Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

www.dewiljes-lab.com

19. Oktober 2022

Universität Potsdam

Organization

Tutorials and Problem sheets

Passing the problem sheets:

- 50% of the points need to be acquired to be allowed to participate in the exam
- Solutions of Problem sheets have to be handed online on Moodle
- possible to work and hand in in pairs

Tutorials: Martin Nicolaus (`jan.martin.nicolaus@uni-potsdam.de`)

- Wednesday 10:15-11:45 room 2.06. 1.01
- Thursday 12:15-13:45 room 2.05. 1.12

Exam: Monday 06.02.2023 at 12:00-14:00

Content

Course content

- Recap foundations of probability theory
- Introduction to concept of learning
- Linear regression
- Batch vs Sequential
- Generalised Linear Regression
- Nonlinear Optimisation - Stochastic gradient descent
- Parametrisation by means of Neural networks
- Classification
- Support Vector Machines
- VC Dimension
- Clustering
- Random Forest models
- Causality
- Principle Component Anaylsis
- Autoencoders
- Gaussian Processes
- Optimal transport
- Generative advisarior networks
- score functions

Brief reminder: foundations of probability theory

Probability space

Def: A probability space consists of three elements $(\Omega, \mathcal{F}, \mathbb{P})$

- A sample space, Ω which is the set of all possible outcomes.
- An event space, which is a set of events $\mathcal{F}$ (σ- algebra), an event being a set $A \subset \Omega$ of outcomes in the sample space
- A probability function $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$

Axioms by Kolmogorov

Axioms: For probability function $\mathbb{P}$ the following holds true:

(A1) $\quad 0 \leq \mathbb{P}(A) \leq 1$

(A2) $\quad \mathbb{P}(\Omega) = 1$

(A3) $\quad \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \quad \forall A, B$ with $A \cap B = \emptyset$

more general :

$$\mathbb{P}(A_1 \cup A_2 \cup \dots) = \sum_{i \geq 1} \mathbb{P}(A_i) \quad \text{for } A_k \cap A_l = \emptyset,\ k \neq l$$

Independence and conditional probability

Def: Two events A and B are called independent if the following equation holds

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Def: Given two events A and B with $\mathbb{P}(B) > 0$, the conditional probability of A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Random Variables

Def: A *random variable* (RV) is measurable function $X : \Omega \rightarrow \mathrm{E}$, from a set of possible outcomes, Ω, to a measurable space E

Example:

- **Discrete RV**:
  - Example tossing a coin, dice roll
  - $X(\omega) \in \underbrace{\{X_1, X_2, ...\}}_{\text{finite countable}}$ for $\omega \in \Omega$
- **Continuous RV:** Example process of measurement or production.
- The result of an experiment is described by a random variable (r.v) X or a set of random variables $(X_1, X_2, X_3, ...)$ is called a random process

Expectation

Definition: The *expectation* of a discrete RV is defined as follows:

$$\mathbb{E}(X) = \bar{X} = \sum_{i=1}^{n} x_i \, \mathbb{P}(X = x_i)$$

Definition: Analogously the *Expectation* of a continuous RV

$$\mathbb{E}(X) = \bar{X} = \int_{-\infty}^{\infty} x \, f(x) dx$$

Markov Inequality

Proposition: Let X be a positive random variable. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{1}$$

Proof:

Cumulative distribution function

Definition: The cumulative distribution function of a real-valued random variable X is the function given by

$$F_X(x) = \mathbb{P}(X \leq x)$$

Definition: The probability density function of a continuous random variable can be determined from the cumulative distribution function by differentiating

$$f(x) = \frac{dF(x)}{dx}$$

Bernoulli distribution

A random variable X is distributed according to the Bernoulli distribution with parameter $p \in (0, 1)$

$$X = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability } 1 - p \end{cases}$$

Proposition:

$$\mathbb{E}[X] = p \tag{2}$$
$$Var(X) = p(1 - p) \tag{3}$$

Notation: $X \sim$ Bernoulli

Normal Distribution

A normal or Gaussian distributed random variable $X : \Omega \to \mathbb{R}$ with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ has the following density

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

and expected value and variance

$$\mathbb{E}[X] = \mu$$
$$Var(X) = \sigma^2$$

$X \sim \mathcal{N}(\mu, \sigma)$

Normal Distribution

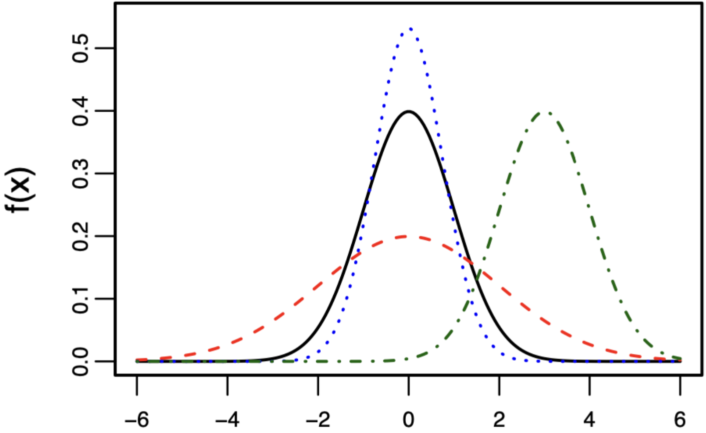


Abbildung 1: $\mu = 0$, $\sigma = 1$ (black), $\mu = 0$, $\sigma = 2$ (red), $\mu = 0$, $\sigma = 0.75$ (blue) and $\mu = 3$, $\sigma = 1$ (green)

Normal Distribution

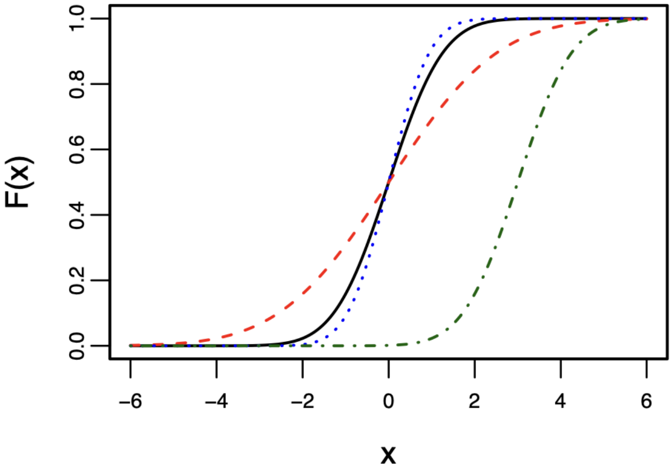


Abbildung 2: $\mu = 0$, $\sigma = 1$ (black), $\mu = 0$, $\sigma = 2$ (red), $\mu = 0$, $\sigma = 0.75$ (blue) and $\mu = 3$, $\sigma = 1$ (green)

Quantile

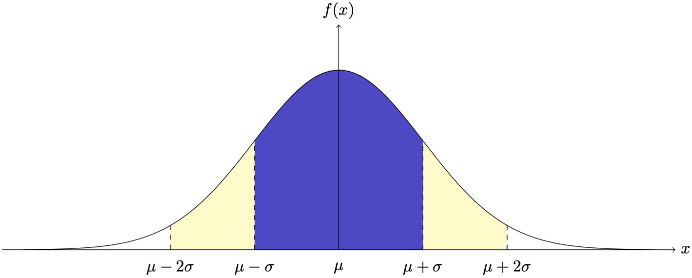


Abbildung 3: 60% of area under the curve (colored in blue) are in the $[\mu - \sigma, \mu + \sigma]$ interval and 95% of the area under the curve are in the interval $[\mu - \sigma, \mu + \sigma]$.

Standard normal distribution

A variable $X : \Omega \to \mathbb{R}$ follows a standard normal distribution, i.e., $X \sim \mathcal{N}(0, 1)$ if the associated density has the following form

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\Big(\frac{x^2}{2}\Big)\Big\}$$

with the associate cumulative distribution

$$\Phi(x) = \int_{-\infty}^{x} \phi(u) du \tag{4}$$

and quantile

$$z_\alpha = \Phi^{-1}(\alpha), \quad \alpha \in (0, 1) \tag{5}$$

Relationship between standard normal distribution and Normal distribution

$$F(x) = \Phi\Big(\frac{x - \mu}{\sigma}\Big) \tag{6}$$

Exponential Distribution

A random variable $X : \Omega \to \mathbb{R}$ follows the exponential distribution with parameters $\lambda > 0$ has the following density and cdf

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda \exp(-\lambda x) & x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp(-\lambda x) & x \geq 0 \end{cases}$$

and expected value and variance

$$\mathbb{E}[X] = \frac{1}{\lambda}$$

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

Notation: $X \sim \text{Exp}(\lambda)$ (often used for waiting times and lifetimes)

Exponential Distribution

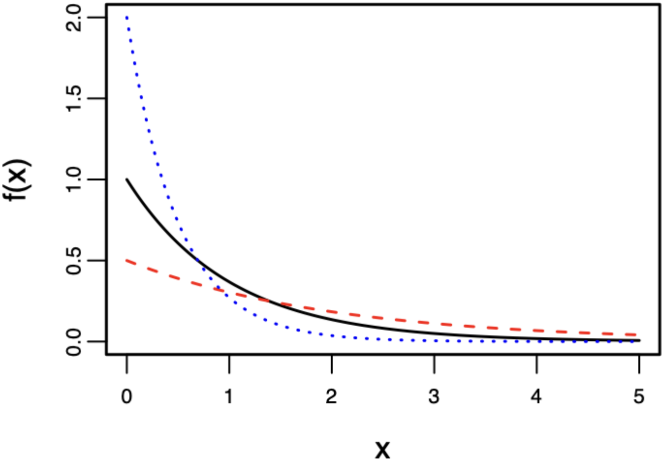


Abbildung 4: $\lambda = 1$ (black), $\lambda = 2$ (blue) and $\lambda = 1/2$ (red).

Exponential Distribution

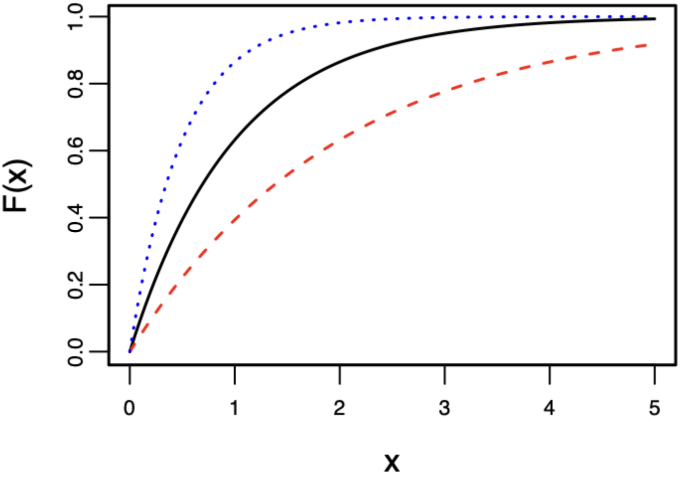


Abbildung 5: $\lambda = 1$ (black), $\lambda = 2$ (blue) and $\lambda = 1/2$ (red).

Example

Setting: The lifetime T of a computer chip is exponentially distributed, i.e., $T \sim \text{Exp}(\lambda)$ with expected lifetime of 15 weeks, i.e., parameter $\lambda = \frac{1}{15}$

Question:

- What is the probability that the computer chip is defect within the first 10 weeks?

- What is the probability that the computer chip will last at least 20 weeks?

Beta distribution

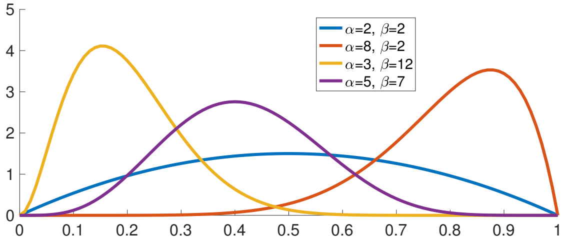
For a and b larger than zero and

$$f(x) = \frac{1}{\mathrm{B}(a,b)} x^{a-1}(1-x)^{b-1}.$$

where the normalization is given by

$$\mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 u^{a-1}(1-u)^{b-1}\,\mathrm{d}u$$

with $\Gamma(n) = (n-1)!$ being the gamma function.



Transformation

Reminder: for arbitrary g the following holds:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{7}$$

Proposition: Let g be a differentiable, strictly monoton function and X and random variable. Then $Y = g(X)$ has the following density

$$f_Y(y) = \left| \frac{1}{g'(g^{-1})(y)} \right| f_X(g^{-1}(y)), y \in E_Y \tag{8}$$

E_Y is given by the value space of X via

$$E_Y = g(E_X) = \{g(x) : x \in E_X\} \tag{9}$$

Variants of convergence

Let X be a random variable and $\{X_n\}_{n\in N}$ a sequence of random variables.

- $\{X_n\}$ converges to X almost surely, $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1 \tag{10}$$

- $\{X_n\}$ converges to X in probability $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0 \tag{11}$$

- $\{X_n\}$ converges to X in law (or in distribution), $X_n \xrightarrow{D} X$, if for any bounded continuous function f

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \tag{12}$$

Proposition: $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$

Samples

Definition: Let $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and $X_1, \dots, X_n$ be associated random variables. Realizations

$$x_1 := X_1(\omega), \dots, x_n := X_n(\omega) \tag{13}$$

are referred to as *samples* and n the sample size.

Estimator

Definition: A measurable function $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is referred to as *sample function*, *estimator* or *statistic*.

Note: we will also consider the composition:

$$\varphi(X) : \Omega \to \mathbb{R}^m \tag{14}$$

$$\omega \mapsto \varphi(X_1(\omega), \ldots, X_n(\omega)) \tag{15}$$

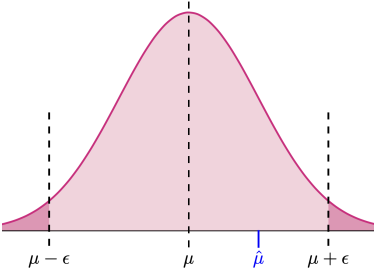
Monte-Carlo Approximation of a Mean

Def: Let X be a RV with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}[X]$ and $x_n \sim X$ be n i.i.d. realizations of X. The empirical mean built on n i.i.d. realizations is defined as

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

Understanding the tail probabilities

How accurately is the empirical estimate $\hat{\mu}$ approximating μ based on a set of samples?



Goals:

- investigate tail probabilities of $\hat{\mu} - \mu$
- derive bounds on $\mathbb{P}(|\hat{\mu} - \mu| \geq \epsilon)$

Subgaussian Random Variables

Subgaussianity

A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2\sigma^2/2\right)$$

Theorem: If X is σ-subgaussian, then for any $\epsilon \geq 0$

$$\mathbb{P}(X \geq \epsilon) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Proof: Let $\lambda > 0$, then

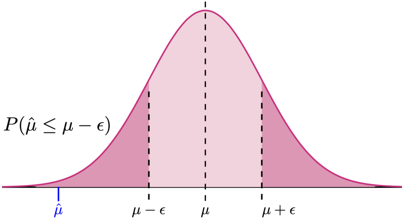
$$\begin{aligned}
\mathbb{P}(X \geq \epsilon) &= \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda\epsilon)) \\
&\leq \mathbb{E}[\exp(\lambda X)]\exp(-\lambda\epsilon) \quad \text{(Markov's inequality)} \\
&\leq \exp(0.5\lambda^2\sigma^2 - \lambda\epsilon) \quad \text{(subgaussianity)} \\
&= \exp(-0.5\epsilon^2/\sigma^2) \quad (\text{choose } \lambda = \epsilon/\sigma^2)
\end{aligned}$$

Confidence bounds

Corollary: Let $X_i - \mu$ be independent and σ-subgaussian for all i. Then

$$\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) \leq \underbrace{\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)}_{\delta}$$

$$\mathbb{P}(\hat{\mu} \leq \mu - \epsilon) \leq \underbrace{\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)}_{\delta}$$



for any $\epsilon \geq 0$.

Then we have

$$\hat{\mu} - \underbrace{\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}}_{\epsilon} \leq \mu \leq \hat{\mu} + \underbrace{\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}}_{\epsilon} \tag{16}$$

with probability at least $1 - \delta$

Monte-Carlo Approximation of a Mean

- Unbiased estimator: $\mathbb{E}[\mu_n] = \mu$ (and $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$])
- Weak law of large numbers: $\mu_n \xrightarrow{P} \mu$
- Strong law of large numbers: $\mu_n \xrightarrow{a.s.} \mu$
- Central limit theorem (CLT): $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$
- Finite sample guarantee:

$$\mathbb{P}\Big[|\frac{1}{n}\sum_{t=1}^{n} X_t - \mathbb{E}[X_1]| > \epsilon\Big] \leq 2\exp\Big(-\frac{2n\epsilon^2}{(b-a)^2}\Big) \tag{17}$$

Empirical variance

Definition: The empirical variance is defined by

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \tag{18}$$

Note: we will also use an analog notation for the random variables:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{19}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[S_n^2] = \sigma^2 \tag{20}$$

Empirical variance

Definition: The empirical variance is defined by

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \tag{21}$$

Note: we will also use an analog notation for the random variables:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \tag{22}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables. Then

$$S_n^2 = \frac{1}{n-1}\Big((\sum_{i=1}^{n} X_i^2) - n\bar{X}_n^2\Big) \tag{23}$$

Empirical variance

Proposition: Let $X_1, \ldots, X_n$ be independent and identical random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$. Then

$$\mathbb{E}[S_n^2] = \sigma^2 \tag{24}$$

Empirical standard deviation

Def: The empirical standard deviation is defined by

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \qquad (25)$$

Order statistic

Def: Let $(x_1, \ldots, x_n) \in \mathbb{R}^n$ be a sample set. One can order the elements in an increasing manner:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \tag{26}$$

Then $x_{(i)}$ is referred to as the i-th order statistic of the sample set.

Sample median

Def: The sample median of a set of samples if given by

$$\mathrm{Med}_n = \mathrm{Med}_n(x_1, \ldots, x_n) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \textit{ uneven} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n \text{ even} \end{cases}$$

Then $x_{(i)}$ is referred to as the i-th order statistic of the sample set.

Truncated mean

Def: The truncated mean samples $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is defined by

$$\frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

Empirische α-Quantile

Def: Let $(x_1, \ldots, x_n) \in \mathbb{R}^n$ be a set of samples and $\alpha \in (0, 1)$. The empirical α Quantil is defined by

$$q_\alpha = \begin{cases} x_{\lfloor n\alpha \rfloor + 1} & \text{falls } n\alpha \notin \mathbb{N} \\ \frac{1}{2}\left(x_{\lfloor n\alpha \rfloor} + x_{\lfloor n\alpha \rfloor + 1}\right) & \text{falls } n\alpha \in \mathbb{N} \end{cases}$$