technische universität
dortmund

Statistical Methods of Data Analysis

Probability

Prof. Dr. Dr. Wolfgang Rhode Dr. Maximilian Linhoff

2023

Overview

Probability Definitions

1D - Distributions
- Probability Distribution, CDF and PDF
- Moments
- Measures of Central Tendency and Dispersion

Common 1D-Distributions
- Discrete Distributions
- Continous Distributions

Overview

Probability Definitions

1D - Distributions
- Probability Distribution, CDF and PDF
- Moments
- Measures of Central Tendency and Dispersion

Common 1D-Distributions
- Discrete Distributions
- Continous Distributions

Frequentist Definition

In a Frequentist way one can introduce probability in two different ways: *With* our *without* a priori knowledge. *A priori* knowledge is information that is independent from the current experience and can be known in advance while *a posteriori* knowledge is inferred from the current experience.

1. **With a priori knowledge:** If an event can result in n equally possible but different outcomes and k of these outcomes have a property A, then the probability for A to happen is

$$P(A) = \frac{k}{n} = \frac{\text{desirable outcomes}}{\text{possible outcomes}}.$$

2. **Without a priori knowledge:** The outcomes A and not-A are *independently* measured n times where A occurs k times. Then the probability for A to happen is

$$P(A) = \lim_{n\to\infty} \frac{k}{n}.$$

Bayesian Definition

In a Bayesian way, the probability $P(A \mid B)$ is a quantitative measure of the assumption A's *plausibility* under the known information given through an assumption B.

- A can be an arbitrary logical assumption
- $P(A \mid B)$ can be computed through **Bayes' Theorem** (valid also in Frequentist statistics)
- Problem: Definition of the term *plausibility*

To state **Bayes' Theorem** we need some foundations first.

Kolmogorov-Axioms

Another effort of finding a definition of probabilities has been made by Andrey Nikolaevich Kolmogorov in 1933 through three (originally five) axioms.
Let Ω be the set of all *possible events* and Σ a set of subsets of Ω. The elements of Σ are called *random events*.

1. The probability $P(A)$ of a set $A \in \Sigma$ is a real number between 0 and 1.
2. The probability for at least one of the possible events to happen is

$$P(\Omega) = 1.$$

3. If some n sets $A_i \in \Sigma$ have no elements in common then their joint probability is

$$P\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P\left(A_i\right).$$

Kolmogorov-Axioms - Example: Ideal Coin Flip

As an example, consider an ideal coin flip. Here we have

$$\Omega = \{\text{heads}, \text{tails}\}$$
$$\Sigma = \{\emptyset, \text{heads}, \text{tails}, \Omega\}.$$

The probabilities for each of Σ's subsets are

$$P(\emptyset) = 0$$
$$P(\text{heads}) = 1 - P(\text{tails})$$
$$P(\Omega) = 1$$

and for a coin flip that is not rigged, we have

$$P(\text{heads}) = P(\text{tails}) = 0.5.$$

Calculating with Probabilities

We will now summarize some calculation rules for Probabilities.

Let A and B be events with probabilities $P(A)$ and $P(B)$ while $P(A \mid B)$ and $P(B \mid A)$ are the probabilities for A to happen if B has already happened and vice versa. $P(A \wedge B)$ is the probability for A *and* B to happen and $P(A \vee B)$[1] is the probability for A *xor* B to happen.

- $P(A \wedge B) = P(A) \cdot P(B \mid A)$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- If A and B are independent then $P(B \mid A) = P(B)$ and therefore $P(A \wedge B) = P(A) \cdot P(B)$
- If A and B are disjoint then $P(A \wedge B) = 0$ and therefore $P(A \vee B) = P(A) + P(B)$
- If A and B are complementary then $P(B) = 1 - P(A)$ and therefore $P(A \vee B) = 1$

[1] The logical *or* symbol $\vee$ is a short notation for the latin "vel" engl. "or". In this context we mean the exclusive or (xor) as can be seen from the slide's second equation.

Calculating with Probabilities - Bayes' Theorem

From the first equation on the last slide,

$$P(A \wedge B) = P(A) \cdot P(B \mid A),$$

and the assumption that $P(A \wedge B) = P(B \wedge A)$, we can directly derive the already mentioned **Bayes' Theorem**.

Bayes' Theorem

Let $P(B \mid A)$ be the conditional probability for B to happen if A has already happened and $P(A)$ and $P(B)$ are the probabilities for A and B to happen respectively then

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}.$$

Calculating with Probabilities - Further Rules for Conditional Probabilities

For conditional probabilities there is a summation rule

$$P(A \mid B) + P(\bar{A} \mid B) = 1$$

where $\bar{A}$ is the complement to A.

Furthermore, there is a product rule

$$\begin{aligned} P(A \wedge B \mid C) &= P(A \mid C) \cdot P(B \mid A \wedge C) \\ &= P(B \mid C) \cdot P(A \mid B \wedge C). \end{aligned}$$

Overview

Probability Definitions

1D - Distributions
- Probability Distribution, CDF and PDF
- Moments
- Measures of Central Tendency and Dispersion

Common 1D-Distributions
- Discrete Distributions
- Continous Distributions

Probability Distribution

A probability distribution P is a mapping of the event-space Σ onto a probability

$$P : \Sigma \to [0, 1].$$

One can distinguish between discrete and continuous probability distributions.

The *(cumulative) distribution function (CDF)* is the cumulative probability up to a value x for a random variable X

$$F(x) = P(X \leq x)$$

- Discrete: $F(x) = \sum_{x_i \in \Sigma \wedge x_i \leq x} P(x_i)$
- Continuous: $F(x) = \int_{-\infty}^{x} f(x)\ \mathrm{d}x$ with $f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x)$ if the *probability density function (PDF)* $f(x)$ exists

Probability Distribution

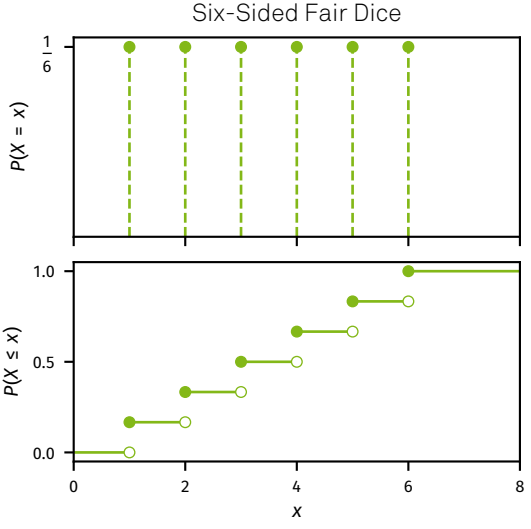
What gives a meaningful probability?

- **For discrete distributions**, one uses the probability distribution $P(x)$ as introduced before
- **For continuous distributions**, the probability to obtain an exact value x is $P(X = x) = 0$, therefore the only meaningful value is a probability on an interval

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$$

Example of a Discrete Distribution: Fair Dice

For a fair dice with six sides, the probability distribution (top) is given through discrete peaks located at the dice's possible faces. The CDF (bottom) is a step function with six steps between **0** and **1**.



Mode

The most probable value of $P(X)$ is called the mode x_{mode} of the distribution

$$x_{\text{mode}} = \text{argmax}\,(P(X))\,.$$

A distribution is called

- **Unimodal**, if there is only one global maximum that then gives the mode
- **Multimodal**, if there are multiple local maxima from which one, the global maximum, is the mode

Median and Quantiles

The median x_{median} of a distribution is the value up to which and from which half of the total probability lies, therefore where the CDF is $F\left(x_{\text{median}}\right) = 0.5$.

A generalized approach are quantiles

- The q-quantile is where $F\left(x_q\right) = q$
- The median is the distributions 0.5- or 50 %-quantile
- Further named quantiles are the lower and upper quartile, the 25 % and 75 % quantiles

Concluding Remarks on CDF and PDF

Some concluding remarks on CDF and PDF for continuous distributions

- $\lim_{x\to\infty} F(X) = 1$ and since $P(A) + P(\bar{A}) = 1$, $\lim_{x\to-\infty} F(X) = 0$
- The probability for X being smaller than some a is

$$P(X \leq a) = \int_{-\infty}^{a} f(x)\,\mathrm{d}x = F(a)$$

  and again since $P(A) + P(\bar{A}) = 1$, the probability for X being higher than some a is

$$P(x > a) = 1 - F(a)$$

- The PDF is normed to unity

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$

Overview

Expected Value and Mean

An important parameter characterizing a PDF is the expected value of the function $h(x)$. Given a continuous PDF $f(x)$, it is defined as

$$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\,\mathrm{d}x\,.$$

The expected value of $h(x) = x$ is also called the mean of a distribution

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

and often written as $\langle X \rangle$.

Connecting the mean to the earlier introduced mode and median, all of them coincidence for a unimodal, continuous and symmetric PDF.

Moments

The distribution mean is also called the first *algebraic* or *raw moment* of a distribution. The n-th raw or algebraic moment μ'_n is the expected value of $h(x) = x^n$, therefore

$$\mu'_n = \mathrm{E}[X^n] = \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x\,.$$

Closely related are the *central moments* μ_n, where $h(x) = (x - \mathrm{E}[x])^n$

$$\mu_n = \mathrm{E}[(X - \mathrm{E}[X])^n] = \int_{-\infty}^{\infty} (x - \mathrm{E}[x])^n f(x)\,\mathrm{d}x\,.$$

Some remarks:

- A PDF is fully defined by all of its moments
- Trivially, the zeroth and first central moments are $\mu_0 = 1$ and $\mu_1 = 0$
- For discrete distributions, one can switch the integrals for summations

Variance and Standard Deviation

The second central moment is called variance of the distribution

$$\mu_2 = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \int_{-\infty}^{\infty} (x - \mathrm{E}[x])^2 f(x)\,\mathrm{d}x \coloneqq \mathrm{Var}(X) \coloneqq \sigma^2(X).$$

- The variance is a measure for the distribution's width
- The square root of the variance is called standard deviation σ

Rules for Mean and Variance

If a variable is scaled with a constant

$$H(x) = cx, \text{with } c = \text{const.}$$

one has

- $\mathrm{E}[cX] = c \cdot \mathrm{E}[X]$
- $\mathrm{Var}[cX] = c^2 \cdot \mathrm{Var}[X]$
- Using the above and expanding the second order term inside the variance's definition, one can derive a form of the Displacement Theorem

$$\mathrm{Var}(x) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$$

Standardized Variables

Using mean and variance, a variable can be transformed into a standardized form U

$$U = \frac{X - \mathrm{E}[X]}{\sigma(X)}$$

- A standardized variable has mean **0**

$$\mathrm{E}[U] = \frac{1}{\sigma(X)} \mathrm{E}[X - \mathrm{E}[X]] = 0$$

- A standardized variable has variance **1**

$$\mathrm{Var}[U] = \frac{1}{\sigma^2(X)} \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \frac{\sigma^2(X)}{\sigma^2(X)} = 1$$

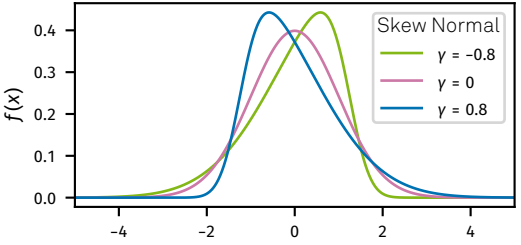
- Standardizing variables yields advantages in data-analysis, as we will see later on

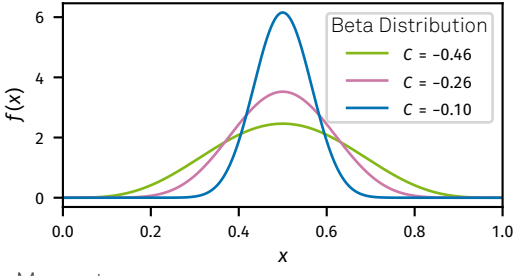
Standardized Moments: Skewness and Kurtosis

Closely related are the *standardized moments*.
The n-th standardized moment is defined as

$$\tilde{\mu}_n = \frac{\mu_n}{\sigma^n}.$$

- The skewness $\gamma := \tilde{\mu}_3$ is a measure for a distribution's symmetry
- $C := \tilde{\mu}_4$ is called kurtosis
- The normal distribution has a kurtosis $C = 3$, in consequence, often the excess kurtosis $\tilde{\mu}_4 - 3$ is used





Root Mean Square

The root mean square (RMS), not to be confused with the root mean square error (RMSE), is defined as

$$\mathrm{RMS}(X) = \sqrt{\mathrm{E}[X^2]} = \sqrt{\sigma^2 + \mathrm{E}[X]^2}.$$

If the expected value is zero this collapses to

$$\mathrm{RMS}(X) = \sigma(X).$$

Due to a bug that propagated from `PAW` to `ROOT`, many particle physicists say RMS but mean the standard deviation σ, even if the expected value is not zero. See

ROOT's documentation on the `GetRMS` function.

Overview

Measures of Central Tendency

Until now, we have discussed moments. They describe a probability distribution and **not** a random sample drawn from it. These are described by *measures of central tendency and dispersion*.

As measures of central tendency, most used are different *means*.

Arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Geometric mean

$$\bar{x}_G = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Harmonic mean

$$\bar{x}_H = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1}$$

Further used are quantiles, including the median, by sorting all x_i and using e.g. for the median

$$x_{0.5} = \begin{cases} x_{\frac{n+1}{2}} & \text{, if n odd} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) & \text{, if n even} \end{cases}$$

Please note: The expected value is also called mean, easily confused with the arithmetic mean.

Measures of Dispersion

Measures of dispersion are the analogy to the distributions variance. There is the

- Empirical variance

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Empirical sample variance

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

In case if no dispersion ($f(x_i) = \delta_{ij}$), the Shannon-Entropy (average information content)

$$H = -\sum_{i=1}^{n} f(x_i)\log_2 f(x_i)$$

is zero. In the case of maximized dispersion, $f(x_i) = \frac{1}{n}$, H becomes $\log_2 n$.

Overview

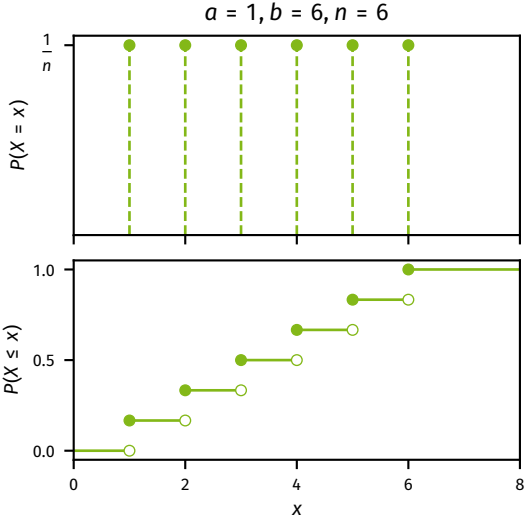
Discrete Uniform Distribution

- Probability distribution:

$$P(X = x) = \frac{1}{n}$$

  with n outcomes in $[a, b]$ at points $x_i = \{a, a + 1, \ldots, b\}$, $a, b \in Z$ and $a < b$.

- Usage: Random experiments with equally possible results like
  - Dice
  - Coin-flip

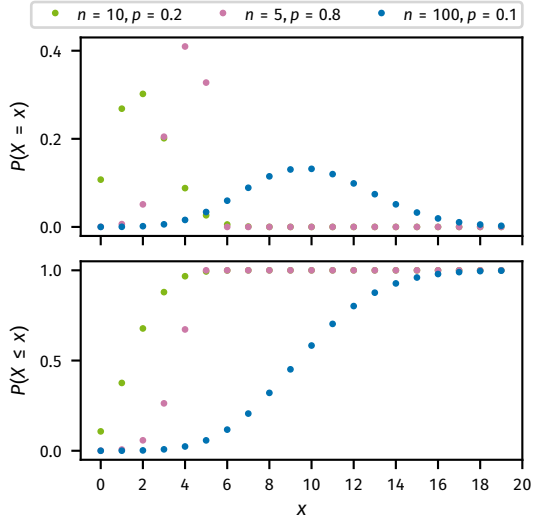


Binomial Distribution

- Probability distribution:

$$P(X = k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

  - n: Total number of events
  - $k \in \{0, 1, \dots, n\}$: Successful events
  - p: Probability for success
- Expected value: np
- Variance: $np(1-p)$
- Usage example: Number of triggered events given a trigger-probability p and n total events

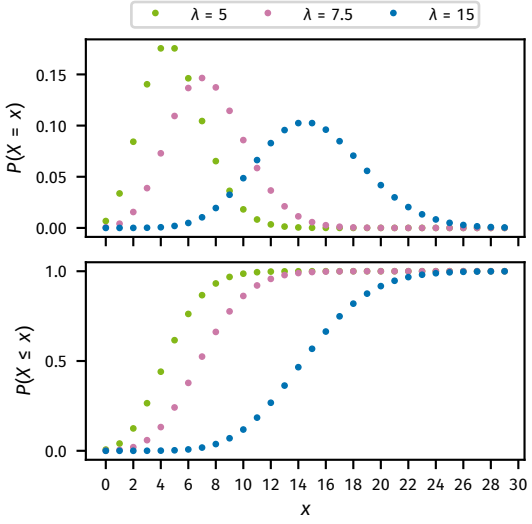


Poission Distribution

- Probability distribution:

$$P(X = k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Expected value: λ
- Variance: λ^2
- Approximation for binomial distribution with high n and small p using $\lambda = np$
- Usage example: Counting experiments, number of events per time-interval



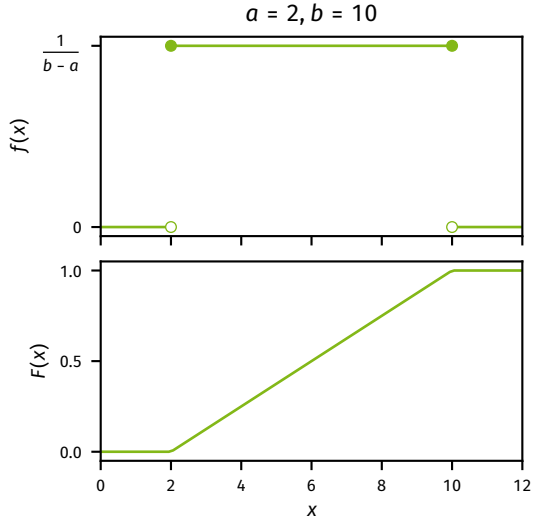
Overview

Continous Uniform Distribution

- PDF:

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

- Expected value: $\mu = \frac{a+b}{2}$
- Variance: $\sigma^2 = \frac{1}{12}(b-a)^2$
- Usage: Transformation into random variates of other distributions



Normal or Gaussian Distribution

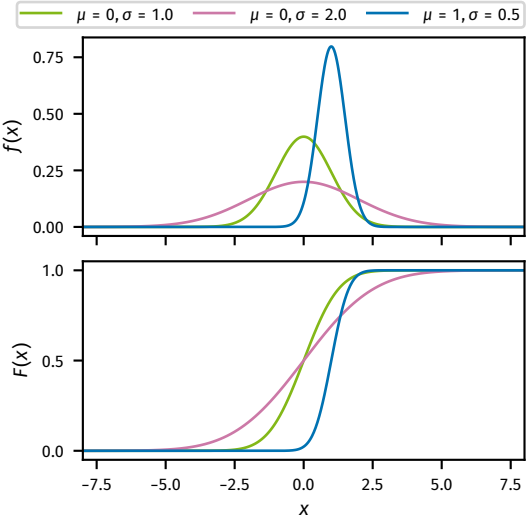
- PDF:

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

- CDF:

$$F(x|\mu,\sigma) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\pi\sigma^2}}\right)\right)$$

- Expected value: μ
- Variance: σ^2
- The distribution is often written as $\mathcal{N}(\mu, \sigma)$
- $\mathcal{N}(\mu = 0, \sigma = 1)$ is called standard-normal



Normal or Gaussian Distribution - Error-function, Central Intervals and Usage

The error-function **erf(x)** has no analytical form so look-up tables or calculators have to be used.

The probability in the intervals centered around the expected value is:

- $[\mu - 1\sigma, \mu + 1\sigma]$: 68.27 %
- $[\mu - 2\sigma, \mu + 2\sigma]$: 95.45 %
- $[\mu - 3\sigma, \mu + 3\sigma]$: 99.73 %
- $[\mu - 4\sigma, \mu + 4\sigma]$: 99.9937 %
- $[\mu - 5\sigma, \mu + 5\sigma]$: 99.999 943 %

The normal distribution is a fundamental distribution in statistics. It is used e.g. for:

- Modelling uncertainties
- Summation of many small deviations, as the sum of many uniformly distributed random variates is normal distributed

For further examples see the jupyter notebook and video `normal_distribution.{ipynb,mp4}`.

χ^2-Distribution

- PDF:

$$f(x|k) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

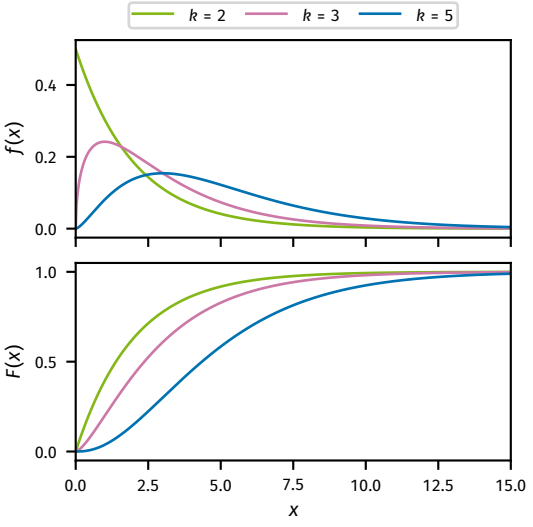
  with the Gamma function Γ and k degrees of freedom

- CDF:

$$F(x|k) = P\left(\frac{k}{2}, \frac{x}{2}\right)$$

  with the incomplete Gamma function P

- Expected value: k
- Variance: $2k$



χ^2-Distribution - Usage

The χ^2-distribution appears in numerous contexts, including

- as the distribution of the sum of k standard-normal distributed random variates
- confidence intervals (SMD B, estimators)
- statistical tests (SMD B, testing)

Exponential Distribution

- PDF:

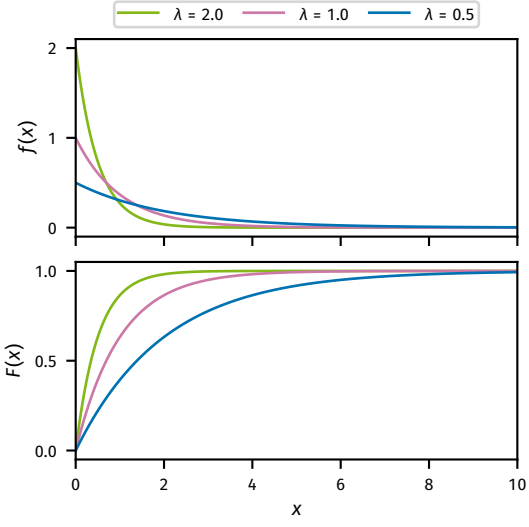
$$f(x \mid \lambda) = \begin{cases} \lambda e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

  with $\lambda > 0$

- CDF:

$$F(x \mid \lambda) = 1 - e^{-\lambda x}$$

- Expected value: λ^{-1}
- Variance: λ^{-2}
- A common-used parametrization is $\tau = \lambda^{-1}$



Exponential Distribution - Usage

The exponential distribution is a typical live-time distribution as the difference between to subsequent uniformly distributed random variates is exponential distributed. This include:

- Radioactive decay
- Particle decay
- The time between two calls in a call-center
- The live-time of some product under the assumption that the failure probability is constant

Exponential and Poission Distribution - The Poission Process

The Poission Process is a statistical process that appears when counting events that happen with a certain rate yet are themselves completely random.

Consider a detector with many cells, n in total, that measure an event with small probability p, then

- the number of events per measurement is Poission distributed with $\lambda_n = np$
- the event's time-stamps are uniformly distributed
- the time between two subsequent events is exponential distributed with $\lambda_{\Delta t} = \frac{1}{\lambda_n}$

For a visualization please see the jupyter notebook and video `poisson_process.{ipynb, mp4}`.

Further Distributions

Further distributions are needed e.g. as test-statistics or for specialized processes.

A great source for information on nearly all distributions including important characteristics as well as algorithms to create random variates following the distributions is the

„Handbook on Statistical Distributions for Experimentalists“

by Christian Walck.