Fachhochschule Dortmund
University of Applied Sciences and Arts

Data Science

08: Random variables

Klaus Kaiser
WiSe 2024 / 2035

Data Science

Evaluation

Data Science

Recap: Random experiment

A **random experiment**, is an experiment for which

- conditions are well-defined
- multiple different outcomes are possible
- it is not predictable which outcome will occur

Example A: Simple dice flip

A dice with six sides is thrown. The result can not be predicted in before.

Example A: Color of a passing car

The color of the next passing car is predicted. The result can not be predicted in before.

Data Science

Recap: Kolmogoroff axioms

A probability measure is a function $P : \mathcal{P}(\Omega) \to \mathbb{R}$ with the following properties:

- $P(A) \geq 0$ for all $A \subset \Omega$
- $P(\Omega) = 1$
- For piece-wise disjoint events $A_1, A_2, \ldots$ there holds
  - $P(A_1 \cup A_2 \cup \cdots \cup A_k) = \sum_{i=1}^{k} P(A_i)$ finite many and
  - $P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$ for countable infinite many.

Some properties of probability measures

- $P(\bar{A}) = 1 - P(A)$ for all $A \in \mathcal{P}(\Omega)$
- $P(\emptyset) = 0$
- If $A \subset B$ then $P(A) \leq P(B)$
- $P(\bar{A}) \leq 1$ for all $A \in \mathcal{P}(\Omega)$

Data Science

Recap: Random experiment

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\#\text{ number of favorable cases}}{\#\text{ number of possible cases}}$$

- For simple cases easy: Throwing a dice has 6 possible results and 3 of them are even - probability that the result is even: $\frac{3}{6}$
- But can get more complex: To color three elements, one can choose from 10 colors - what is the probability that at least one element is red?

Data Science

Recap: Random experiment

Number of possible samples with k observations out of n objects is given by the following table

	Repeats allowed	No Repeats ($k \leq n$)
Combinations (order doesn't matter)	$\binom{n+k-1}{k}$	$\binom{n}{k}$
Permutations (order matters)	n^k	$\frac{n!}{(n-k)!}$

Data Science

Recap

Events A and B are called **(stochastic) independent** if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$

Otherwise A and B are called **(stochastic) dependent.**

The following is equivalent ($\Leftrightarrow$) to A and B are stochastically independent

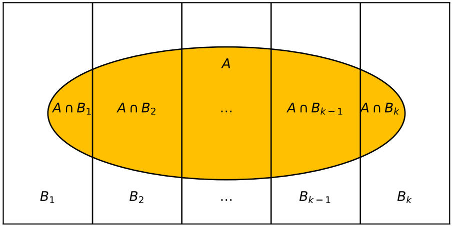
- $\Leftrightarrow$ $P(B|A) = P(B)$
- $\Leftrightarrow$ $P(A|B) = P(A)$
- $\Leftrightarrow$ $\overline{A}$ and $\overline{B}$ are stochastically independent
- $\Leftrightarrow$ $\overline{A}$ and B are stochastically independent
- $\Leftrightarrow$ A and $\overline{B}$ are stochastically independent

Data Science

Recap

In the following: We assume that there are $B_1, \dots, B_k$ events which fulfill the following conditions:

- $B_i \cap B_j = \emptyset$ for $i \neq j$, i.e. the events are piece-wise disjoint
- $\bigcap_{i=1}^{k} B_i = \Omega$, i.e. the events cover the complete sample space



Law of total probability

Under the given conditions, there holds for every $A \subset \Omega$:

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$$

Bayes' Theorem

Under the given conditions, there holds for every $A \subset \Omega$

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

Data Science

Today

Random variables and **distribution functions**

Random Variables

Often: Results of random process can be mapped to numbers

A random variable provides an assignment of results to numbers

- **Discrete random variable:** Countable many possible values
- **Continuous random variable:** Any value in an interval possible

- Throwing a dice:
  - Is the thrown value larger than 2? $X : \{1, 2, 3, 4, 5, 6\} \to \{0, 1\}, X(\omega) = \begin{cases} 0 & \omega \leq 2 \\ 1 & \omega > 2 \end{cases}$
  - **Of interest:** Probability $P(X = 1)$ respectively $P(X = 0)$.
- Overweight of persons:
  - Height ω_H (cm) and weight ω_W (kg) of persons: $\Omega_0 = \{\omega = (\omega_H, \omega_W) | \omega_H > 0, \omega_W > 0\}$

$$X : \Omega_0 \to \mathbb{R}^+, X(\omega) = \frac{Weight(kg)}{(Height(m))^2} = \frac{\omega_W}{(\omega_H/100)^2}$$

  - **Of interest:** Does a person have normal weight? Probability $P(18.5 \leq X \leq 25)$

Data Science

Random Variables

Given a random process with sample space Ω. A function X, mapping every possible result $\omega \in \Omega$ to a real number is called **random variable**:

$$X : \Omega \to \mathbb{R}, \omega \mapsto X(\omega) = x.$$

x is also known as the **realization of the random variable**.

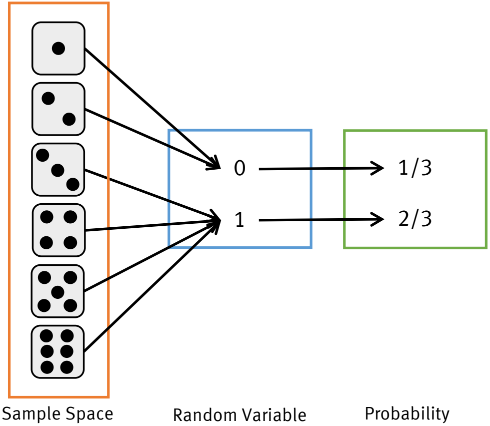
Data Science

Random Variables

Let X be a real valued random variable, then we call the probability measure $P(X \in A), A \subset \mathbb{R}$ **probability distribution** of X

Data Science

Random Variables Discrete random variable



Example

X: The value of a thrown dice is larger than 2:

$$P(X(\omega) = 0) = P(\omega \leq 2) = p = \frac{1}{3}$$

$$\begin{aligned} P(X(\omega) = 1) &= P(\omega \geq 2) \\ &= 1 - P(\omega \leq 2) = 1 - p \\ &= \frac{2}{3} \end{aligned}$$

$$P(X \in \{0, 1\}) = 1$$

$$P(X \notin \{0, 1\}) = 0$$

Random Variables

Discrete distributions

A random variable X has a **discrete distribution** or the distribution of X is called **discrete distribution** if the sample space

$$\{x|X(\omega) = x, \omega \in \Omega_0\}$$

has countable many elements. The set

$$\Omega = \{x|P(X = x) > 0, x \in \mathbb{R}\}$$

is called **support** of X.

The support contains all possible realizations of X.

Data Science

Discrete distributions

The **probability density** (short: density) of a discrete random variable X is defined by

$$f(x) = \begin{cases} P(X = x) & x \in \Omega \\ 0 & otherwise \end{cases}.$$

Properties of the density a discrete random variable X

- $\sum_{x \in \Omega} f(x) = 1$
- $P(x \in A) = \sum_{x \in A} f(x), A \subset \Omega$

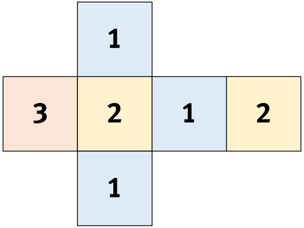
Example: 1-2-3 dice

Consider a dice with six sides and three values. The value 1 is on three sides, the value 2 on two sides and the value 3 on one side. X **gives the thrown value**:

- Support: $\Omega = \{1, 2, 3\}$
- Probability density:
  - $P(X = 1) = f(1) = \frac{3}{6} = \frac{1}{2}$
  - $P(X = 2) = f(2) = \frac{2}{6} = \frac{1}{3}$
  - $P(X = 3) = f(3) = \frac{1}{6}$

  or

$$f(x) = \frac{4 - x}{6} \text{ for } x = 1, 2, 3$$



Let X be a discrete random variable, then the function

$$F : \mathbb{R} \to \mathbb{R}, x \mapsto F(x) = P(X \leq x)$$

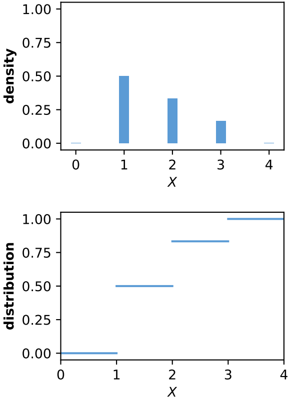
is called (cumulative) **distribution function** of X.

Properties

- $F(X) = P(X \leq x) = \sum_{\{z|z \leq x, z \in \Omega\}} f(z)$
- $0 \leq F(X) \leq 1$
- $F(X)$ monotone increasing

Data Science

Discrete distributions



Example: 1-2-3 dice

X: Number of the dice, $\Omega = \{1, 2, 3\}$

- $P(X < 1) = 0$
- $F(1) = P(X \leq 1) = f(1) = \frac{1}{2}$
- $F(2) = P(X \leq 2) = f(1) + f(2) = \frac{5}{6}$
- $F(3) = P(X \leq 3) = f(1) + f(2) + f(3) = 1$

Data Science

Discrete distributions

For a discrete random variable X with probability density $f(x)$ and support Ω we call the sum

$$E(X) = \sum_{x \in \Omega} x \cdot f(x)$$

the **expected value** of X. The sum

$$\sigma^2 = Var(X) = \sum_{x \in \Omega} (x - E(x))^2 \cdot f(x),$$

is called **variance**. The square root $\sqrt{\sigma^2}$ is called standard deviation of X.

- The expected value can be seen as the mean value in the long run (many repetitions)
- The variance is a measure of the spread around the expected value

Task: Compute the expected value and variance of the *1-2-3 dice* example!

Data Science

Discrete distributions

Task: Compute the expected value and variance of the *1-2-3 dice* example! X: Number of dice, $\Omega = \{1, 2, 3\}$

$$E(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{2}{6} + 3 \cdot \frac{1}{6} = \frac{10}{6} = \frac{5}{3} = 1.66\ldots$$

$$\begin{aligned} Var(X) &= \left(1 - \frac{5}{3}\right)^2 \frac{1}{2} + \left(2 - \frac{5}{3}\right)^2 \frac{1}{3} + \left(3 - \frac{5}{3}\right)^2 \frac{1}{6} \\ &= \frac{4}{9} \cdot \frac{3}{6} + \frac{1}{9} \cdot \frac{2}{6} + \frac{16}{9} \cdot \frac{1}{6} = \frac{5}{9} \end{aligned}$$

$$\sigma = \sqrt{Var(X)} = \frac{\sqrt{5}}{3}$$

Let’s take a look

Data Science

Discrete distributions

Let X be a random variable and $a, b \in \mathbb{R}$ constant values. Then there holds:

- $Y = aX + b$ is a random variable with

$$E(Y) = E(aX + b) = aE(X) + b$$
$$Var(Y) = Var(aX + b) = a^2 Var(X)$$

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

Data Science

Discrete distributions

Depending on the situation, e.g. the observed random process, there are different proper distribution functions for a random variable.

Let $\mathcal{D}$ be a distribution named *dist* with probability density f and distribution function F. If a random variable follows this distribution we call the random variable *dist*-distributed or $X \sim \mathcal{D}$.

In the following: Some example distributions!

Data Science

Discrete distributions

A **Bernoulli-experiment** is a random experiment with the following structure:

- each experiment only considers whether event A occurs or not ($\overline{A}$ occurs)
- Then $P(A) = p$ and $P(\overline{A}) = 1 - p$

If a Bernoulli experiment is carried out n times independently of each other, then the distribution of the number of successes follows a binomial distribution

Examples

- Urn model with replacements
- n-times throwing a dice: Probability that the 6 is thrown x-times
- Number of heals of n treated patients, Number of defect parts in the case of n produced parts.

Data Science

Discrete distributions

$Bin(n, p)$-distribution

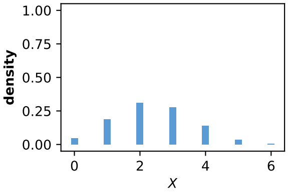
A discrete random variable X with support $\{0, 1, \ldots, n\}$ has a **binomial distribution** ($X \sim Bin(n, p)$) with parameter n and p, if the probability density of X is given by:

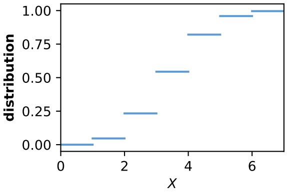
$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0, 1, \ldots, n\} \\ 0 & otherwise \end{cases}$$

For $X \sim Bin(n, p)$ there holds $E(X) = np$ and $Var(X) = np(1-p)$.

Data Science

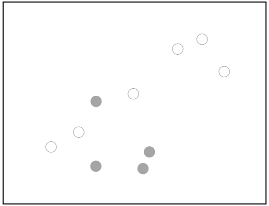
Discrete distributions





Example: Urn model with 10 balls (4 white, 6 black)

- Drawing six balls with repetition: X: number of white balls.
- $f(x) = P(X = x) = \binom{6}{x} 0.4^x 0.6^{6-x}$
- $X \sim Bin(6, 0.4)$



Data Science

Discrete distributions

In the case of a **singe Bernoulli-experiment** the random variable can be given as

$$X = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{if } \overline{A} \text{ occurs} \end{cases}$$

with distribution function

$$f(x) = P(X = x) = \begin{cases} p^x (1-p)^{1-x} & x \in \{0, 1\} \\ 0 & \textit{otherwise} \end{cases}$$

which equals to the distribution $Bin(1, p)$ ($X \sim Bin(1, p)$).

Reminder: Laplace-experiment

A random experiment is called **Laplace experiment**, if all possible results of the experiment have the same chance to occur.

Examples

- Throwing a dice, coin or equals
- Drawing a card from a pile
- Spinning the fortune wheel

Data Science

Discrete distributions

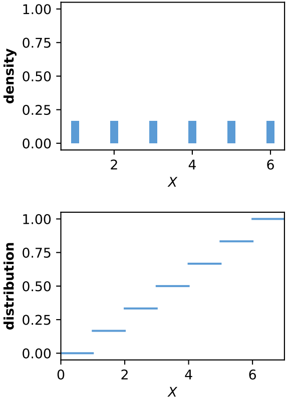
$DU(m)$-distribution

A discrete random variable X with finite support $\Omega = \{x_1, \ldots, x_m\}$ has a **(discrete) uniform distribution** ($X \sim DU(m)$), if the probability density function is given by

$$f(x) = \begin{cases} \frac{1}{m}, & x \in \Omega \\ 0 & otherwise \end{cases}$$

For X discrete uniform distributed with support $\Omega = \{1, \ldots, m\}$ there holds:

$$E(X) = \frac{m+1}{2}, E(X^2) = \frac{(m+1)(2m+1)}{6}, Var(X) = \frac{m^2-1}{12}$$



Example: Throwing a fair dice

X="Thrown value", $\Omega = \{1, \dots, 6\}$, $f(x) = \frac{1}{6} \forall x \in \Omega$.

$$E(X) = \sum_{x \in \Omega} x \cdot f(x) = \frac{1}{6} \cdot (1 + \dots + 6) = 3.5$$

$$Var(x) = \sum_{x \in \Omega} (x - 3.5)^2 \cdot f(x) = \dots = \frac{35}{12}$$

The distributions shown are just two examples. There are other distributions for a wide variety of requirements!

Further distributions

- **Poisson distribution:** e.g. counting rare events in a defined period
- **hypergeometrical distribution:** e.g. urn model without replacements
- **Geometric distribution:** e.g. number of tries till first success
- ...

Random Variables

Continuous distributions

Data Science

Continuous distributions

A random variable X has a **continuous distribution**, if there exists a function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) \geq 0$ for all $x \in \mathbb{R}$ such that

$$P(X \leq x) = F(X) = \int_{-\infty}^{x} f(t)dt$$

for all $x \in \mathbb{R}$ holds.

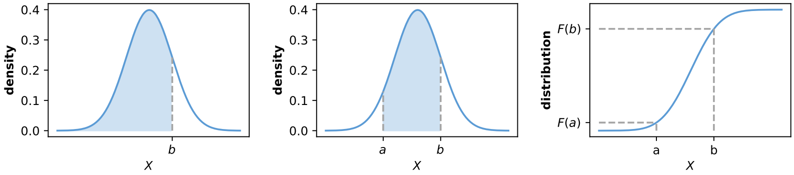
The function $f(x)$ is called **(probability) density** of X and $F(X)$ is called **distribution function**.

Note: $\{x|X(\omega) = x, \omega \in \Omega_0\}$ is a range or a union of ranges.

Data Science

Continuous distributions

- $F(x)$ is continuous and monotone increasing with values in $[0, 1]$.
  $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$ and $F(\infty) = \lim_{x \to \infty} F(x) = 1$
- $P(X \leq b) = P(X < b) = F(b) = \int_{-\infty}^{b} f(t)dt$
- $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$



Properties of densitiy function f of a continuous random variable

- Standardization: $\int\limits_{-\infty}^{\infty} f(x)dx = 1$
- $f(x) = \frac{dF(x)}{dx} = F'(x)$ for all x with f continuous in x.
- The probability of an event A is given by

$$P(X \in A) = \int\limits_{A} f(x)dx$$

- $f(x) \neq P(X = x)$ and $P(X = x) = 0$ for all $x \in \mathbb{R}$

The support Ω is given by all values $x \in \mathbb{R}$ with $f(x) > 0$, i.e.

$$\Omega = \{x | f(x) > 0\}$$

Data Science

Continuous distributions

Task: For which a is the following function a density function of a random variable?

$$f(x) = \begin{cases} 2x & 0 \leq x \leq a \\ 0 & otherwise \end{cases}$$

Data Science

Continuous distributions

Task: For which a is the following function a density function of a random variable?

$$f(x) = \begin{cases} 2x & 0 \le x \le a \\ 0 & otherwise \end{cases}$$

Compute the integral to check for which a it equals to 1:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{a} 2xdx = [x^2]_0^a = a^2$$

For $\int_{-\infty}^{\infty} f(x)dx = 1$ and $0 \le a$ there must hold $a = 1$.

Task: What is the distribution function of X, if the following density function is given:

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

Data Science

Continuous distributions

Task: What is the distribution function of X, if the following density function is given:

$$f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

Distribution function:

$$F(X) = \int_{-\infty}^{x} (t)dt = \begin{cases} \int_{-\infty}^{x} 0dt & x < 0 \\ \int_{-\infty}^{0} 0dt + \int_{0}^{x} 2tdt = x^2 & 0 \leq x \leq 1 \\ \int_{-\infty}^{0} 0dt + \int_0^1 2tdt + \int_{1}^{x} 0dt = 1 & x > 1 \end{cases}$$

Expected value and variance

Let X be a continuous random variable with density f, then the **expected value** of X is given by

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

The **variance** of X is defined by

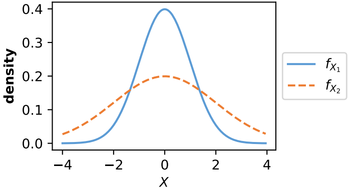
$$\sigma^2 = Var(x) = E((X - E(X))^2) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)dx$$

The positive square root σ is denoted **standard deviation** of X.

Data Science

Continuous distributions

Similar to the discrete case, the expected value gives the mean value in the long run and the variance defines the spread around the mean value.



Density functions of the random variables X_1 and X_2 with

$$E(X_1) = E(X_2) \text{ and } Var(X_1) > Var(X_2)$$

Let’s take a look

Properties of expected value and variance

For $a, b \in \mathbb{R}$ being constant and $g(x)$ a real function, there holds

- $E(aX + b) = aE(X) + b$
- $E(g(X)) = \int\limits_{-\infty}^{\infty} g(x)f(x)dx$
- $Var(X) = E(X^2) - (E(X))^2$ (Steiner's theorem)
- $Var(aX + b) = a^2 Var(X)$

Task: Compute the expected value and variance of a continuous random variable with the following density function:

$$f(x) = \begin{cases} 2x & 0 \le x < 1 \\ 0 & \textit{otherwise} \end{cases}$$

Data Science

Continuous distributions

Task: Compute the expected value and variance of a continuous random variable with the following density function:

$$f(x) = \begin{cases} 2x & 0 \leq x < 1 \\ 0 & otherwise \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{1} 2x^2 dx = [\frac{2}{3}x^3]_0^1 = \frac{2}{3}$$

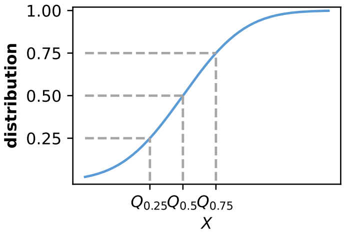
$$Var(X) = E(X^2) - (E(X))^2 = \int_{0}^{1} x^2 f(x)dx - (\frac{2}{3})^2 = \int_{0}^{1} 2x^3 dx - \frac{4}{9} = \frac{1}{18}$$

Data Science

Continuous distributions

The **p-quantile** of the distribution variable of the random variable X is the value Q_p for which there holds

$$p = \int_{-\infty}^{Q_p} f(x)dx = P(x \leq Q_p) = F(Q_p).$$



The 0.5-quantile of X is called **median** of X. The 0.25- and 0.75-quantil $Q_{0.25}$ and $Q_{0.75}$ is called **upper** and **lower quartile.**

Data Science

Continuous distributions

Task: Compute the median of the random variable X with the following distribution function:

$$F(X) = \begin{cases} 1 & x > 1 \\ x^2 & 0 \leq x \leq 1 \\ 0 & x < 0 \end{cases}$$

Data Science

Continuous distributions

Task: Compute the median of the random variable X with the following distribution function:

$$F(X) = \begin{cases} 1 & x > 1 \\ x^2 & 0 \leq x \leq 1 \\ 0 & x < 0 \end{cases}$$

1. Find $Q_{0.5}$ such that $F(Q_{0.5}) = 0.5$
2. $F(Q_{0.5}) = Q_{0.5}^2$
3. $Q_{0.5}^2 = 0.5$ only if $Q_{0.5} = \sqrt{0.5}$

Thus, the median of the random variable X equals to $\sqrt{0.5}$.

Uniform distribution

A continuous random variable X has a **(continuous) uniform distribution** (rectangular distribution, $X \sim Unit(a, b)$) with parameter $a, b \in \mathbb{R}$ and $a < b$, if the probability density function is given by

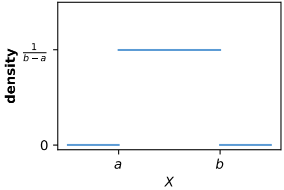
$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & otherwise \end{cases}$$

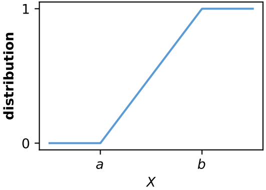
For $X \sim Unif(a, b)$ there holds:

$$E(X) = \frac{a+b}{2} \text{ and } Var(X) = \frac{(b-a)^2}{12}$$

Data Science

Continuous distributions





Distribution function of a uniform distribution:

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

Example

Waiting time for metro without knowledge of the timetable.

Data Science

Continuous distributions

Often: One wants to generate values following an arbitrary distribution F. This can be derived from a uniform distribution.

Let $U \sim Unif(0,1)$, F a distribution function and F^{-1} the corresponding inverse distribution function. Then, the random variable $X = F^{-1}(U)$ has the distribution function F.

- Start with random numbers $u_1, \ldots, u_n$ from a $Unif(0,1)$-distribution (e.g. by a list of pseudo random numbers)
- Compute $x_1 = F^{-1}(u_1), \ldots, x_n = F^{-1}(u_n)$

Normal distribution is the "most important" distribution in statistics!

- Known with different names: Gaussian distribution, bell-shaped curve, ...
- Many variables in natural science are normally distributed:
  - people's heights, IQ scores, examination grades, sizes of snowflakes, lifetimes of lightbulbs, weights of loaves of bread, milk production of cows, ...
  - **errors in measurements**

Central limit theorem

In short: The sum of many random variables with arbitrary distribution is nearly normal distributed.

Data Science

Continuous distributions

Definition: Normal distribution

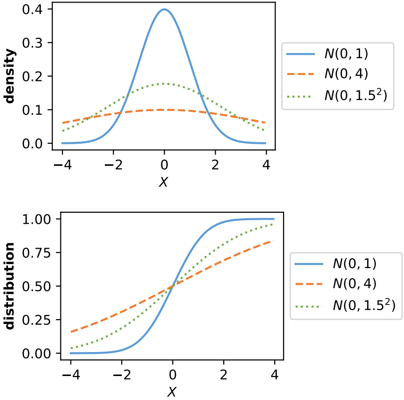
A continuous random variable X has a **normal distribution** ($X \sim N(\mu, \sigma^2)$) with parameter $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if the probability density is given by

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

For $X \sim N(\mu, \sigma^2)$ there holds $E(X) = \mu$ and $Var(X) = \sigma^2$

Data Science

Continuous distributions



curve sketching of f:

- $f(x) > 0$ for all $x \in \mathbb{R}$
- $\lim_{x \to -\infty} f(x) = \lim_{x \to \infty} f(x) = 0$
- global maximum in μ
- symmetric around μ:

$$f(\mu - x) = f(\mu + x) \text{ for all } x > 0$$

- Two turning points in $w_1 = \mu - \sigma$ and $w_2 = \mu + \sigma$

Data Science

Continuous distributions

The normal distribution function $F(X) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\mu\sigma}} exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$ can only be approximated numerically!

For transformations $Y = aX + b$ with $X \sim N(\mu, \sigma^2)$ with constant values $a, b \in \mathbb{R}$ there holds $Y \sim N(a\mu + b, a^2\sigma^2)$

Standarization

A variable $X \sim N(\mu, \sigma^2)$ can be transformed into a standardized normal distributed variable:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

with distribution function $F(x) = P(X \leq x) = \Phi(\frac{x-\mu}{\sigma})$.

Definition standard normal distribution

$X \sim N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 1$ is called **standard normal distribution**. The corresponding density function is denoted with φ and Φ, i.e.

$$\varphi = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ and } \Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$$

Remark

$\Phi(-x) = 1 - \Phi(x)$

For $X \sim N(0, 1)$ there holds $E(X) = 0$ and $Var(X) = 1$

Data Science

Continuous distributions

The distribution function of the standard normal distribution $\Phi(x)$ and the quantile $\Phi(z_\beta) = P(Z \leq z_\beta) = \beta$ are given in different books.

Table of Standard Normal Probabilities

Z	F(Z)	Z	F(Z)	Z	F(Z)	Z	F(Z)
-3.00	0.0013	-1.48	0.0694	0.04	0.5160	1.56	0.9406
-2.96	0.0015	-1.44	0.0749	0.08	0.5319	1.60	0.9452
-2.92	0.0018	-1.40	0.0808	0.12	0.5478	1.64	0.9495
-2.88	0.0020	-1.36	0.0869	0.16	0.5636	1.68	0.9535
-2.84	0.0023	-1.32	0.0934	0.20	0.5793	1.72	0.9573
-2.80	0.0026	-1.28	0.1003	0.24	0.5948	1.76	0.9608
-2.76	0.0029	-1.24	0.1075	0.28	0.6103	1.80	0.9641
-2.72	0.0033	-1.20	0.1151	0.32	0.6255	1.84	0.9671
-2.68	0.0037	-1.16	0.1230	0.36	0.6406	1.88	0.9699
-2.64	0.0041	-1.12	0.1314	0.40	0.6554	1.92	0.9726
-2.60	0.0047	-1.08	0.1401	0.44	0.6700	1.96	0.9750
-2.56	0.0052	-1.04	0.1492	0.48	0.6844	2.00	0.9772
-2.52	0.0059	-1.00	0.1587	0.52	0.6985	2.04	0.9793
-2.48	0.0066	-0.96	0.1685	0.56	0.7123	2.08	0.9812
-2.44	0.0073	-0.92	0.1788	0.60	0.7257	2.12	0.9830
-2.40	0.0082	-0.88	0.1894	0.64	0.7389	2.16	0.9846

Example

Let X be the height of a six years old child. Assumption: $X \sim N(100, 25)$.
What is the probability that such a child is at most 110 cm large?

$$P(X \leq 110) = P\left(\frac{X - 100}{5} \leq \frac{110 - 100}{5}\right) = \Phi(2) = 0.977$$

- Probability that a child is at most 90 cm large?
- Probability that a child is between 90 and 110 cm large?

Hint:

x	-0.36	0	0.5	0.78	1	2	2.5
$\Phi(x)$	0.359	0.500	0.692	0.782	0.841	0.977	0.994

Data Science

Continuous distributions

The distributions shown are just two or three examples. There are other distributions for a wide variety of requirements!

Further

- **Exponential distribution** Survival times (e.g. for devices), Waiting times between two Poission events.
- **F / t / χ^2-distribution:** typical distributions of test statistics
- **logistic distribution:** modelling of dosis-effect relationship
- ...

Random Variables

Overview

Data Science

Overview: Discrete distribution functions

Name	Density	Support	Expected value	Variance
Binomial				
$Bin(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,1,\ldots,n\}$	np	$np(1-p)$
Discrete uniform				
$DU(m)$	$\frac{1}{m}$	$\{0,1,\ldots,m\}$	$\frac{m+1}{2}$	$\frac{m^2-1}{12}$

Data Science

Overview: Continuous distribution functions

Name	Density	Support	Expected value	Variance
Normal				
$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\phi\sigma}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mathbb{R}$	μ	σ^2
Standard normal				
$N(0,1)$	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$	$\mathbb{R}$	0	1
continuous uniform				
$Unif(a,b)$	$\frac{1}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Summary & Outlook

Data Science

Summary & Outlook: Summary

- You know the basics of random variables and their distributions
- You are able to compute the expected value and variance of random variables
- You know different types of distribution functions
- You know the normal distribution and are able to work with it

Statistical tests and **linear regression**

References

Data Science

Summary & Outlook: List of images

- https://commons.wikimedia.org/wiki/File:Table_of_Standard_Normal_Probabilities.png

Parts of the lecture base on the lecture ”Statistics” (FH Dortmund)

by

Prof. Dr. Sonja Kuhnt and Prof. Dr. Nadja Bauer.