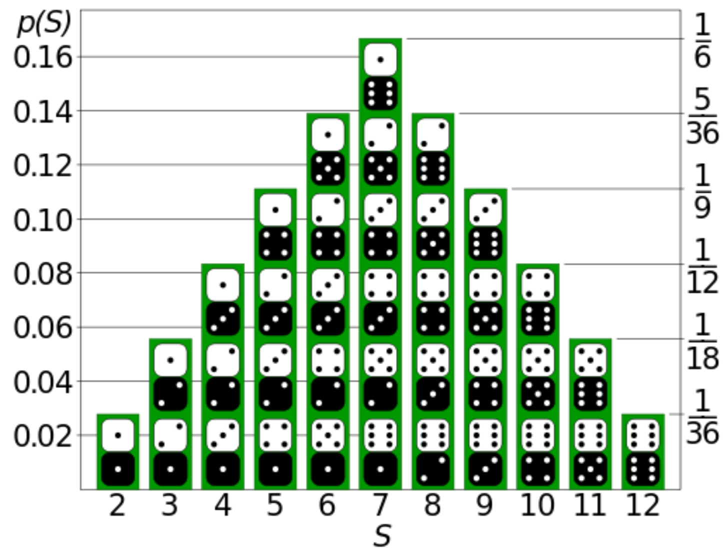
$$\Pr(X = x) = \sum_{\{\omega \in \Omega \mid X(\omega) = x\}} \Pr(\{\omega\})$$



Probability mass function

Example: rolling two dice

- Let S be the sum of numbers on the two dice.
- $S_S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.
- These values are not equally likely!



Source: Wikipedia

Value (s)	2	3	4	5	6	7	8	9	10	11	12
Pr(S = s)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

The table provides a representation of the probability mass function for S.

Cumulative distribution function

- **Cumulative distribution function** (**CDF**) is the probability that a real-valued random variable X will take a value less than or equal to x.

$$F_X(x) = \Pr(X \le x) = \sum_{\{y \in S_X \mid y \le x\}} \Pr(X = y)$$

- **Example:** sum of two dice

PMF

s	2	3	4	5	6	7	8	9	10	11	12
Pr(S = s)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

CDF

s	2	3	4	5	6	7	8	9	10	11	12
$F_S(s)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

Expectation of a random variable

- The **expected value** (expectation, mean, average, first moment) of a random variable is the weighted average of all possible values that this random variable can take on, where the weights are the respective probabilities:

$$\mu = E(X) = \sum_{x \in S_X} x \cdot \Pr(X = x)$$

- Remarks:
  - The expected value is **not** a random variable!
  - It represents the value of the random variable, that is expected to get on average.

Variance and standard deviation

- **Variance** (second central moment) measures how far a set of numbers are spread out.
- It is the **mean of the squared deviations** from the expected value E(X):

$$\sigma^2 = Var(X) = E[(X - E(X))^2)] = \sum_{x \in S_X} \Pr(X = x) \cdot (x - E(X))^2$$

$$= E(X^2) - (E(X))^2$$

- **Standard deviation**: a measure of dispersion of a set of data values

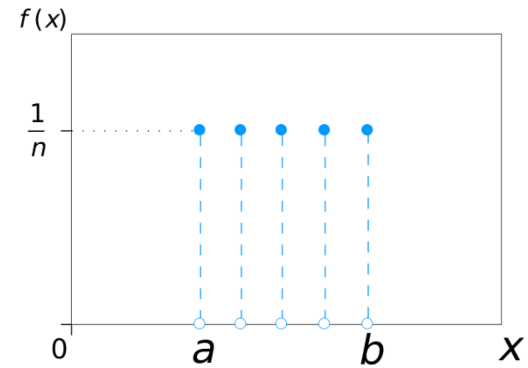
$$\sigma = SD(X) = \sqrt{Var(X)}$$

Uniform distribution

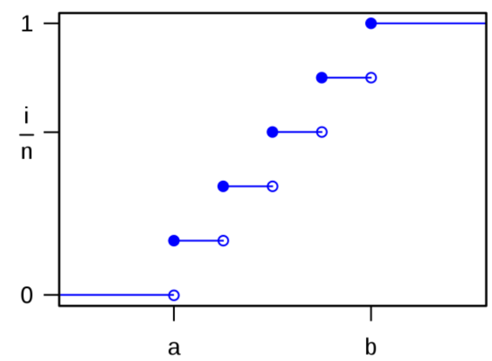
The **discrete uniform distribution** is a symmetric probability distribution whereby a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n.

- Sample space: $S_X = \{1, 2, \dots, n\}$
- PMF: $\Pr(X = k) = \begin{cases} \frac{1}{n} & k \in S_X \\ 0 & k \notin S_X \end{cases}$
- CDF: $\Pr(X \leq k) = \frac{k}{n} \quad k \in S_X$
- Expectation: $E(X) = \frac{n+1}{2}$
- Variance: $Var(X) = \frac{n^2-1}{12}$

PMF



CDF



Binomial distribution

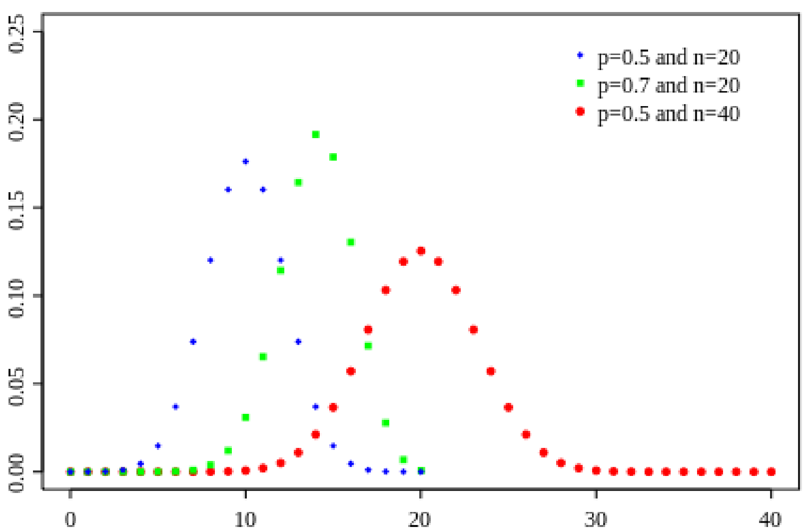
The **binomial distribution** with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent experiments, each of which yields success with probability *p*.

- $X \sim B(n,p)$
- The experiment consists of n repeated trials.
- Each trial can result in **just two** possible outcomes: a success and a failure.
- The probability of success *p* is the same on every trial and the probability of failure is $1-p$.
- The trials are independent.

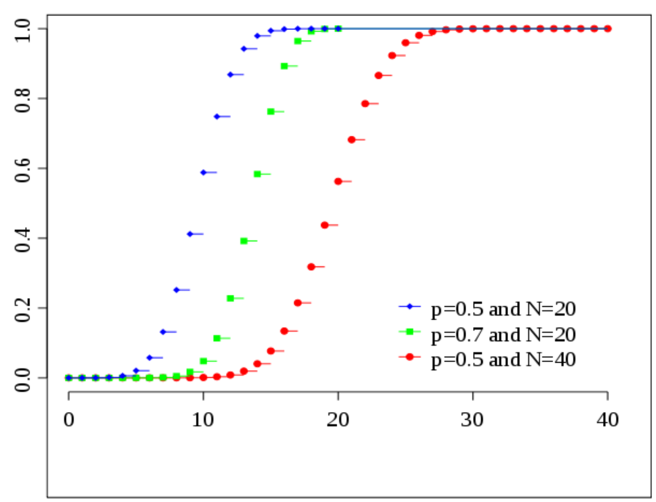
Binomial distribution

- Sample space: $S_X = \{0, 1, 2, \ldots, n\}$
- PMF: $\Pr(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k},\ k \in S_X$
- CDF: $\Pr(X \le k) = \sum_{i=0}^{k} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i},\ \ \mathrm{k} \in S_X$
- Expectation: $E(X) = n \cdot p$
- Variance: $Var(X) = n \cdot p \cdot (1-p)$

PMF



CDF



Continuous random variables

- **Continuous random variables** are appropriate if the result of an experiment is a continuous measurement, rather than a count of a discrete set.
  - **Example:** in the context of Systems Biology, such measurements often correspond to measurements of **time**.

- If X is a continuous random variable with sample space S_X, then for aby particular $x \in S_X$, we generally have that $\Pr(X = x) = 0$.
  - **Intuition:** the sample space is so "large" and every possible outcome so "small" that the probability of any particular value is vanishingly small.
  - Therefore, the probability mass function introduced for discrete random variables is inappropriate for describing continuous random variables.

Probability density function

- **Probability density function** (**PDF**): If X is a continuous random variable, then there exists a function $f_X(x)$, called the probability density function, which satisfies the following:
  1. $f_X(x) \geq 0, \quad \forall x;$
  2. $\int_{-\infty}^{\infty} f_X(x)\, dx = 1;$
  3. $\Pr(a \leq X \leq b) = \int_a^b f_X(x)\, dx$ for any $a \leq b$.
- Consequently:

$$\Pr(x \leq X \leq x + \delta x) = \int_x^{x+\delta x} f_X(y)\, dy \cong f_X(x)\, \delta x \quad \text{(for small } \delta x\text{)}$$

$$\Longrightarrow f_X(x) \cong \frac{\Pr(x \leq X \leq x + \delta x)}{\delta x}$$

- If $f_X: \mathbb{R} \to \mathbb{R}$, then in the points where f_X is continuous we have that

$$f_X(x) = \lim_{\delta x \to 0} \frac{\Pr(x \leq X \leq x + \delta x)}{\delta x}.$$

Probability density function

Remarks:

- PDF values are **not** probabilities! The density can take values greater than 1.
- $\Pr(X \leq k) = \Pr(X < k)$ for continuous random variables.

Cumulative distribution function

- **Cumulative distribution function** (**CDF**) for continuous random variables:

$$F_X(x) = \Pr(X \leq x) = \Pr(-\infty \leq X \leq x) = \int_{-\infty}^{y} f_X(y)dy.$$

- CDF values are probabilities!
- The distribution function is defined for all $x \in \mathbb{R}$, even if the sample space S_X is not the whole of the real line.
- $F_X(x) \in [0,1]$
- $F_X(-\infty) = 0$ and $F_X(\infty) = 1$
- If $a < b$, then $F_X(a) \leq F_X(b)$
- When X is contiunous, $F_X(x)$ is continuous. Also, by the *Fundamental Theorem of Calculus*, we have

$$\frac{d}{dx} F_X(x) = f_X(x).$$

Expectation & variance

- The **expectation** of a continuous random variable X is given by

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

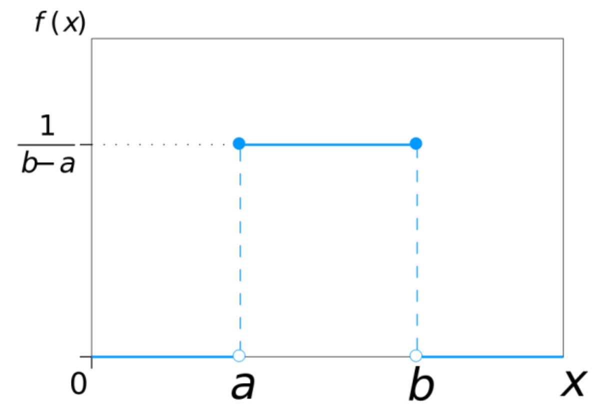
- The variance is given by

$$Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 \cdot f_X(x)\, dx = \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx - [E(X)]^2 = E(X^2) - [E(X)]^2$$

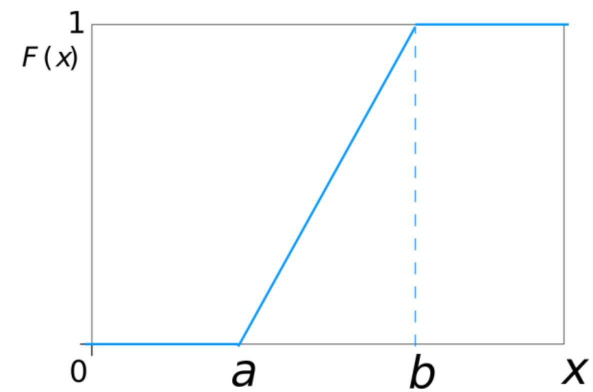
Uniform distribution

- **Uniform distribution**, X ~ U(a,b), is a distribution that has a constant probability.
- PDF: $f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$
- CDF: $F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$
- Expectation: $E(X) = \frac{a+b}{2}$
- Variance: $Var(X) = \frac{(b-a)^2}{12}$

PMF



CDF



Normal (Gaussian) distribution

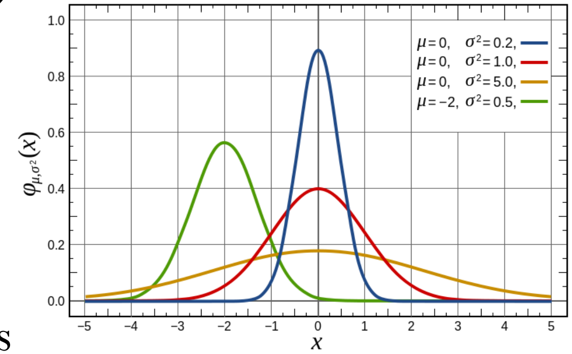
- A random variable X has a **normal distribution** with parameters μ and σ^2, denoted $X \sim N(\mu, \sigma^2)$ if it has pdf

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$$

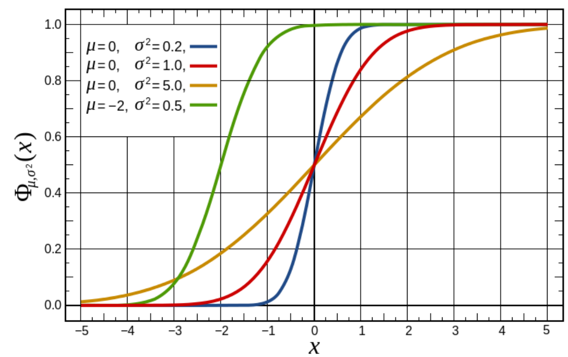
  for $\sigma > 0$.

- It turns out that sums of random quantities often approximately follow a normal distribution.
  - Very important in statistics.
- CDF: $\frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
- Expectation: $E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)dx = \mu$
- Variance: $Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)dx = \sigma^2$

PMF



CDF



Exponential distribution

- The random variable X has an **exponential distribution** with parameter $\lambda > 0$, denoted $X \sim Exp(\lambda)$, if it has PDF

$$f_X(\mathrm{x}) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

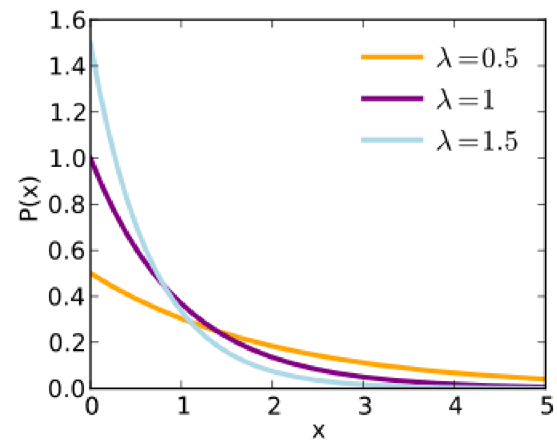
- The exponential distribution is the most important continuous distribution in the theory of discrete-event stochastic simulation.

- CDF: $F_X(\mathrm{x}) = \begin{cases} 0 & x < 0, \\ 1 - e^{-\lambda \cdot x} & x \geq 0. \end{cases}$

- Expectation: $E(X) = \int_0^{\infty} x \cdot \lambda \cdot e^{-\lambda \cdot x}\, dx = \frac{1}{\lambda}$

- Variance: $Var(X) = \frac{1}{\lambda^2}$

PMF



CDF