



Statistics for Data Science

MSc Data Science WiSe 2020/21

Univ.-Prof. Dr. Dirk Ostwald

(1) Introduction

Introduction

- Data science
- Statistics
- Statistics for Data Science
- Exercises

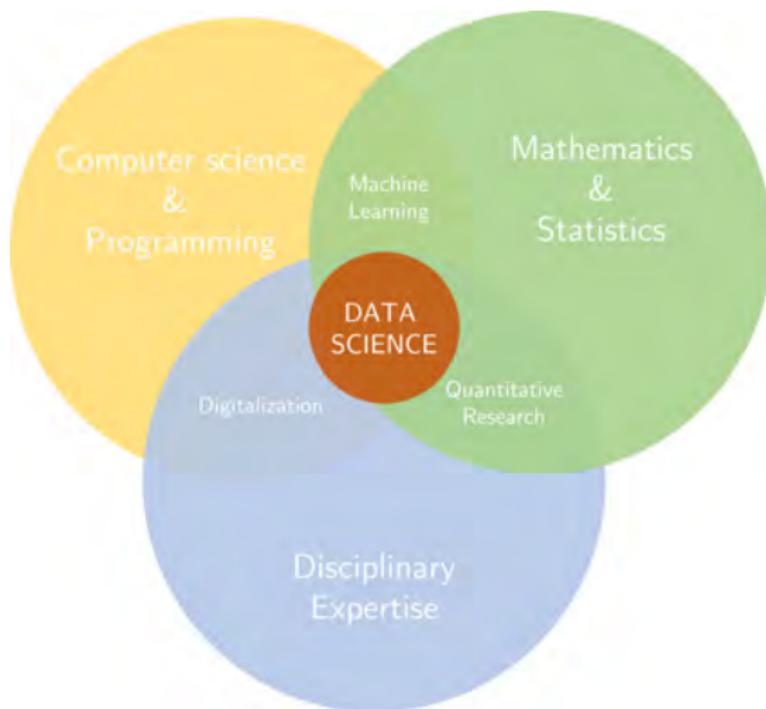
Introduction

- **Data science**
- Statistics
- Statistics for Data Science
- Exercises

Data science

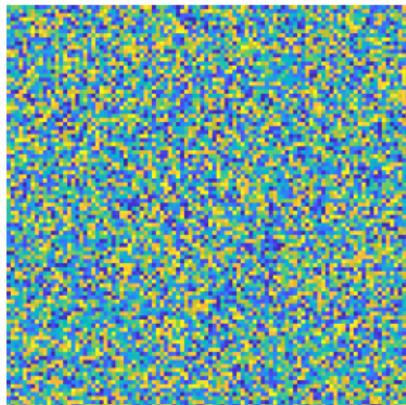
The art of creating meaning from data

Data science



Data analysis is data reduction

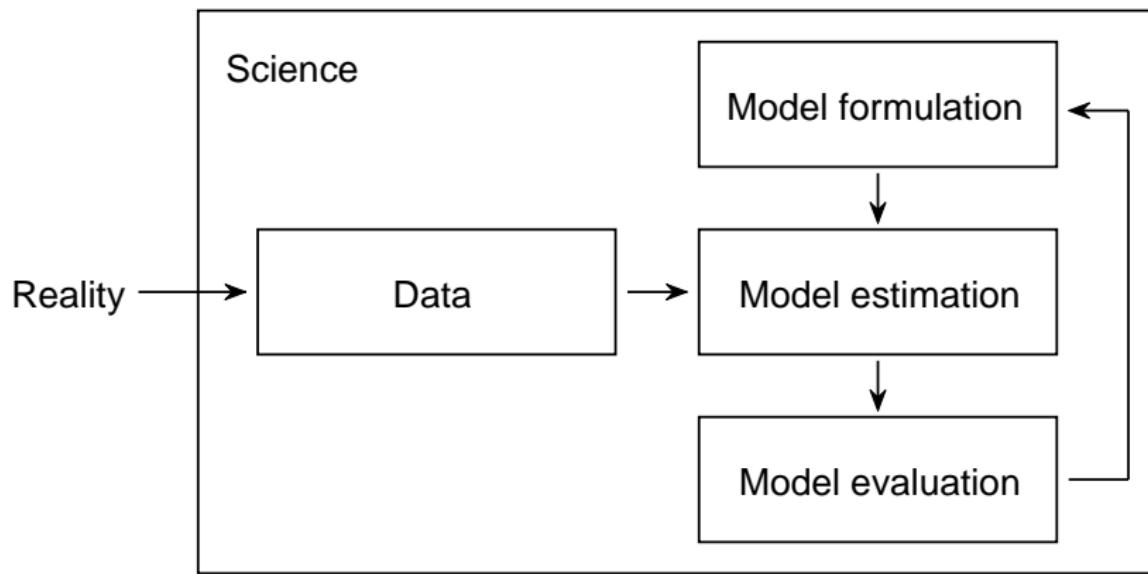
Raw Data



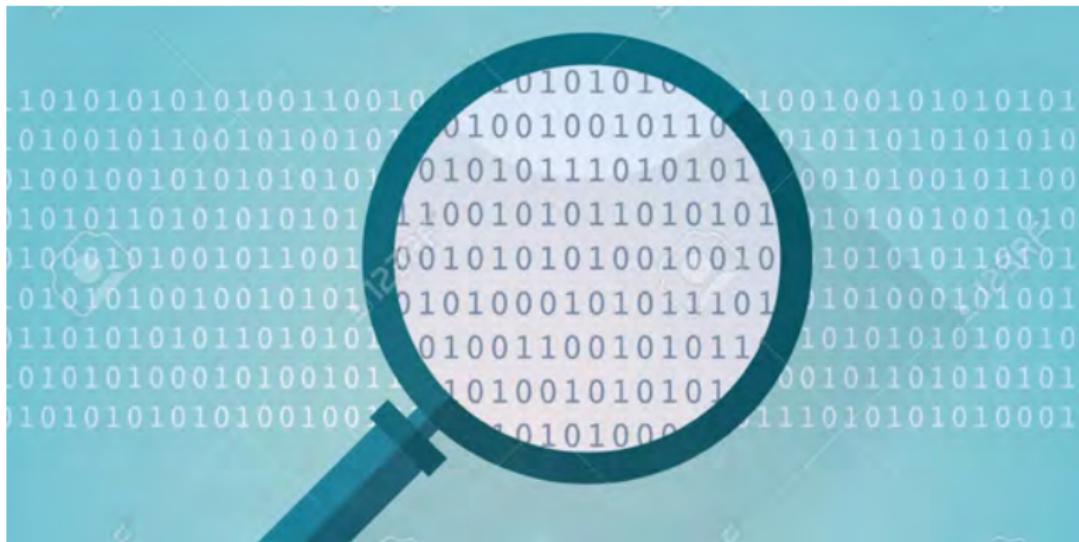
Reduced Data



Data analysis is model-based



Data analysis is an interpretative device



Data science = Statistics = Machine learning = Artificial intelligence

Statistics

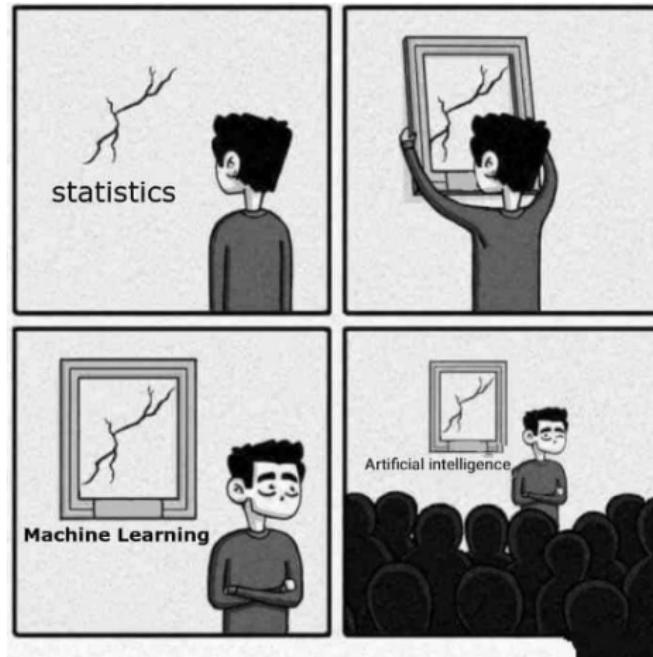
- Probabilistic models
- Theoretical analysis
- Optimality
- Asymptotics
- Science philosophy

Machine learning

- Deterministic models
- Classification
- Bayesian models
- Benchmarking
- Applications

Artificial intelligence

- Deep learning
- Reinforcement
- Decisions
- Data analysis
- Hype



www.instagram.com/sandserifcomics/



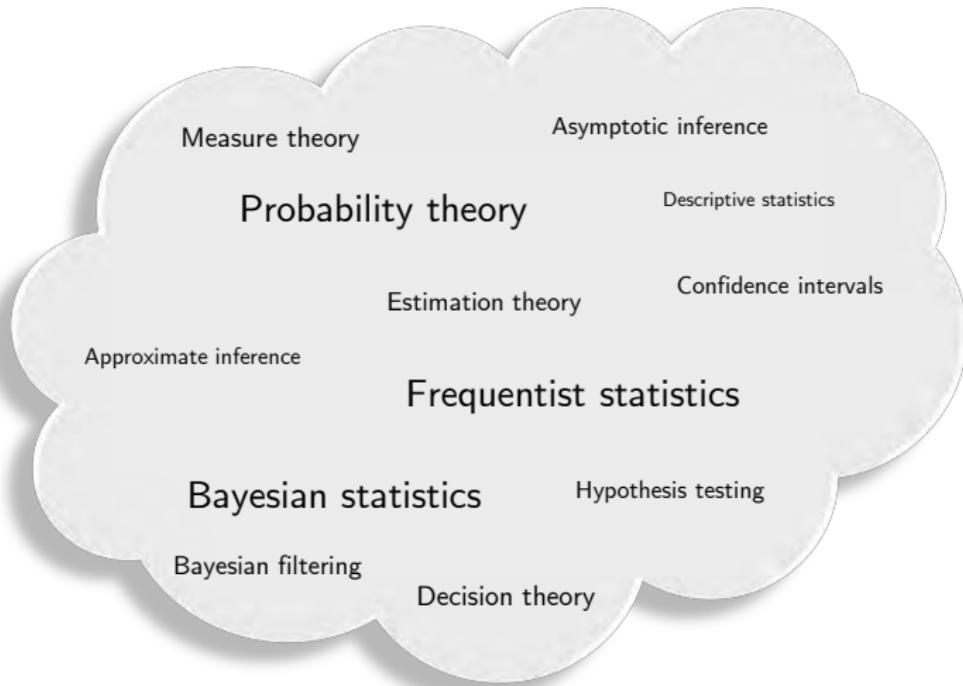
www.oak-tree.tech/blog/ml-models

Introduction

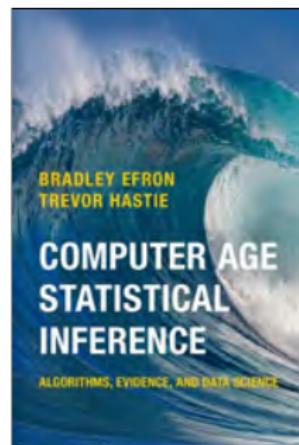
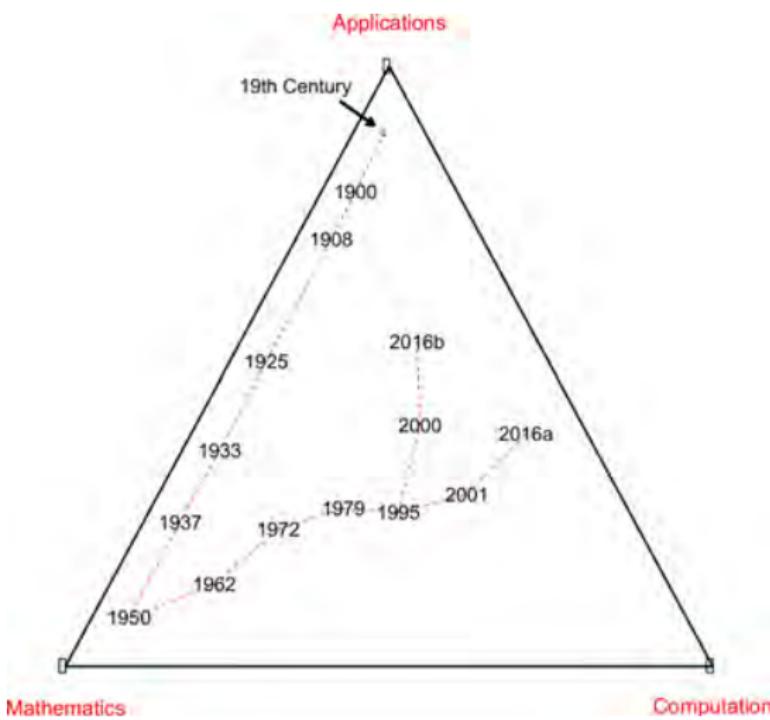
- Data science
- **Statistics**
- Statistics for Data Science
- Exercises

Statistics

The art of creating meaning from data
and quantifying its associated uncertainty



Historical development of statistics



Historical development of statistics

- 1900 Karl Pearson's chi-square test
- 1908 Student's t statistic
- 1925 Fisher's Statistical Methods for Research Workers
- 1933 Neyman and Pearson's optimal hypothesis testing
- 1937 Neyman's confidence intervals
- 1950 Wald's statistical decision theory
- 1950 Savage's & de Finetti's Bayesian decision theory
- 1961 Raiffa & Schlaifer's Applied statistical decision theory
- 1962 Tukey's The future of data analysis
- 1971 Lindley's Bayesian statistics
- 1972 Cox's proportional hazards
- 1979 Bootstrap and MCMC
- 1995 False discovery rate and LASSO
- 1996 Support vector machines
- 2000 Microarray and neuroimaging multiple testing
- 2010 Resurgence of neural networks as deep learning
- 2015 Data science

Central postulates of Probability theory

- Chance processes can be described mathematically.
- Mathematics can be used to make predictions about random events.
- Reasoning about uncertain events is naturally related to measuring volumes.

Central postulates of Frequentist inference

- Probabilities are interpreted as limiting relative frequencies and are considered objective properties of the real world.
- Parameters are fixed, unknown constants, referred to as *true, but unknown* values. No probability statements are made about parameters.
- Statistical procedures are designed to have good long run frequency properties and are typically assessed by studying their sampling distributions.

Central postulates of Bayesian inference

- Probabilities are interpreted as degrees of belief, not limiting frequencies. Statements like “the probability that it will rain this afternoon is 0.5” are meaningful.
- Parameters are fixed, unknown constants, about which probabilistic statements quantifying our uncertainty about their value can be made.
- Probabilistic statements about parameters are made with the help of probability distributions, from which further inferences, such as point or interval estimates, can be derived.

Statistics known as Machine learning

- Principal component and independent component analysis
- Logistic regression and linear discriminants
- Support vector machines, kernel methods
- Latent variable and graphical models
- Generalized linear models, neural networks, deep learning
- Gaussian process regression

Statistics known as Artificial intelligence

- Markov decision processes
- Partially observable Markov decision processes
- Reinforcement learning

Examples of community nomenclatures

Statistics	Machine Learning	Meaning
Data	Training data	Data
Estimation	Learning, Training	Using data to estimate parameters
Frequentist inference	-	Optimal many samples methods
Bayesian inference	Bayesian inference	Data-based uncertainty updating
Covariates	Features	Structural and known data predictors

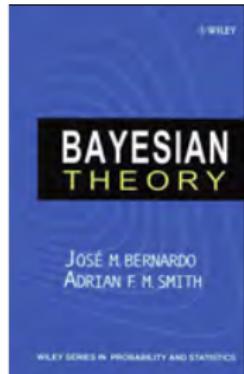
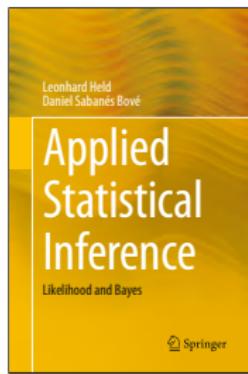
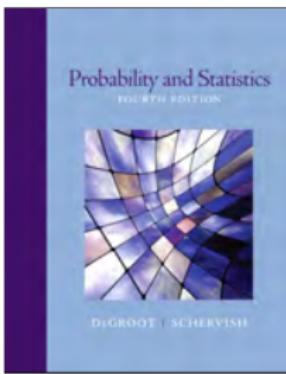
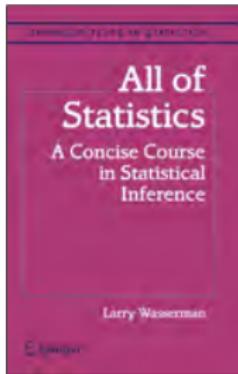
Introduction

- Data science
- Statistics
- **Statistics for Data Science**
- Exercises

Statistics for Data Science

Unit	Date	Theme
(1) Introduction	06.11.2020	
(2) Probability spaces	13.11.2020	Probability theory
(3) Random variables	20.11.2020	Probability theory
(4) Joint distributions	27.11.2020	Probability theory
(5) Transformations	04.12.2020	Probability theory
(6) Expectation and covariance	11.12.2020	Probability theory
(7) Inequalities and limits	18.11.2020	Probability theory
(8) Foundations and maximum likelihood	08.01.2021	Frequentist inference
(9) Finite-sample estimator properties	15.01.2021	Frequentist inference
(10) Asymptotic estimator properties	22.01.2021	Frequentist inference
(11) Confidence intervals	29.01.2021	Frequentist inference
(12) Hypothesis testing	05.02.2021	Frequentist inference
(13) Conjugate inference	12.02.2021	Bayesian inference
(14) Numerical methods	19.02.2021	Bayesian inference
(15) Variational inference	26.02.2021	Bayesian inference

Key references



Course components

Component	Aims
Vorlesung	Core course content, knowledge acquisition, breadth
Übung	Active participation, knowledge consolidation, depth
Study questions	Focus, memorization support, examination
Theoretical exercises	Theoretical depth, self-study
Programming exercises	Intuition, Python training, application
Exam	Knowledge reproduction

Course requirements for MSc Data Science students

Vorlesung

- Written exam, pass or fail, 30 questions, 90 min
- 2 questions for each of the 15 units, 1 point per question
- Exam question pool := Study questions
- The study questions pool provided with the lecture slides is final
- ≥ 15 points and ≥ 1 point on the 2 question of each unit to pass
- Exam date 05.03.2021, exam resit date 26.03.2021

Übung

- Presentation of one theoretical or programming exercise in class
- Python script with honest solution attempts of all programming exercises
- Programming exercises honest solutions attempts deadline 26.03.2021

Course requirements for Non-MSc Data Science students

Vorlesung

- Graded exam
- 2 questions for each of the 15 units, 1 point per question
- 30 questions, 90 min
- Exam question pool := Study questions
- The study questions pool provided with the lecture slides is final
- ≥ 15 points to pass.
- Exam date 05.03.2021, exam resit date 26.03.2021

Übung

- Python script with honest solution attempts of all programming exercises
- Programming exercises honest solutions attempts deadline 26.03.2021

Übung active participation criteria

- 15 min presentation of one theoretical or programming exercise in class
 - Theoretical exercise ⇒ Beamer presentation
 - Programming exercise ⇒ Jupyter Notebook
- Except on 13.11.2020, active participation can be failed
- Think of the presentation and subsequent discussion as a mini oral exam
- Binary feedback will be provided only on the Tuesday before the presentation
- The link to the exercise pool is provided by email
- The link to the exercise sign up form is provided by email
- The programming exercise pool provided with the lecture slides is final

Introduction

- Data science
- Statistics
- Statistics for Data Science
- **Exercises**

Study questions

1. Give a definition of Data Science.
2. Give a definition of Statistics.
3. Name three central postulates of Probability theory.
4. Name three central postulates of Frequentist inference.
5. Name three scientists involved in the development of Frequentist statistics.
6. Name three central postulates of Bayesian inference.
7. Name three scientists involved in the development of Bayesian statistics.
8. Name five typical topics in Statistics.
9. Name three topics commonly discussed in Machine Learning.
10. Name three topics commonly discussed in Artificial Intelligence.

Exercises

Programming exercises

1. Sample a univariate Gaussian using `scipy.stats`.
2. Evaluate the PDF of a univariate Gaussian using `scipy.stats`.
3. Visualize the PDF of a univariate and a normalized sample histogram of samples from a univariate Gaussian with identical parameters on top of each other using Matplotlib.

(2) Probability spaces

Bibliographic remarks

The presented material is standard and can be found in any introductory textbook to statistical inference. Wasserman (2004, Chapter 1) and Casella and Berger (2012, Sections 1.1 - 1.3) and are closest in spirit. Further excellent introductions to modern probability theory include Billingsley (1995), Fristedt et al. (1998), and Rosenthal (2006).

Probability spaces

- Probability spaces
- Elementary probabilities
- Exercises

Probability spaces

- **Probability spaces**
- Elementary probabilities
- Exercises

Definition (Probability space)

A *probability space* is a triple $(\Omega, \mathcal{A}, \mathbb{P})$, where

- Ω is a set of elementary outcomes ω ,
- \mathcal{A} is a σ -algebra, i.e., \mathcal{A} is a set with the following properties,
 - $\Omega \in \mathcal{A}$,
 - \mathcal{A} is closed under the formation of complements, i.e., if $A \in \mathcal{A}$, then also $A^c := \Omega \setminus A \in \mathcal{A}$ for all $A \in \mathcal{A}$,
 - \mathcal{A} is closed under countable unions, i.e., if $A_1, A_2, \dots \in \mathcal{A}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.
- \mathbb{P} is a mapping $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ with the following properties:
 - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{A}$.
 - $\mathbb{P}(\Omega) = 1$.
 - If $A_1, A_2, \dots \in \mathcal{A}$ are pairwise disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$, which is referred to as σ -*additivity of \mathbb{P}* .

Probability spaces

FOUNDATIONS OF THE THEORY OF PROBABILITY

BY
A. N. KOLMOGOROV

Second English Edition

TRANSLATED BY
NATHAN MORRISON

WITH AN AFTERWORD BY
A. T. BHARUCHA-REID
UNIVERSITY OF SINGAPORE

CHELSEA PUBLISHING COMPANY
NEW YORK
1956

§ 3. Notes on Terminology

Remark 1. If two separate statements are each practically reliable, then we may say that simultaneously they are both reliable, although the sum of reliability of each individual involved in the process. If, however, the number of such statements is very large, then from the practical reliability of each, one cannot deduce anything about the simultaneous correctness of all of them. Therefore from the principle stated in (a) it does not follow that in a very large number of cases, in each of which the ratio n/m will differ only slightly from $P(A)$.

Remark 2. To an impossible event (an empty set) corresponds, in accordance with our axiom, the probability $P(\emptyset) = 0$, but the converse statement $P(A) = 0$ does not necessarily imply the impossibility of A . When $P(A) = 0$ from principle (b) all we can assert is that when the conditions Ω are realized but once, event A is practically impossible. It does not at all assert, however, that in a sufficiently long series of trials that event A will never occur. On the other hand, one can easily verify the principle (a) merely that when $P(A) = 0$ and n is very large, the ratio n/m will be very small (it might, for example, be equal to 1%).

§ 3. Notes on Terminology

We have defined the objects of our future study, random events, as sets. However, in the theory of probability many set-theoretic concepts are designated by other terms. We shall give here a brief list of such concepts.

Theory of Sets

1. A and B do not intersect, $AB = \emptyset$.
2. $AB \dots N = \emptyset$.
3. $AB \dots N = X$.
4. $A \dot{+} B \dot{+} \dots \dot{+} N = X$.

Random Events

1. Events A and B are incompatible.
2. Events A, B, \dots, N are incompatible.
3. Event X is defined as the simultaneous occurrence of events A, B, \dots, N .
4. Event X is defined as the occurrence at least one of the events A, B, \dots, N .

¹ Cf. 4.4. Formula (3).

§ 3. Elementary Theory of Probability

Theory of Sets

5. The complementary set \bar{A} .

6. $A = \emptyset$.

7. $A = E$.

8. The system \mathbb{W} of the sets A_1, A_2, \dots, A_n forms a decomposition of the set E if $A_1 + A_2 + \dots + A_n = E$.

9. The system \mathbb{W} of the sets A_1, A_2, \dots, A_n occurs. We therefore call A_1, A_2, \dots, A_n , the possible results of experiment E .

10. If B is a subset of A : $B \subset A$.

11. From the occurrence of event B follows the inevitable occurrence of A .

§ 4. Immediate Corollaries of the Axioms; Conditional Probabilities; Theorem of Bayes

From $A + \bar{A} = E$ and the Axiom IV and V it follows that

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

$$P(\bar{A}) = 1 - P(A). \quad (2)$$

Since $\bar{B} = \emptyset$, then, in particular,

$$P(\emptyset) = 0. \quad (3)$$

If A, B, \dots, N are incompatible, then from Axiom V follows the formula (the Addition Theorem):

$$P(A + B + \dots + N) = P(A) + P(B) + \dots + P(N). \quad (4)$$

If $P(A) > 0$, then the quotient

$$P_B(A) = \frac{P(AB)}{P(B)} \quad (5)$$

is defined to be the conditional probability of the event B under the condition A .

From (5) it follows immediately that

"The purpose of this monograph is to give an axiomatic foundation for the theory of probability. The author set himself the task of putting in their natural place, among the general notions of modern mathematics*, the basic concepts of probability theory-concepts which until recently were considered to be quite peculiar". (*e.g., set theory, mappings, Lebesgue integrals)

Kolmogorov, A.N. (1933) *Grundbegriffe der Wahrscheinlichkeitsrechnung*

Selected remarks

- Probability spaces are used as abstract models of random experiments.
- Probability spaces are special cases of measure spaces $(\Omega, \mathcal{A}, \mu)$.
- Measure spaces are mathematical models for assigning volume to sets.
- Elementary outcomes are “realized” according to $\mathbb{P}(\{\omega\})$.
- Probability spaces unify finite, countable (\mathbb{N}), and uncountable (\mathbb{R}) outcome sets.
- Probability spaces offer a language spanning discrete and continuous probability.
- σ -algebras are “complete sets of events (which comprise elementary outcomes)”.
- For finite outcome spaces, the typical σ -algebra is the power set $\mathcal{P}(\Omega)$.
- For the outcome space \mathbb{R} , the *Borel* σ -algebra fulfills the same function.
- The probabilistic characteristics of $(\Omega, \mathcal{A}, \mathbb{P})$ are defined by \mathbb{P} .
- Probability spaces often disappear behind random variables and distributions.
- Probability spaces are useful in the definition of random fields, Markov kernels, ...

Example (Throwing a dice)

- It is reasonable to consider the elementary outcome set $\Omega := \{1, 2, 3, 4, 5, 6\}$.
- The elementary outcome set $\Omega := \{\cdot, \dots, \dots, \dots, \dots, \dots, \dots\}$ also works.
- The power set $\mathcal{P}(\{1, 2, 3, 4, 5, 6\})$ contains all possible events A_i , for example:

Any number occurs	$A_1 = \Omega$
A number larger than 4 occurs	$A_2 = \{5, 6\}$
An even number occurs	$A_3 = \{2, 4, 6\}$
A six occurs	$A_4 = \{6\}$
One, three, or six occurs	$A_5 = \{1, 3, 6\}$
No number occurs	$A_6 = \emptyset$

- \mathbb{P} can be defined by specification of $\mathbb{P}(\{\omega\})$ for all $\omega \in \Omega$.
- Because the $\omega \in \mathcal{A}$ for which $\omega \in \Omega$ are pairwise disjoint, the probabilities of all events $A \in \mathcal{A}$ can be evaluated based on $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.
- A fair dice would have $\mathbb{P}(\{\omega\}) := 1/6$ for all $\omega \in \Omega$.
- A biased dice could have

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{6\}) := 1/9, \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) := 2/9.$$

Probability spaces

- Probability spaces
- **Elementary probabilities**
- Exercises

Definition (Probability measure)

Let Ω denote an outcome space and \mathcal{A} denote a σ -algebra on Ω . A function

$$\mathbb{P} : \mathcal{A} \rightarrow \mathbb{R}, A \mapsto \mathbb{P}(A) \quad (1)$$

that satisfies the following axioms

- (1) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{A}$,
- (2) $\mathbb{P}(\Omega) = 1$, and
- (3) if $A_1, A_2, \dots \in \mathcal{A}$ are disjoint, then $\mathbb{P}(\cup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ (σ -additivity)

is called a *probability measure* or *probability*.

Remarks

- $\mathbb{P}(A)$ can be interpreted as the idealized long run frequency of observing A .
- $\mathbb{P}(A)$ can be interpreted as the subjective degree of belief that A is true.
- Frequentist and Bayesian interpretations use the same formal framework.

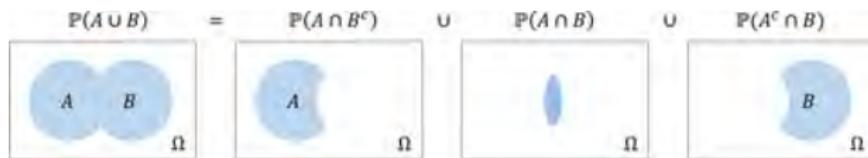
Elementary probabilities

Some properties of probabilities

- $\mathbb{P}(\emptyset) = 0$ and $0 \leq \mathbb{P}(A) \leq 1$
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Exemplary proof

With the additivity of \mathbb{P} for disjoint events, we have:



$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) \\&= \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B) + \mathbb{P}(A \cap B) - \mathbb{P}(A \cap B) \\&= \mathbb{P}((A \cap B^c) \cup (A \cap B)) + \mathbb{P}((A^c \cap B) \cup (A \cap B)) - \mathbb{P}(A \cap B) \\&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\end{aligned}\tag{2}$$

□

Definition (Independent events)

Two events $A \in \mathcal{A}$ and $B \in \mathcal{A}$ are independent, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B). \quad (3)$$

A set of events $\{A_i | i \in I\} \subset \mathcal{A}$ for an arbitrary index set is independent, if for every finite subset $J \subseteq I$

$$\mathbb{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbb{P}(A_j). \quad (4)$$

Remarks

- Independence of events often arises by design of the probabilistic model.
- Independence models the absence of deterministic and stochastic influences.
- Independence may follow from the design of a probabilistic model.
- Disjoint events with positive probability are not independent:

$\mathbb{P}(A)\mathbb{P}(B) > 0$, but $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$, thus $\mathbb{P}(A \cap B) \neq \mathbb{P}(A)\mathbb{P}(B)$.

- The arbitrary subset condition for $|I|$ events ensures their pairwise independence.

Definition (Conditional probability)

If $\mathbb{P}(B) > 0$, then the conditional probability of $A \in \mathcal{A}$ given $B \in \mathcal{A}$ is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (5)$$

For any fixed B , $\mathbb{P}(\cdot|B)$ is a probability measure, i.e., $\mathbb{P}(\cdot|B) \geq 0$, $\mathbb{P}(\Omega|B) = 1$, and if $A_1, A_2, \dots \in \mathcal{A}$ are disjoint, $\mathbb{P}(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$.

Remarks

- $\mathbb{P}(A|B)$ is “the fraction of times A occurs among those in which B occurs”.
- $\mathbb{P}(A|B)$ is the normalized version of $\mathbb{P}(A \cap B)$.
- Defining the joint probability $\mathbb{P}(A \cap B)$ defines $\mathbb{P}(A|B)$.
- The rules of probability apply to the events on the left of the vertical bar.
- Usually $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$, e.g., $\mathbb{P}(\text{Death}|\text{Hanging}) \neq \mathbb{P}(\text{Hanging}|\text{Death})$.
- An extension to $\mathbb{P}(B) = 0$ is possible, but technically more challenging.

Elementary probabilities

Theorem (Conditional probability for independent events)

If $A, B \in \mathcal{A}$ are independent, then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A). \quad (6)$$

Remark

- Given independence, knowledge of B does not change the probability of A .

Theorem (Joint and conditional probabilities)

For any $A, B \in \mathcal{A}$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A). \quad (7)$$

Remark

- Joint probabilities can be constructed from conditional and total probabilities.

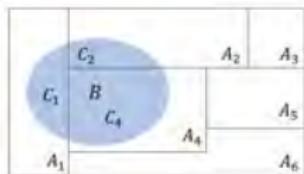
Theorem (Law of total probability)

Let A_1, \dots, A_k be a partition of Ω , i.e., $\cup_{i=1}^k A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for all $1 \leq i, j \leq k$ with $i \neq j$. Then, for any event $B \in \mathcal{A}$

$$\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i). \quad (8)$$

Proof

For $i = 1, \dots, k$, let $C_i := B \cap A_i$, so $\cup_{j=1}^k C_j = B$ and $C_i \cap C_j = \emptyset$ for $1 \leq i, j \leq k, i \neq j$.



Thus, $\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(C_i) = \sum_{i=1}^k \mathbb{P}(B \cap A_i) = \sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i)$.

□

Remark

- The total probability of B as a weighted average of conditional probabilities of B .

Theorem (Bayes theorem)

Let A_1, \dots, A_k be a partition of Ω with $\mathbb{P}(A_i) > 0$. If $\mathbb{P}(B) > 0$, then for each $i = 1, \dots, k$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \quad (9)$$

Proof

Using the definition of conditional probability twice and the law of total probability, we have

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^k \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \quad (10)$$

□

Remarks

- There is nothing “Bayesian” about Bayes theorem.
- Bayes theorem is a means to compute conditional probabilities.
- $\mathbb{P}(A_i)$ is often called *prior* and $\mathbb{P}(A_i|B)$ *posterior*.
- $\mathbb{P}(B|A_i)$ is sometimes called *likelihood* and $\mathbb{P}(B)$ *evidence*.

Probability spaces

- Probability spaces
- Elementary probabilities
- **Exercises**

Exercises

Study questions

1. Write down the definition of a probability space.
2. Give two interpretations for the probability $\mathbb{P}(A)$ of an event A .
3. Sketch the probability space model of throwing a dice.
4. Write down the definition of a probability measure.
5. For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, let $A, B \in \mathcal{A}$. What is the probability of the event that A or B are true?
6. Write down the definition of the independence of two events A and B and the definition of the independence of a set of events $\{A_i | i \in I\}$ with index set I .
7. Write down the definition of the conditional probability of an event given another B .
8. State the law of total probability.
9. What is the conditional probability of an event given an event B , if A and B are independent events? Justify your answer.
10. Write down and prove Bayes theorem.

Theoretical exercises

1. Develop a probability space model of tossing a coin (Casella and Berger, 2012, Example 1.2.5).
2. Develop a probability space model of throwing two dice (DeGroot and Schervish, 2012, Example 1.6.5).
3. Develop a probability space model of tossing a fair coin twice. Consider the events “heads appears on the first toss”, “heads appears on the second toss”, and “both tosses have the same outcome”. Show that these three events are pairwise independent, but that all three events are not independent (DeGroot and Schervish, 2012, Example 2.2.4).

Theoretical exercise 1

- It is reasonable to consider the elementary outcome set $\Omega := \{H, T\}$.
- The elementary outcome set $\Omega := \{0, 1\}$ also works.
- The power set σ -algebra $\mathcal{A} := \mathcal{P}(\{H, T\})$ contains all possible events,

Neither heads or tails occurs	$A_1 = \emptyset$
Heads occurs	$A_2 = \{H\}$
Tails occurs	$A_3 = \{T\}$
Heads or tails occurs	$A_4 = \{H, T\}$

- \mathbb{P} can be defined by specification of $\mathbb{P}(\{\omega\})$ for all $\omega \in \Omega$.
- Because the $\omega \in \mathcal{A}$ for which $\omega \in \Omega$ are pairwise disjoint, the probabilities of all events $A \in \mathcal{A}$ can be evaluated based on $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.
- A fair coin would have $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\})$, a biased coin would have $\mathbb{P}(\{H\}) \neq \mathbb{P}(\{T\})$.
- Note that $\mathbb{P}(A_4) = \mathbb{P}(\{H, T\}) = 1$.

Theoretical exercise 2

- It is reasonable to consider the elementary outcome set $\Omega := \{(d_1, d_2) | d_1 \in \mathbb{N}_6, d_2 \in \mathbb{N}_6\}$ with cardinality $|\Omega| = 36$.
- The power set σ -algebra $\mathcal{A} := \mathcal{P}(\Omega)$ contains all possible events, for example

Three pips on first throw $A_1 = \{(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6)\}$

Sum of pips equals four $A_2 = \{(1, 3), (3, 1), (2, 2)\}$

Doubles (Pasch) $A_3 = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

- \mathbb{P} can be defined by specification of $\mathbb{P}(\{\omega\})$ for all $\omega \in \Omega$ and for fair dice $\mathbb{P}(\{(d_1, d_2)\}) = 1/36$ for all $(d_1, d_2) \in \Omega$.
- Because the $\omega \in \mathcal{A}$ for which $\omega \in \Omega$ are pairwise disjoint, the probabilities of all events $A \in \mathcal{A}$ can be evaluated based on $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$. For example, the probability for the event that the sums of the pips equals for is

$$\begin{aligned}\mathbb{P}(A_2) &= \mathbb{P}(\{(1, 3), (3, 1), (2, 2)\}) \\&= \mathbb{P}(\{(1, 3)\}) \cup \{(3, 1)\} \cup \{(2, 2)\}) \\&= \mathbb{P}(\{(1, 3)\}) + \mathbb{P}(\{(3, 1)\}) + \mathbb{P}(\{(2, 2)\}) \\&= 1/36 + 1/36 + 1/36 \\&= 1/12.\end{aligned}$$

Theoretical exercise 3

- It is reasonable to consider the elementary outcome set $\Omega := \{HH, HT, TH, TT\}$.
- The power set σ -algebra $\mathcal{A} = \mathcal{P}(\{HH, HT, TH, TT\})$ contains all possible events.
- For a fair coin $\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = 1/4$
- Events and probabilities of interest:

“Heads appear on first toss”

$$A_1 = \{HH, HT\} \Rightarrow \mathbb{P}(A_1) = \mathbb{P}(\{HH\} \cup \{HT\}) = \mathbb{P}(\{HH\}) + \mathbb{P}(\{HT\}) = 1/2$$

“Heads appear on second toss”

$$A_2 = \{HH, TH\} \Rightarrow \mathbb{P}(A_2) = \mathbb{P}(\{HH\} \cup \{TH\}) = \mathbb{P}(\{HH\}) + \mathbb{P}(\{TH\}) = 1/2$$

“Both tosses have the same outcome”

$$A_3 = \{HH, TT\} \Rightarrow \mathbb{P}(A_3) = \mathbb{P}(\{HH\} \cup \{TT\}) = \mathbb{P}(\{HH\}) + \mathbb{P}(\{TT\}) = 1/2$$

Theoretical exercise 3 (cont.)

- Pairwise independence of A_1, A_2, A_3 :

$$A_1 \cap A_2 = \{HH, HT\} \cap \{HH, TH\} = \{HH\}$$

$$\Rightarrow \mathbb{P}(A_1 \cap A_2) = \mathbb{P}(\{HH\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A_1)\mathbb{P}(A_2)$$

$$A_1 \cap A_3 = \{HH, HT\} \cap \{HH, TT\} = \{HH\}$$

$$\Rightarrow \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(\{HH\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A_1)\mathbb{P}(A_3)$$

$$A_2 \cap A_3 = \{HH, TH\} \cap \{HH, TT\} = \{HH\}$$

$$\Rightarrow \mathbb{P}(A_2 \cap A_3) = \mathbb{P}(\{HH\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A_2)\mathbb{P}(A_3)$$

- Non-independence of A_1, A_2, A_3 :

$$A_1 \cap A_2 \cap A_3 = \{HH, HT\} \cap \{HH, TH\} \cap \{HH, TT\} = \{HH\}$$

$$\Rightarrow \mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\{HH\}) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$$

Exercises

Programming exercises

1. (Dice experiment 1) Consider the probability space model of tossing a fair dice. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$ be two events. Then, $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 2/3$ and $\mathbb{P}(A \cap B) = 1/3$. Since $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, the events A and B are independent. Simulate draws from the outcome space and verify that $\hat{\mathbb{P}}(A \cap B) = \hat{\mathbb{P}}(A)\hat{\mathbb{P}}(B)$, where $\hat{\mathbb{P}}(E)$ denotes the proportion of times an event E occurs in the simulation. Document your results.
2. (Dice experiment 2) Consider the probability space model of tossing a fair dice. Identify two events A and B that are not independent. Analytically, evaluate $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ and verify these values by means of simulation. Document your results.
3. (Coin experiment) Consider the probability space model of tossing a fair coin twice, i.e. a uniform probability measure on $\Omega = \{HH, HT, TH, TT\}$, where H indicates heads and T indicates tails. Simulate draws from this probability space and verify that the events “ H appears on the first toss”, “ H appears on the second toss”, and “both tosses have the same outcome” each have probability $1/2$. Document your results.

(3) Random variables

Bibliographic remarks

The presented material is standard and can be found in any probability or statistics textbook. Wasserman (2004, Sections 2.1 - 2.8) and DeGroot and Schervish (2012, Chapter 3) provide the main source for the material covered here.

Random variables

- Construction, definition, and notation
- Probability mass functions
- Probability density functions
- Cumulative distribution functions
- Exercises

Random variables

- **Construction, definition, and notation**
- Probability mass functions
- Probability density functions
- Cumulative distribution functions
- Exercises

Construction of random variables and distributions

- Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let $X : \Omega \rightarrow \mathcal{X}$ be a function.
- Let \mathcal{S} be a σ -algebra on \mathcal{X} .
- For every $S \in \mathcal{S}$ let the *preimage* of S be

$$X^{-1}(S) := \{\omega \in \Omega | X(\omega) \in S\}. \quad (11)$$

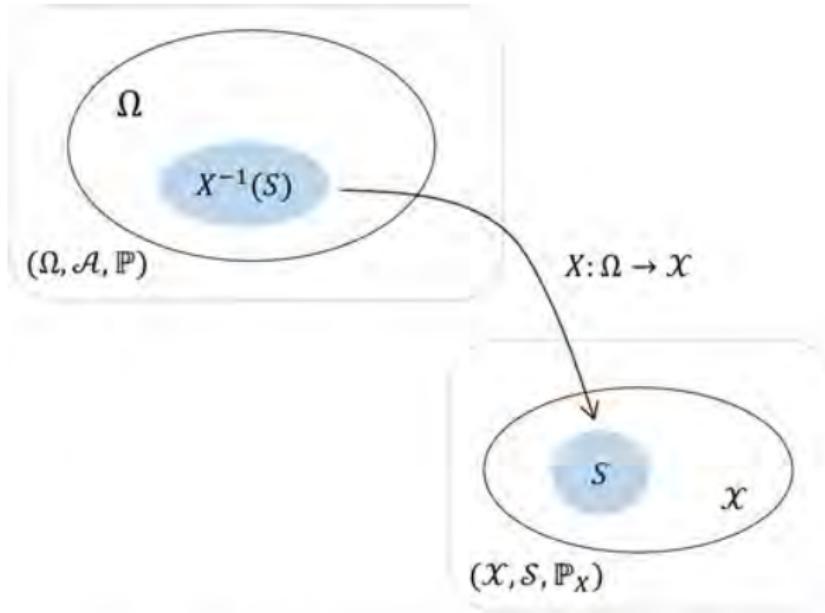
- If $X^{-1}(S) \in \mathcal{A}$ for all $S \in \mathcal{S}$, then X is called *measurable*.
- Let $X : \Omega \rightarrow \mathcal{X}$ be measurable. All $S \in \mathcal{S}$ get allocated the probability

$$\mathbb{P}_X : \mathcal{S} \rightarrow [0, 1], S \mapsto \mathbb{P}_X(S) := \mathbb{P}(X^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in S\}) \quad (12)$$

- X is called a *random variable* and \mathbb{P}_X is called the *distribution* of X .
- $(\mathcal{X}, \mathcal{S}, \mathbb{P}_X)$ is a probability space.
- With $\mathcal{X} = \mathbb{R}$ and $\mathcal{S} = \mathcal{B}$ the probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ takes center stage.

Construction, definition, and notation

Construction of random variables and distributions



$$\mathbb{P}(X^{-1}(S)) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in S\}) =: \mathbb{P}_X(S)$$

Definition (Random variable)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote a probability space and let $(\mathcal{X}, \mathcal{S})$ denote a measurable space, i.e., \mathcal{X} is a set and \mathcal{S} is a σ -algebra on \mathcal{X} . A *random variable* is a function

$$X : \Omega \rightarrow \mathcal{X}, \omega \mapsto X(\omega), \quad (13)$$

with the *measurability property*

$$\{\omega \in \Omega | X(\omega) \in S\} \in \mathcal{A} \text{ for all } S \in \mathcal{S}. \quad (14)$$

Remarks

- Random variables are neither “random” nor “variables”.
- Intuitively, $\omega \in \Omega$ gets randomly selected according to \mathbb{P} and $X(\omega)$ realized.
- The distributions (probability measures) of random variables are central.
- We call \mathcal{X} the *outcome set of the random variable* X .

Construction, definition, and notation

Let $(\Omega, \mathcal{A}, \mathbb{P})$ and $(\mathcal{X}, \mathcal{S}, \mathbb{P}_X)$ denote probability spaces relating to $X : \Omega \rightarrow \mathcal{X}$. The following notations for events are conventional:

$$\{X \in S\} := \{\omega \in \Omega | X(\omega) \in S\}, S \subset \mathcal{X},$$

$$\{X = x\} := \{\omega \in \Omega | X(\omega) = x\}, x \in \mathcal{X},$$

$$\{X \leq x\} := \{\omega \in \Omega | X(\omega) \leq x\}, x \in \mathcal{X},$$

$$\{X < x\} := \{\omega \in \Omega | X(\omega) < x\}, x \in \mathcal{X}.$$

These conventions entail the following conventions for distributions:

$$\mathbb{P}_X(X \in S) = \mathbb{P}(\{X \in S\}) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \in S\}), S \subset \mathcal{X}$$

$$\mathbb{P}_X(X \leq x) = \mathbb{P}(\{X \leq x\}) = \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\}), x \in \mathcal{X}.$$

Often, the random variable subscript for distribution symbols is omitted:

$$\mathbb{P}(X \in S) = \mathbb{P}_X(X \in S), S \subset \mathcal{X},$$

$$\mathbb{P}(X \leq x) = \mathbb{P}_X(X \leq x), x \in \mathcal{X}.$$

Construction, definition, and notation

Interlinking probability measures, random variables, and PMF/PDFs

- The distributions of random variables are central in probability and statistics.
- Distributions of random variables are usually defined by PMFs and PDFs.
- But what about probability spaces and probability measures then?
- Probability measures \Leftrightarrow Random variable distributions

\mathbb{P} is a probability measure on the Borel σ -algebra.

$\Leftrightarrow (\Omega, \mathcal{F}, \mathbb{P}) := (\mathbb{R}, \mathcal{B}, \mathbb{P})$ is a probability space

$\Leftrightarrow X : \Omega \rightarrow \mathbb{R}, \omega \mapsto X(\omega) := \omega$ is a random variable

$\Leftrightarrow \mathbb{P}_X = \mathbb{P}.$

$\Leftrightarrow \mathbb{P}$ is the distribution of a random variable

- The focus will be on distributions in the remainder of the course.

Random variables

- Construction, definition, and notation
- **Probability mass functions**
- Probability density functions
- Cumulative distribution functions
- Exercises

Definition (Discrete random variable, probability mass function)

A random variable X is discrete, if it takes on countably many values in $\mathcal{X} := \{x_1, x_2, \dots\}$. The probability mass function of X is defined as

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \mathbb{P}(X = x). \quad (15)$$

Remarks

- A set is countable, if it is finite or bijectively related to \mathbb{N} .
- A PMF is non-negative: $p(x) \geq 0$ for all $x \in \mathcal{X}$.
- A PMF is normalized: $\sum_{x \in \mathcal{X}} p(x) = 1$.

Probability mass functions

Example (Bernoulli random variable)

Let X be a random variable with outcome set $\mathcal{X} = \{0, 1\}$ and probability mass function

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \mu^x(1 - \mu)^{1-x} \text{ for } \mu \in [0, 1]. \quad (16)$$

Then X is said to be distributed according to a *Bernoulli distribution* with parameter $\mu \in [0, 1]$, for which we write $X \sim \text{Bern}(\mu)$. We denote the probability mass function of a Bernoulli random variable by

$$\text{Bern}(x; \mu) := \mu^x(1 - \mu)^{1-x}. \quad (17)$$

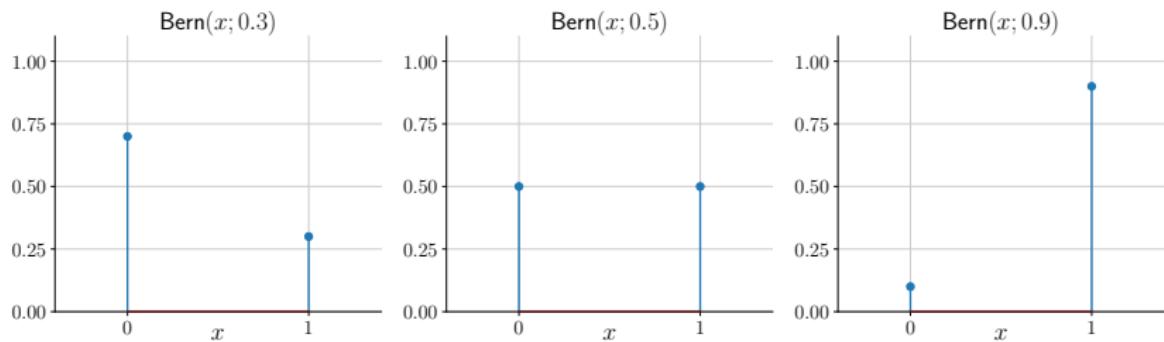
Remarks

- A Bernoulli random variable can be used to model a single biased coin flip with outcomes “failure” 0 and “success” 1.
- μ is the probability for X to take the value 1,

$$\mathbb{P}(X = 1) = \mu^1(1 - \mu)^{1-1} = \mu. \quad (18)$$

Probability mass functions

Example (Bernoulli distributions)



Probability mass functions

Example (Binomial random variable)

Let X be a random variable with outcome set $\mathcal{X} := \mathbb{N}_n^0$ and probability mass function

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \binom{n}{x} \mu^x (1 - \mu)^{n-x} \text{ for } \mu \in [0, 1]. \quad (19)$$

Then X is said to be distributed according to a *Binomial distribution* with parameters $\mu \in [0, 1]$ and $n \in \mathbb{N}$, for which we write $X \sim \text{Bin}(\mu, n)$. We denote the probability mass function of a Binomial random variable by

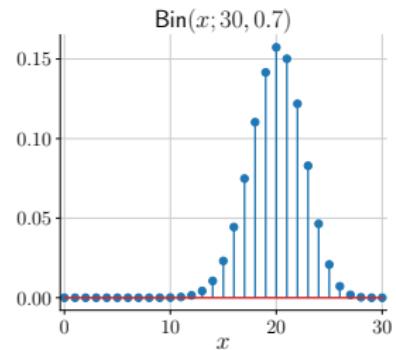
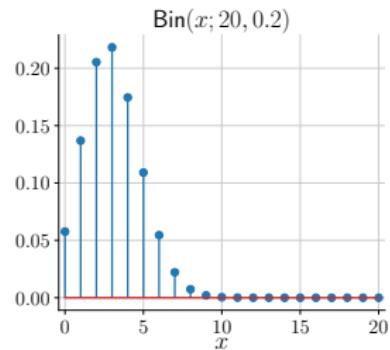
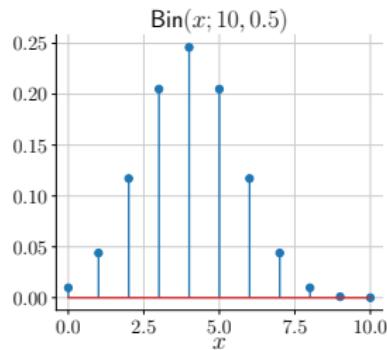
$$\text{Bin}(x; \mu, n) := \binom{n}{x} \mu^x (1 - \mu)^{n-x} \quad (20)$$

Remark

- $\text{Bin}(x; \mu, 1) = \text{Bern}(x; \mu)$.

Probability mass functions

Example (Binomial distributions)



Example (Discrete uniform random variable)

Let X be a discrete random variable with a finite outcome set \mathcal{X} and probability mass function

$$p : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) := \frac{1}{|\mathcal{X}|}. \quad (21)$$

Then X is said to be distributed according to a *discrete uniform distribution*, for which we write $X \sim U(|\mathcal{X}|)$. We abbreviate the PMF of a discrete uniform random variable by

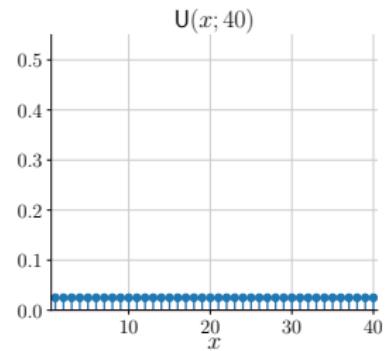
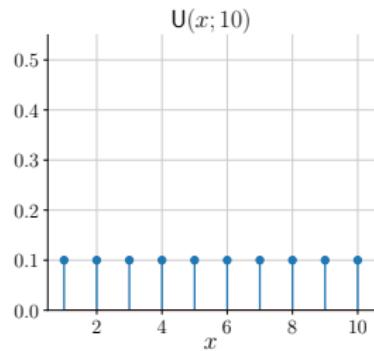
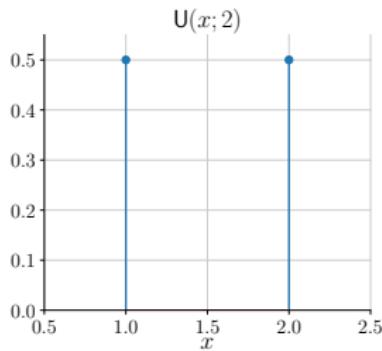
$$U(x; |\mathcal{X}|) := \frac{1}{|\mathcal{X}|}. \quad (22)$$

Remark

- $\text{Bern}(x; 0.5) = U(x; |\mathcal{X}|)$ for $\mathcal{X} = \{0, 1\}$.
- E.g., $\text{Bin}(x; 0.5) = U(x; |\mathcal{X}|)$ for $\mathcal{X} = \{0, 1\}$.

Probability mass functions

Example (Discrete uniform distributions)



Random variables

- Construction, definition, and notation
- Probability mass functions
- **Probability density functions**
- Cumulative distribution functions
- Exercises

Definition (Continuous random variable, probability density function)

A random variable X is continuous, if there exists a function

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) \quad (23)$$

such that

- $p(x) \geq 0$ for all $x \in \mathbb{R}$,
- $\int_{-\infty}^{\infty} p(x) dx = 1$,
- $\mathbb{P}(a \leq X \leq b) = \int_a^b p(x) dx$ for all $a, b \in \mathbb{R}$, $a \leq b$.

Remarks

- PDFs can take on values larger than 1 and $\mathbb{P}(X = a) = \int_a^a p(x) dx = 0$.
- Probabilities are obtained from PDFs by integration.
- (Probability) mass = (probability) density \times (set) volume.

Example (Gaussian random variable, standard normal variable)

Let X be a random variable with outcome set \mathbb{R} and probability density function

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (24)$$

Then X is said to be distributed according to a *Gaussian distribution* with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, for which we write $X \sim N(\mu, \sigma^2)$. We abbreviate the PDF of a Gaussian random variable by

$$N(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \quad (25)$$

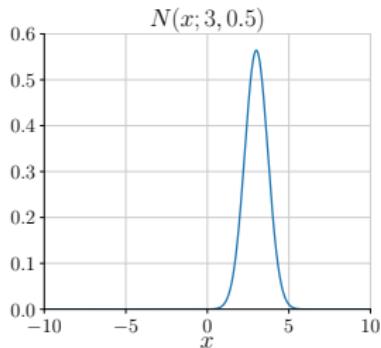
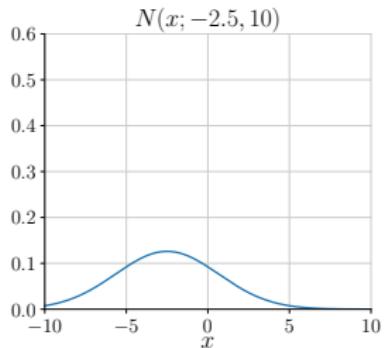
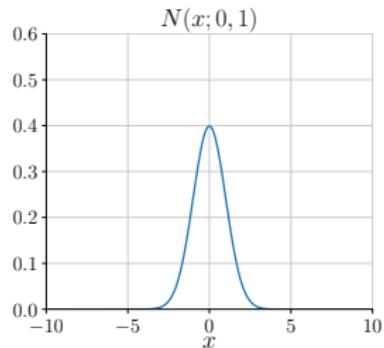
A Gaussian random variable with $\mu = 0$ and $\sigma^2 = 1$ is said to be distributed according to a *standard normal distribution* and is often referred to as a *Z variable*.

Remarks

- The parameter μ specifies the location of highest probability density.
- The parameter σ^2 specifies the width of the distribution.

Probability density functions

Example (Gaussian distributions)



Example (Gamma random variable)

Let X be a random variable with outcome set $\mathcal{X} := \mathbb{R}_{>0}$ and probability density function

$$p : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right). \quad (26)$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a *Gamma distribution* with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$, for which we write $X \sim G(\alpha, \beta)$. We denote the PDF of a Gamma random variable by

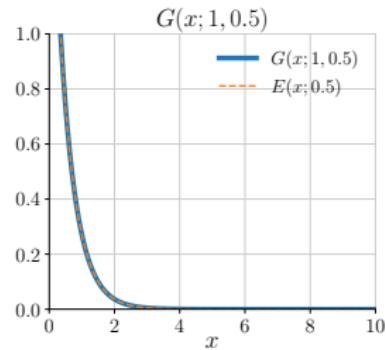
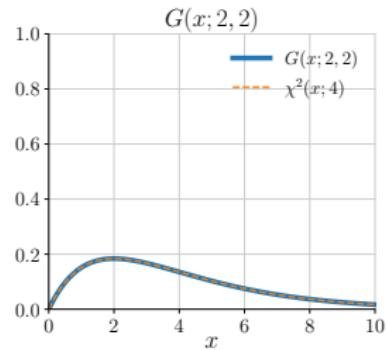
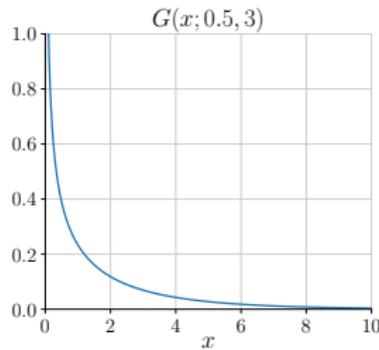
$$G(x; \alpha, \beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right). \quad (27)$$

Remarks

- $G\left(\frac{n}{2}, 2\right), n \in \mathbb{N}$ is the χ^2 distribution $\chi^2(n)$ with n degrees of freedom.
- $G(1, \beta), \beta > 0$ is the exponential distribution $E\left(\frac{1}{\beta}\right)$ with rate parameter $\frac{1}{\beta}$.

Probability density functions

Example (Gamma distributions)



Example (Beta random variable)

Let X be a random variable with outcome set $\mathcal{X} := [0, 1]$ and probability density function

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \text{ for } \alpha, \beta \in \mathbb{R}_{>0}, \quad (28)$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a *Beta distribution* with parameters α, β , for which we write $X \sim \text{Beta}(\alpha, \beta)$. We denote the probability density function of a Beta random variable by

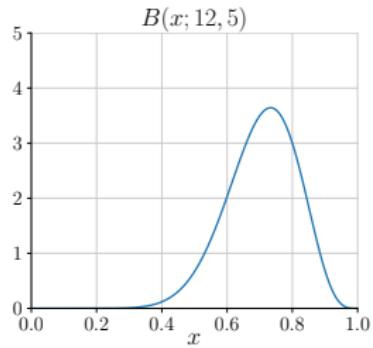
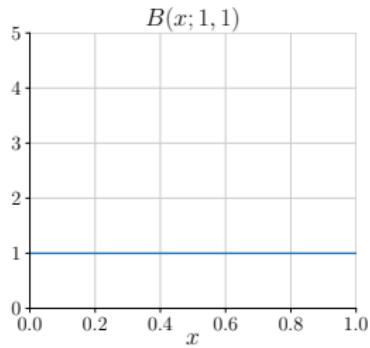
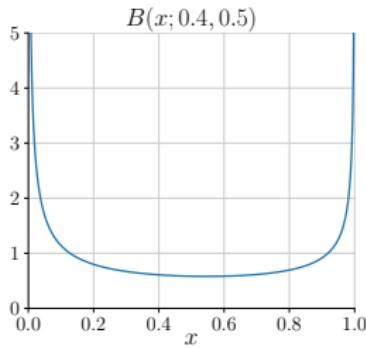
$$\text{Beta}(x; \alpha, \beta) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}. \quad (29)$$

Remarks

- A Beta random variable can be used to model the distribution of a probability.
- For $\alpha < 1, \beta < 1$ the outcome set is $\mathcal{X} :=]0, 1[$.

Probability density functions

Example (Beta distributions)



Example (Uniform random variables)

Let X be a continuous random variable with probability density function

$$p : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, x \mapsto p(x) := \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}. \quad (30)$$

Then X is said to be distributed according to a *continuous uniform distribution* with parameters a and b , for which we write $X \sim U(a, b)$. We abbreviate the PDF of a continuous uniform random variable by

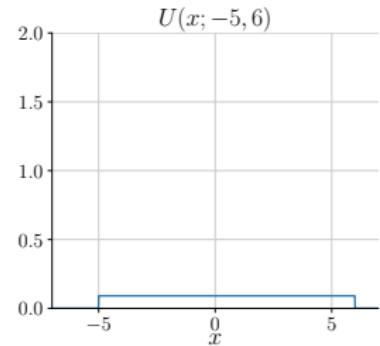
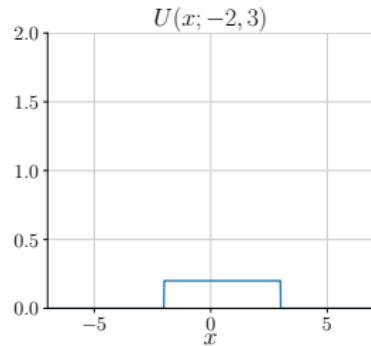
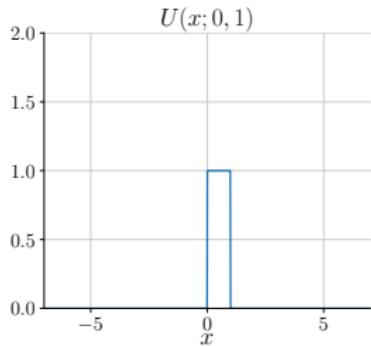
$$U(x; a, b) := \frac{1}{b - a}. \quad (31)$$

Remarks

- $\text{Beta}(x; 1, 1) = U(x; 0, 1)$.

Probability density functions

Example (Continuous uniform distributions)



Random variables

- Construction, definition, and notation
- Probability mass functions
- Probability density functions
- **Cumulative distribution functions**
- Exercises

Definition (Cumulative distribution function)

The cumulative distribution function (CDF) of a random variable X is defined as

$$P : \mathbb{R} \rightarrow [0, 1], x \mapsto P(x) := \mathbb{P}(X \leq x). \quad (32)$$

Remarks

- CDFs are defined for both discrete and continuous random variables.
- $P(x)$ is defined for every $x \in \mathbb{R}$, even if $x \notin \mathcal{X}$.
- CDFs allow for evaluating interval probabilities.

Theorem (Cumulative distribution function properties)

Let X denote a random variable and let P denotes its cumulative distribution function. Then P has the following properties

- (1) P is *nondecreasing*, i.e., if $x_1 < x_2$, then $P(x_1) \leq P(x_2)$.
- (2) $\lim_{x \rightarrow -\infty} P(x) = 0$ and $\lim_{x \rightarrow \infty} P(x) = 1$.
- (3) P is *right-continuous*, i.e., $P(x) = P(x^+) = \lim_{y \rightarrow x, y > x} P(y)$ for all $x \in \mathbb{R}$

Remarks

- The listed properties can also be used to define the very notion of CDFs.
- (3) \Leftrightarrow No CDF jumps occur when limit points are approached from the right.

Cumulative distribution functions

Proof

(1) We first recall that for events $A \subset B$ it holds that $\mathbb{P}(A) \leq \mathbb{P}(B)$. We next note that for $x_1 < x_2$,

$$\{X \leq x_1\} = \{\omega \in \Omega | X(\omega) \leq x_1\} \subset \{\omega \in \Omega | X(\omega) \leq x_2\} = \{X \leq x_2\} \quad (33)$$

Thus

$$\mathbb{P}(\{X \leq x_1\}) \leq \mathbb{P}\{X \leq x_2\} \Rightarrow P(x_1) \leq P(x_2). \quad (34)$$

(2) Omitted.

(3) Define

$$P(x^+) = \lim_{y \rightarrow x, y > x} P(y). \quad (35)$$

Let $y_1 > y_2 > \dots$ such that $\lim_{n \rightarrow \infty} y_n = x$. Then

$$\{X \leq x\} = \cap_{n=1}^{\infty} \{X \leq y_n\} \quad (36)$$

We thus have

$$P(x) = \mathbb{P}(\{X \leq x\}) = \mathbb{P}(\cap_{n=1}^{\infty} \{X \leq y_n\}) = \lim_{n \rightarrow \infty} \mathbb{P}(\{X \leq y_n\}) = P(x^+), \quad (37)$$

where we leave the third equality unjustified.

Cumulative distribution functions

Cumulative distribution functions of discrete random variables

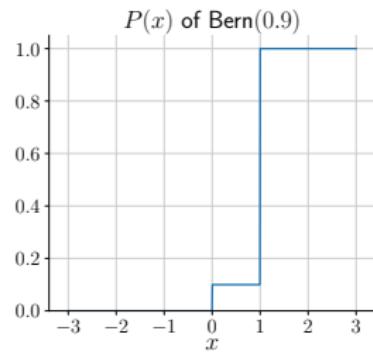
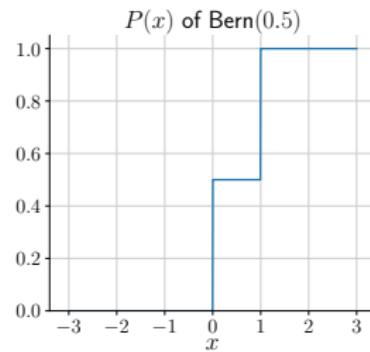
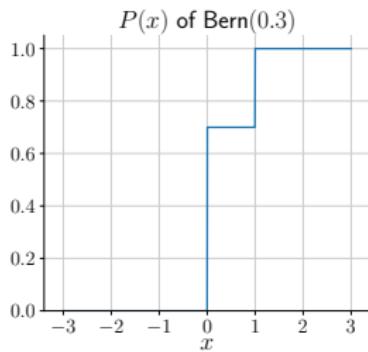
- If $a < b$ and $\mathbb{P}(a < X < b) = 0$, then P_X is constant and horizontal on $]a, b[$.
- At every point x with $\mathbb{P}(X = x) > 0$, the CDF jumps by the amount $\mathbb{P}(X = x)$.
- \Leftrightarrow At every point x with $p(x) > 0$, the CDF jumps by the amount $p(x) > 0$.
- In general, a discrete random variable with outcome space \mathbb{N}_0 has CDF

$$P : \mathbb{R} \rightarrow [0, 1], x \mapsto P(x) := \sum_{k=0}^{\lfloor x \rfloor} \mathbb{P}(X = k) \quad (38)$$

where $\lfloor x \rfloor$ is the floor function.

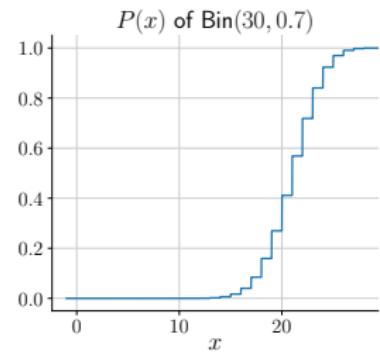
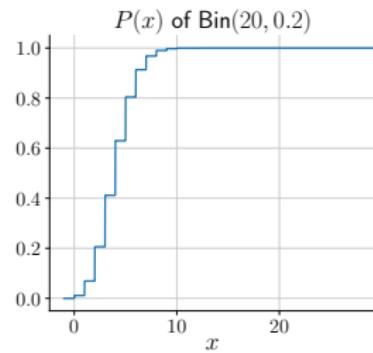
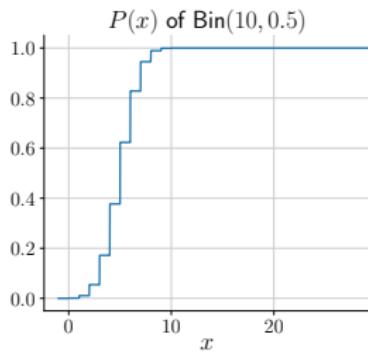
Cumulative distribution functions

Example (Bernoulli distributions)



Cumulative distribution functions

Example (Binomial distributions)



Theorem (CDFs of continuous random variables)

Let X be a continuous random variable with PDF p and CDF P . Then

$$P(x) = \int_{-\infty}^x p(t) dt \text{ and } p(x) = \frac{d}{dx} P(x). \quad (39)$$

Proof

We first note that because $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$, the CDF of X does not have jumps, i.e., it is continuous over the real line. By definition of the CDF and PDF, the anti-derivative nature of the CDF with respect to the PDF follows immediately. The derivative nature of p with respect to P follows by the fundamental theorem of calculus.

□

Remarks

- The CDF is the anti-derivative of the PDF, the PDF is the derivative of the CDF.
- A more general variant of the theorem is afforded by the Radon-Nikodym theorem.

Example (Gaussian distributions)

Let X have a Gaussian distribution $N(\mu, \sigma^2)$.

- The PDF of X is

$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right).$$

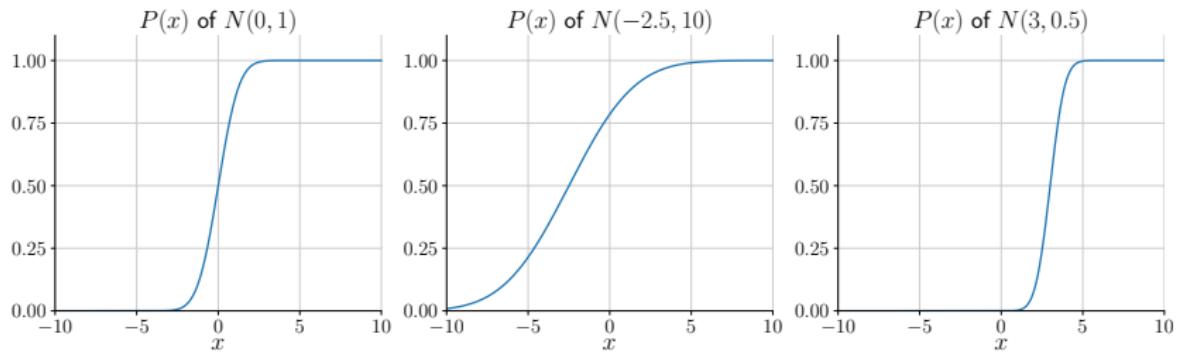
- The CDF of X is

$$P : \mathbb{R} \rightarrow]0, 1[, x \mapsto P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\left(-\frac{1}{2\sigma^2}(\xi - \mu)^2\right) d\xi.$$

- An analytical form of the Gaussian CDF does not exist, numerical approximations do.
- For example, for $\mu = 1, \sigma^2 = 1$, $p(2) = 0.24$ and $P(2) = 0.84$.
- The PDF and CDF of $N(0, 1)$ are often denoted by ϕ and Φ , respectively.

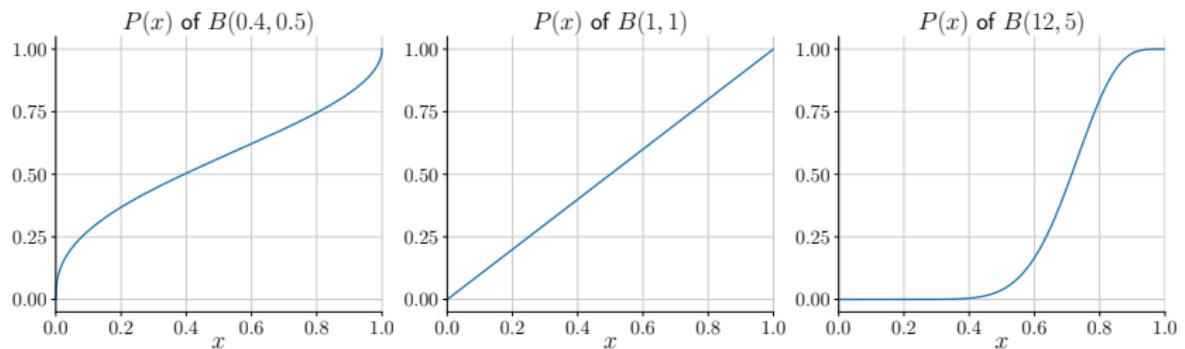
Cumulative distribution functions

Example (Gaussian distributions (cont.))



Cumulative distribution functions

Example (Beta distributions)



Definition (Inverse cumulative distribution function)

Let X be a continuous random variable with CDF P . Then the function

$$P^{-1} :]0, 1[\rightarrow \mathbb{R}, q \mapsto P^{-1}(q) := \{x \in \mathbb{R} | P(x) = q\} \quad (40)$$

is called inverse cumulative distribution function of X .

Remarks

- P^{-1} is the inverse function of P , i.e., $P_X^{-1}(P(x)) = x$.
- Notably, $P(x) = q \Leftrightarrow \mathbb{P}(X \leq x) = q$.
- For $q \in]0, 1[, P^{-1}(q)$ is thus that value x of X , such that $\mathbb{P}(X \leq x) = q$.
- For example, if $Z \sim N(0, 1)$ with CDF Φ , then $\Phi^{-1}(0.975) = 1.960$.

Example (Gaussian distributions)

Let X have a Gaussian distribution $N(\mu, \sigma^2)$

- The CDF of X is

$$P : \mathbb{R} \rightarrow]0, 1[, x \mapsto \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\left(-\frac{1}{2\sigma^2}(\xi - \mu)^2\right) d\xi \quad (41)$$

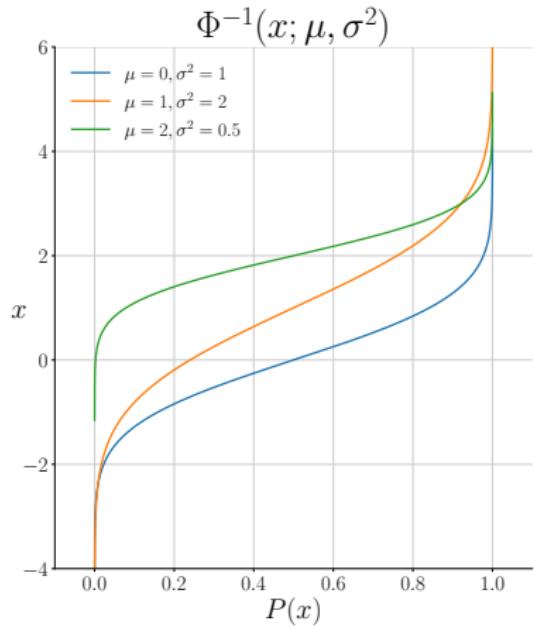
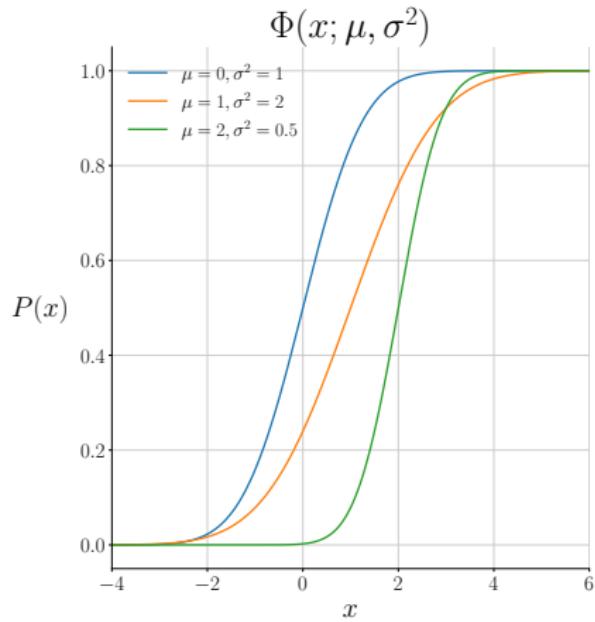
- The inverse CDF of X is

$$P^{-1} :]0, 1[\rightarrow \mathbb{R}, q \mapsto P^{-1}(q) = \{x \in \mathbb{R} | P(x) = q\}. \quad (42)$$

- For example, for $\mu = 1, \sigma^2 = 1$, $P(2) = 0.84$, and $P^{-1}(0.84) = 2$.
- The CDF and inverse CDF of $N(0, 1)$ are often denoted by Φ and Φ^{-1} , respectively.
- Typical examples for CDF and inverse CDF values of $N(0, 1)$ are
 - $\Phi(1.645) = 0.950$, $\Phi^{-1}(0.950) = \Phi^{-1}(1 - 0.050) = 1.640$.
 - $\Phi(1.960) = 0.975$, $\Phi^{-1}(0.975) = \Phi^{-1}(1 - \frac{0.050}{2})$.

Cumulative distribution functions

Example (Gaussian distributions)



Definition (Quantile function)

Let X be a random variable with CDF P . Let $q \in]0, 1[$ and define $P^{-1}(q)$ to be the smallest value x such that $P(x) \geq q$, formally

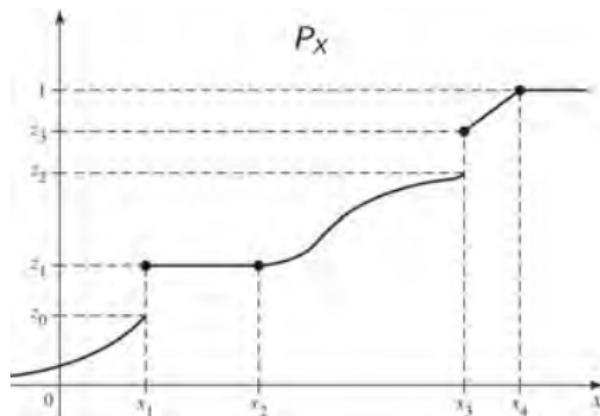
$$P_X^{-1} :]0, 1[\rightarrow \mathbb{R}, q \mapsto P^{-1}(q) := \inf\{x \in \mathbb{R} | P(x) \geq q\}. \quad (43)$$

Remarks

- The quantile function is a generalized inverse cumulative distribution function.
- The quantile function exists for continuous and discrete distributions.
- Right-continuity ensures that the smallest x with $P(x) \geq q$ exists for all $0 < q < 1$.
- $P^{-1}(0)$ does not exist, because $\lim_{x \rightarrow -\infty} P(x) = 0$.
- $P^{-1}(1)$ may or may not exist.

Cumulative distribution functions

Example (Quantile function, DeGroot and Shervish, 2012, Figure 3.6)



- For $q = 1$, the smallest x with $P(x) \geq 1$ is x_4 . Thus $P^{-1}(1) = x_4$.
- $P^{-1}(z_3) = x_3$ and $P^{-1}(z_2) = x_2$.
- For $z_0 \leq q \leq z_1$, the smallest x with $P(x) \geq q$ is x_1 .
 - For every $x < x_1$, $P(x) < z_0 \leq q$ and $P(x_1) = z_1$.
 - $P(x) = z_1$ for all x between x_1 and x_2 , but x_1 is the smallest.

Random variables

- Construction, definition, and notation
- Probability mass functions
- Probability density functions
- Cumulative distribution functions
- **Exercises**

Study Questions

1. Write down the definition of a random variable.
2. Write down the definition of a discrete random variable and a probability mass function (PMF).
3. Write down the definition of a continuous random variable and a probability density function (PDF).
4. Write down the definition of the cumulative distribution function (CDF) of a random variable.
5. Express the value $P(x)$ of the CDF of a discrete random variable X in terms of its PMF.
6. Express the value $P(x)$ of the CDF of a continuous random variable X in terms of its PDF.
7. Express the value $p(x)$ of the PDF of a continuous random variable X in terms of its CDF.
8. Write down the PDF and CDF of a Gaussian random variable
9. State three properties of CDFs.
10. Write down the definitions of the inverse CDF and the quantile functions.

Theoretical Exercises

1. Develop a probability space model of throwing two dice and the probability space model that results from defining a random variable that evaluates the sum of the pips (e.g., Moeschlin, 2000b, Beispiel 3.1.1).
2. Let X denote a random variable with outcome space \mathcal{X} and let P denote its cumulative distribution function. Prove the following properties of P (DeGroot and Schervish, 2012, Theorems 3.3.1 and 3.3.2).
 - $\mathbb{P}(X > x) = 1 - P(x)$ for all $x \in \mathcal{X}$.
 - $\mathbb{P}(x_1 < X \leq x_2) = P(x_2) - P(x_1)$ for all $x_1, x_2 \in \mathcal{X}$ with $x_1 < x_2$.
3. Introduce the left-continuous, right-continuous, and normalized version of the cumulative distribution function (e.g. Brunner et al. (2018, Chapter 2.1)).

Theoretical Exercise 1

- The probability space model $(\Omega, \mathcal{A}, \mathbb{P})$ of throwing two dice is sensibly defined as
 - $\Omega := \{(d_1, d_2) | d_1 \in \mathbb{N}_6, d_2 \in \mathbb{N}_6\}$ and $\mathcal{A} := \mathcal{P}(\Omega)$.
 - $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ with $\mathbb{P}(\{(d_1, d_2)\}) = 1/36$ for all $(d_1, d_2) \in \Omega$ and σ -additivity of \mathbb{P} .
- Let $\mathcal{X} := \{2, 3, \dots, 12\}$. Then the random variable modelling the evaluation of the sum of pips is sensibly defined as

$$X : \Omega \mapsto \mathcal{X}, (d_1, d_2) \mapsto X((d_1, d_2)) := d_1 + d_2. \quad (44)$$

- $\mathcal{S} := \mathcal{P}(\mathcal{X})$ is a sensible σ -algebra on \mathcal{X} .
- We tabulate the distribution \mathbb{P}_X of X for all elementary outcomes $\{x\} \in \mathcal{S}$ below. With the σ -additivity of \mathbb{P}_X , we have then established the probability space model $(\mathcal{X}, \mathcal{S}, \mathbb{P}_X)$.

Exercises

Theoretical Exercise 1 (cont.)

$\mathbb{P}_X(\{2\})$	$= \mathbb{P}(X^{-1}(\{2\}))$	$= \mathbb{P}(\{(1, 1)\})$	$= \frac{1}{36}$
$\mathbb{P}_X(\{3\})$	$= \mathbb{P}(X^{-1}(\{3\}))$	$= \mathbb{P}(\{(1, 2), (2, 1)\})$	$= \frac{2}{36}$
$\mathbb{P}_X(\{4\})$	$= \mathbb{P}(X^{-1}(\{4\}))$	$= \mathbb{P}(\{(1, 3), (3, 1), (2, 2)\})$	$= \frac{3}{36}$
$\mathbb{P}_X(\{5\})$	$= \mathbb{P}(X^{-1}(\{5\}))$	$= \mathbb{P}(\{(1, 4), (4, 1), (2, 3), (3, 2)\})$	$= \frac{4}{36}$
$\mathbb{P}_X(\{6\})$	$= \mathbb{P}(X^{-1}(\{6\}))$	$= \mathbb{P}(\{(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)\})$	$= \frac{5}{36}$
$\mathbb{P}_X(\{7\})$	$= \mathbb{P}(X^{-1}(\{7\}))$	$= \mathbb{P}(\{(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)\})$	$= \frac{6}{36}$
$\mathbb{P}_X(\{8\})$	$= \mathbb{P}(X^{-1}(\{8\}))$	$= \mathbb{P}(\{(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)\})$	$= \frac{5}{36}$
$\mathbb{P}_X(\{9\})$	$= \mathbb{P}(X^{-1}(\{9\}))$	$= \mathbb{P}(\{(3, 6), (6, 3), (4, 5), (5, 4)\})$	$= \frac{4}{36}$
$\mathbb{P}_X(\{10\})$	$= \mathbb{P}(X^{-1}(\{10\}))$	$= \mathbb{P}(\{(4, 6), (6, 4), (5, 5)\})$	$= \frac{3}{36}$
$\mathbb{P}_X(\{11\})$	$= \mathbb{P}(X^{-1}(\{11\}))$	$= \mathbb{P}(\{(5, 6), (6, 5)\})$	$= \frac{2}{36}$
$\mathbb{P}_X(\{12\})$	$= \mathbb{P}(X^{-1}(\{12\}))$	$= \mathbb{P}(\{(6, 6)\})$	$= \frac{1}{36}$

Exercises

Theoretical Exercise 2

Theorem (Exceedance probabilities)

Let X denote a random variable with outcome space \mathcal{X} and let P denote its cumulative distribution function. Then

$$\mathbb{P}(X > x) = 1 - P(x) \text{ for all } x \in \mathcal{X}. \quad (45)$$

Proof

The events $\{X > x\}$ and $\{X \leq x\}$ are disjoint and

$$\Omega = \{\omega \in \Omega | X(\omega) > x\} \cup \{\omega \in \Omega | X(\omega) \leq x\} = \{X > x\} \cup \{X \leq x\}. \quad (46)$$

With the σ -additivity of \mathbb{P} , we thus have

$$\begin{aligned} \mathbb{P}(\Omega) &= 1 \\ \Leftrightarrow \mathbb{P}(\{X > x\} \cup \{X \leq x\}) &= 1 \\ \Leftrightarrow \mathbb{P}(\{X > x\}) + \mathbb{P}(\{X \leq x\}) &= 1 \\ \Leftrightarrow \mathbb{P}(\{X > x\}) &= 1 - \mathbb{P}(\{X \leq x\}) \\ \Leftrightarrow \mathbb{P}(\{X > x\}) &= 1 - P(x). \end{aligned} \quad (47)$$

□

Theoretical Exercise 2

Theorem (Interval probabilities)

Let X denote a random variable with outcome space \mathcal{X} and let P denote its cumulative distribution function. Then

$$\mathbb{P}(x_1 < X \leq x_2) = P(x_2) - P(x_1) \text{ for all } x_1, x_2 \in \mathcal{X} \text{ with } x_1 < x_2. \quad (48)$$

Proof

Consider the events $\{X \leq x_1\}$, $\{x_1 < X \leq x_2\}$, and $\{X \leq x_2\}$ with

$$\{X \leq x_1\} \cap \{x_1 < X \leq x_2\} = \emptyset \text{ and } \{X \leq x_1\} \cup \{x_1 < X \leq x_2\} = \{X \leq x_2\}. \quad (49)$$

With the σ -additivity of \mathbb{P} , we have

$$\begin{aligned} \mathbb{P}(\{X \leq x_1\} \cup \{x_1 < X \leq x_2\}) &= \mathbb{P}(\{X \leq x_2\}) \\ \Leftrightarrow \mathbb{P}(\{X \leq x_1\}) + \mathbb{P}(\{x_1 < X \leq x_2\}) &= \mathbb{P}(\{X \leq x_2\}) \\ \Leftrightarrow \mathbb{P}(\{x_1 < X \leq x_2\}) &= \mathbb{P}(\{X \leq x_2\}) - \mathbb{P}(\{X \leq x_1\}) \\ \Leftrightarrow \mathbb{P}(\{x_1 < X \leq x_2\}) &= P(x_2) - P(x_1). \end{aligned} \quad (50)$$

□

Theoretical Exercise 3

- The most popular CDF is the *right-continuous version of the CDF*,

$$F^+ : \mathbb{R} \rightarrow [0, 1], x \mapsto F^+(x) := \mathbb{P}(X \leq x). \quad (51)$$

- Equivalently, one can define the *left-continuous version of the CDF*

$$F^- : \mathbb{R} \rightarrow [0, 1], x \mapsto F^-(x) := \mathbb{P}(X < x). \quad (52)$$

- The proof of the left-continuity of F^- is omitted here.
- For continuous random variables, $F^-(x) = F^+(x)$ for all $x \in \mathbb{R}$.
- For discrete random variables at x_0 with $\mathbb{P}(X = x_0) > 0$, it holds that

$$F^+(x_0) > F^-(x_0) \text{ and } \mathbb{P}(X = x_0) = F^+(x_0) - F^-(x_0). \quad (53)$$

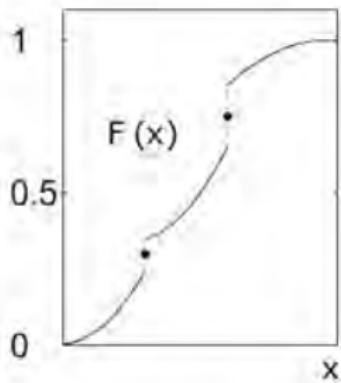
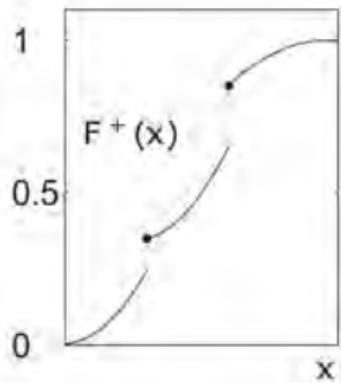
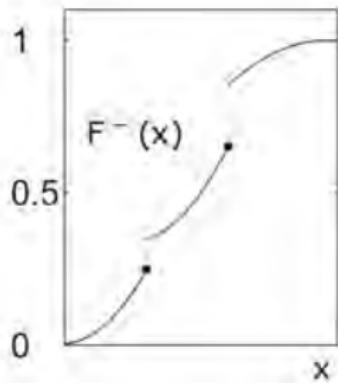
- The *normalized version of the CDF* is defined as

$$F : \mathbb{R} \rightarrow [0, 1], x \mapsto F(x) := \frac{1}{2} (F^+(x) + F^-(x)). \quad (54)$$

- F is useful for dissolving ties in rank-order statistics (Brunner et al., 2018).

Exercises

CDF versions (Brunner et al., 2018, Figure 2.1)

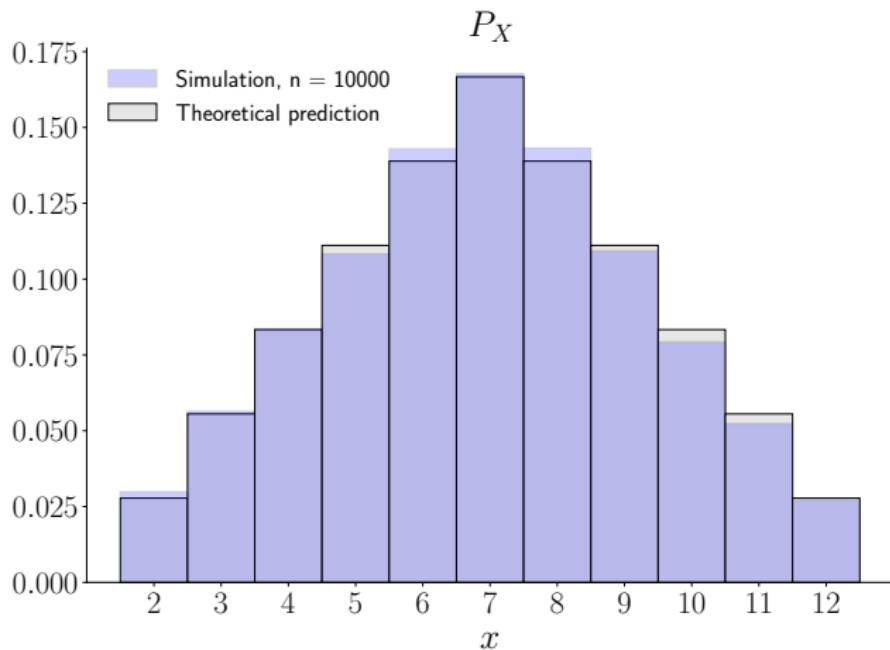


Programming Exercises

1. Simulate the probability space model of throwing two dice and the random variable corresponding to the sum of the pips. Visualize a normalized histogram of simulated outcomes of this random variable and compare it to the theoretical prediction.
2. Visualize the PMF of a Bernoulli random variable and a normalized histogram of many samples of a Bernoulli random variable with identical parameter setting on top of each other.
3. Visualize the PDF of a Gaussian random variable and a normalized histogram of many samples of a Gaussian random variable with identical parameter settings on top of each other.

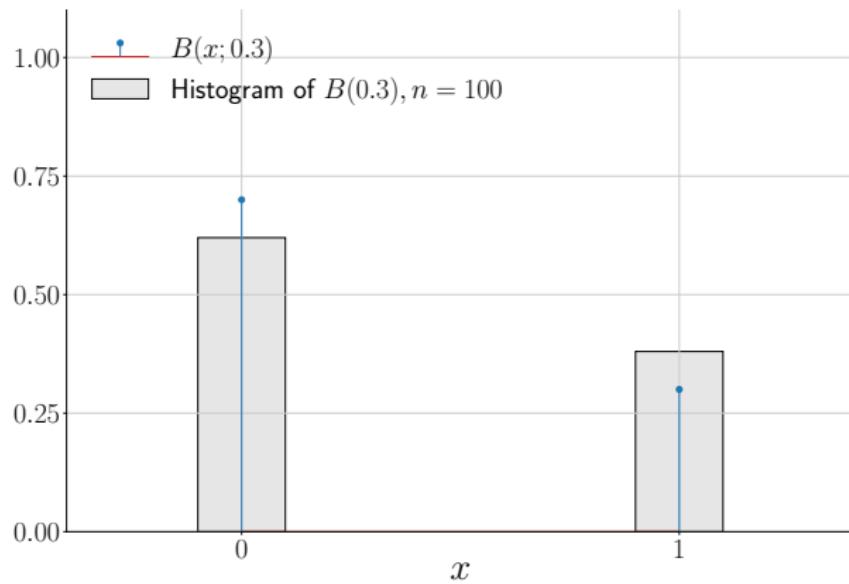
Exercises

Programming Exercise 1



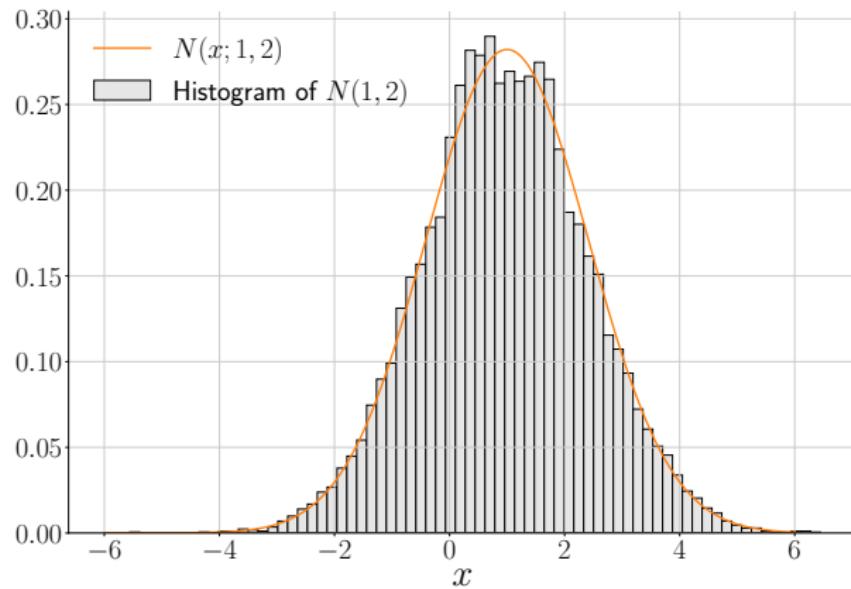
Exercises

Programming Exercise 2



Exercises

Programming Exercise 3



(4) Joint distributions

Bibliographic remarks

Multivariate distributions are indispensable for statistical inference and probability-oriented machine learning. The presented material is standard and can be found in any introductory textbook to statistical inference. DeGroot and Schervish (2012, Sections 3.4 - 3.7) and Wasserman (2004, Sections 2.9 - 2.10) provide the main sources for the material covered here.

Joint distributions

- Joint distributions
- Marginal distributions
- Independent random variables
- Conditional distributions
- Multivariate distributions
- Exercises

Joint distributions

- **Joint distributions**
- Marginal distributions
- Independent random variables
- Conditional distributions
- Multivariate distributions
- Exercises

Definition (Joint distribution)

Let X and Y denote two random variables with outcome spaces \mathcal{X} and \mathcal{Y} , respectively. The *joint distribution* or *bivariate distribution* of X and Y is the probability measure

$$\mathbb{P}_{X,Y}((X, Y) \in A) \text{ for all } A \subset \mathcal{X} \times \mathcal{Y} \quad (55)$$

such that $\{(X, Y) \in A\}$ is an event.

Remarks

- We focus on the specification of joint distributions using PMFs and PDFs.
- We will omit the X, Y subscript of $\mathbb{P}_{X,Y}$ in the following.

Definition (Discrete joint distributions)

Let X and Y be random variables and consider the ordered pair (X, Y) . If there are only finitely or at most countably many different possible values (x, y) for the pair (X, Y) , then X and Y have a *discrete joint distribution*.

Remarks

- X and Y have a discrete joint distribution, if both X and Y are discrete.
- It is conventional to write, e.g., $\{X = x, Y = y\}$ for $\{(X, Y) = (x, y)\}$.

Definition (Joint probability mass function)

Let X and Y be discrete random variables with outcome spaces \mathcal{X} and \mathcal{Y} , respectively. Then a *joint probability mass function (PMF)* or *bivariate PMF* of X and Y is defined as

$$p : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1], (x, y) \mapsto p(x, y) := \mathbb{P}(X = x, Y = y). \quad (56)$$

Remarks

- $\{(X = x, Y = y)\}$ means that $X = x$ AND $Y = y$.
- A joint PMF is non-negative: $p(x, y) \geq 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.
- A joint PMF is normalized: $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) = 1$.

Example (Joint probability mass function)

Let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{Y} = \{1, 2, 3, 4\}$. Then a joint PMF of X and Y is given by

$$p : \{1, 2, 3\} \times \{1, 2, 3, 4\} \rightarrow [0, 1], (x, y) \mapsto p(x, y) \quad (57)$$

with $p(x, y)$ specified according to the table below:

$p(x, y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$
$x = 1$	0.1	0.0	0.2	0.1
$x = 2$	0.1	0.2	0.0	0.0
$x = 3$	0.0	0.1	0.1	0.1

Note that $\sum_{x=1}^3 \sum_{y=1}^4 p(x, y) = 1$.

Definition (Continuous joint distribution)

Two random variables X and Y have a *continuous joint distribution*, if there exists a non-negative function $p : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}$ such that for every subset $A \subseteq \mathbb{R}^2$

$$\mathbb{P}((X, Y) \in A) = \iint_A p(x, y) dx dy. \quad (58)$$

Remarks

- The function $p : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}$ is central.
- The integral is formed over the set A .
- If $A = \{(x, y)\}$ for $(x, y) \in \mathbb{R}^2$, then

$$\mathbb{P}((X, Y) \in \{(x, y)\}) = \int_x^x \int_y^y p(\xi, v) d\xi dv = 0. \quad (59)$$

- Hence $\mathbb{P}(X = x, Y = y) = 0$ for all $(x, y) \in \mathbb{R}^2$.

Definition (Joint probability density function)

Let X and Y be continuous random variables with joint distribution \mathbb{P} . Then a function

$$p : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}, (x, y) \mapsto p(x, y) \quad (60)$$

is called a *joint probability density function (PDF)* for (X, Y) , if

- $p(x, y) \geq 0$ for all $(x, y) \in \mathbb{R}^2$,
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) dx dy = 1$, and
- for any set $A \subseteq \mathbb{R}^2$ it holds that $\mathbb{P}((X, Y) \in A) = \iint_A p(x, y) dx dy$.

Example (The bivariate Gaussian distribution)

Let X and Y have a continuous joint distribution with joint probability density function

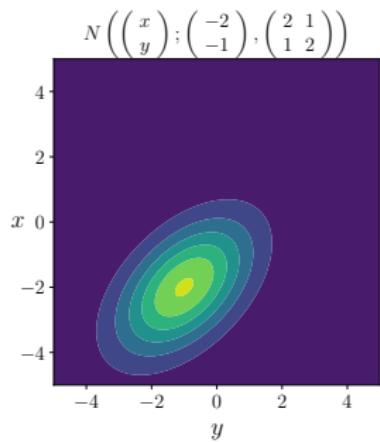
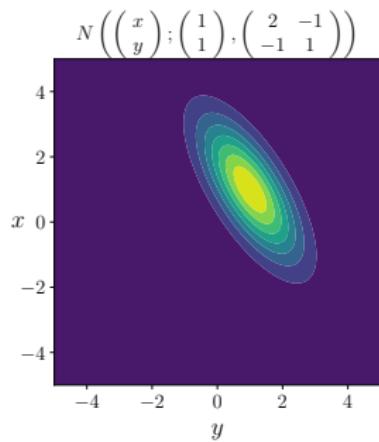
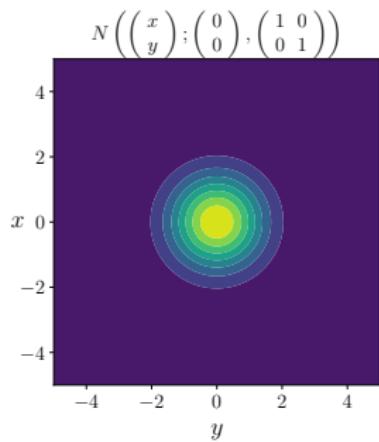
$$p : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, (x, y) \mapsto p(x, y)$$

$$:= \frac{1}{2\pi} \left| \begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{YX}^2 & \sigma_Y^2 \end{pmatrix} \right|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \right)^T \begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{YX}^2 & \sigma_Y^2 \end{pmatrix}^{-1} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \right) \right).$$

Then X and Y are said to be distributed according to a *bivariate Gaussian distribution* with expectation parameters μ_X, μ_Y , variance parameters $\sigma_X^2 > 0, \sigma_Y^2 > 0$ and covariance parameter $\sigma_{XY}^2 = \sigma_{YX}^2$.

Joint distributions

Example (The bivariate Gaussian distribution)



Definition (Joint cumulative distribution function)

The *joint cumulative distribution function* of two random variables X and Y is defined as the function

$$P : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1], (x, y) \mapsto P(x, y) := \mathbb{P}(X \leq x, Y \leq y). \quad (61)$$

Remarks

- For each fixed $x \in \mathcal{X}$, P is monotonically increasing in y .
- For each fixed $y \in \mathcal{Y}$, P is monotonically increasing in x .

Joint distributions

- Joint distributions
- **Marginal distributions**
- Independent random variables
- Conditional distributions
- Multivariate distributions
- Exercises

Definition (Marginal probability mass and probability density functions)

If the discrete random variables X and Y have a joint distribution with joint PMF $p_{X,Y}$, then the marginal PMFs of X and Y are defined as

$$p_X(x) := \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \text{ and } p_Y(y) := \sum_{x \in \mathcal{X}} p_{X,Y}(x,y), \quad (62)$$

respectively. Similarly, if the continuous random variables X and Y have a joint distribution with joint PDF $p_{X,Y}$, then the marginal PDFs of X and Y are defined as

$$p_X(x) := \int_{-\infty}^{\infty} p_{X,Y}(x,y) dy \text{ and } p_Y(y) := \int_{-\infty}^{\infty} p_{X,Y}(x,y) dx, \quad (63)$$

respectively.

Remarks

- We omit the formal definition of a marginal distribution.
- The X, Y subscripts at $p_{X,Y}, p_X, p_Y$ are commonly omitted.

Marginal distributions

Example (Marginal probability mass function)

Consider the earlier example of a joint PMF. The marginal PMFs of this joint distribution evaluate as documented below.

$p_{X,Y}(x,y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$p_X(x)$
$x = 1$	0.1	0.0	0.2	0.1	0.4
$x = 2$	0.1	0.2	0.0	0.0	0.3
$x = 3$	0.0	0.1	0.1	0.1	0.3
$p_Y(y)$	0.2	0.3	0.3	0.2	

Note that $\sum_{x=1}^3 p_X(x) = 1$ and $\sum_{y=1}^3 p_Y(y) = 1$.

Example (Marginal probability density function)

Let X and Y be distributed according to a bivariate Gaussian distribution with expectation parameters μ_X, μ_Y , variance parameters σ_X^2, σ_Y^2 and covariance parameter $\sigma_{XY}^2 = \sigma_{YX}^2$. Then X and Y have the marginal probability density functions

$$p_X : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto p_X(x) := \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2\sigma_X^2}(x - \mu_X)^2\right)$$

$$p_Y : \mathbb{R} \rightarrow \mathbb{R}_{>0}, y \mapsto p_Y(y) := \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{1}{2\sigma_Y^2}(y - \mu_Y)^2\right)$$

The marginal distributions of bivariate Gaussian distributions are thus univariate Gaussian distributions.

Joint distributions

- Joint distributions
- Marginal distributions
- **Independent random variables**
- Conditional distributions
- Multivariate distributions
- Exercises

Definition (Independent random variables)

Two random variables X and Y with outcome spaces \mathcal{X} and \mathcal{Y} , respectively, are independent, if for every $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B). \quad (64)$$

Remarks

- The definition implies that $\{X \in A\}$ and $\{Y \in B\}$ are independent events.
- Event independence implies that $\mathbb{P}(\{X \in A\} | \{Y \in B\}) = \mathbb{P}(\{X \in A\})$.
⇒ Knowing that $\{Y \in B\}$ does not change the probability that $\{X \in A\}$.

Theorem (Independent discrete and continuous random variables)

Let X and Y be discrete random variables with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y , respectively. Then X and Y are independent, if and only if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ for all } (x,y) \in \mathcal{X} \times \mathcal{Y}. \quad (65)$$

Similarly, let X and Y be continuous random variables with joint PDF $p_{X,Y}$ and marginal PDFs p_X and p_Y , respectively. Then X and Y are independent, if and only if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ for all } (x,y) \in \mathbb{R}^2. \quad (66)$$

Remarks

- For a proof, see DeGroot and Schervish (2012, Corollary 3.5.1.).
- The property $p(x,y) = p(x)p(y)$ is referred to as “factorization”.
- Independence is equivalent to the factorization of joint PMFs/PDFs.

Independent random variables

Example (Independent discrete random variables)

Consider the earlier example of a joint PMF and its associated marginal PMFs.

$p_{X,Y}(x,y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$p_X(x)$
$x = 1$	0.1	0.0	0.2	0.1	0.4
$x = 2$	0.1	0.2	0.0	0.0	0.3
$x = 3$	0.0	0.1	0.1	0.1	0.3
$p_Y(y)$	0.2	0.3	0.3	0.2	

As

$$p_{X,Y}(1,1) = 0.1 \neq 0.08 = p_X(1)p_Y(1) \quad (67)$$

the random variables X and Y are not independent.

For the same marginal PMFs, a joint PMF of independent random variables X and Y is

$p_{X,Y}(x,y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$p_X(x)$
$x = 1$	0.08	0.12	0.12	0.08	0.4
$x = 2$	0.06	0.09	0.09	0.06	0.3
$x = 3$	0.06	0.09	0.09	0.06	0.3
$p_Y(y)$	0.2	0.3	0.3	0.2	

Independent random variables

Theorem (Independent Gaussian variables)

Let X and Y have a bivariate Gaussian distribution and assume that $\sigma_{XY}^2 = \sigma_{YX}^2 = 0$. Then X and Y are independent.

Proof

We show that for a bivariate Gaussian distribution with $\sigma_{XY}^2 = \sigma_{YX}^2 = 0$ factorizes into the product of its marginal distributions.

$$\begin{aligned} p_{X,Y}(x, y) &= \frac{1}{2\pi} \left| \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix} \right|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \right)^T \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}^{-1} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \right) \right) \\ &= \frac{1}{2\pi\sqrt{\sigma_X^2\sigma_Y^2}} \exp \left(-\frac{1}{2} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix}^T \begin{pmatrix} 1/\sigma_X^2 & 0 \\ 0 & 1/\sigma_Y^2 \end{pmatrix} \begin{pmatrix} x - \mu_X \\ y - \mu_Y \end{pmatrix} \right) \\ &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp \left(\begin{pmatrix} -\frac{1}{2}(x - \mu_X) & -\frac{1}{2}(y - \mu_Y) \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma_X^2}(x - \mu_X) \\ \frac{1}{\sigma_Y^2}(y - \mu_Y) \end{pmatrix} \right) \quad (68) \\ &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp \left(-\frac{1}{2\sigma_X^2}(x - \mu_X)^2 - \frac{1}{2\sigma_Y^2}(y - \mu_Y)^2 \right) \\ &= \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp \left(-\frac{1}{2\sigma_X^2}(x - \mu_X)^2 \right) \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp \left(-\frac{1}{2\sigma_Y^2}(y - \mu_Y)^2 \right) \\ &= p_X(x)p_Y(y) \end{aligned}$$

Joint distributions

- Joint distributions
- Marginal distributions
- Independent random variables
- **Conditional distributions**
- Multivariate distributions
- Exercises

Definition (Conditional PMF and conditional discrete distribution)

For a joint distribution of two discrete random variables with joint PMF $p_{X,Y}$, the *conditional probability mass function* of Y given $X = x$ is

$$p_{Y|X}(\cdot|x) : \mathcal{Y} \rightarrow [0, 1], y \mapsto p_{Y|X}(y|x) := \frac{p_{X,Y}(x,y)}{p_X(x)} \text{ for } p_X(x) > 0. \quad (69)$$

The discrete distribution with PMF $p_{Y|X}(\cdot|X = x)$ is called the the *conditional distribution* of Y given that $X = x$. The conditional distribution of X given $Y = y$ is defined analogously.

Conditional distributions

Example (Discrete conditional distributions)

Consider the earlier example of a joint PMF and its associated marginal PMFs.

$p_{X,Y}(x,y)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$p_X(x)$
$x = 1$	0.1	0.0	0.2	0.1	0.4
$x = 2$	0.1	0.2	0.0	0.0	0.3
$x = 3$	0.0	0.1	0.1	0.1	0.3
$p_Y(y)$	0.2	0.3	0.3	0.2	

Then the conditional distributions of Y given X are given by

$p_{Y X}(y x)$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	
$p_{Y X}(y x=1)$	$\frac{0.1}{0.4} = \frac{1}{4}$	$\frac{0.0}{0.4} = 0$	$\frac{0.2}{0.4} = \frac{1}{2}$	$\frac{0.1}{0.4} = \frac{1}{4}$	$\sum_{y=1}^4 p_{Y X}(y x) = 1$
$p_{Y X}(y x=2)$	$\frac{0.1}{0.3} = \frac{1}{3}$	$\frac{0.2}{0.3} = \frac{2}{3}$	$\frac{0.0}{0.3} = 0$	$\frac{0.0}{0.3} = 0$	$\sum_{y=1}^4 p_{Y X}(y x) = 1$
$p_{Y X}(y x=3)$	$\frac{0.0}{0.3} = 0$	$\frac{0.1}{0.3} = \frac{1}{3}$	$\frac{0.1}{0.3} = \frac{1}{3}$	$\frac{0.1}{0.3} = \frac{1}{3}$	$\sum_{y=1}^4 p_{Y X}(y x) = 1$

Note the qualitative similarity of $p_{X,Y}(x,y)$ and $p_{Y|X}(y|x)$.

Definition (Conditional PDF and conditional continuous distribution)

For a joint distribution of two continuous random variables with joint PDF $p_{X,Y}$, the *conditional probability density function* of Y given X is

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \text{ for } p_X(x) > 0. \quad (70)$$

Remarks

- $p_X(x) > 0$ refers to a probability density, not a probability.
- Formally relating conditional PDFs to conditional probabilities is demanding.
- A rigorous framework is afforded by the theory of *Markov kernels*.

Joint distributions

- Joint distributions
- Marginal distributions
- Independent random variables
- Conditional distributions
- **Multivariate distributions**
- Exercises

Definition (Joint cumulative distribution function)

The joint cumulative distribution function of n random variables X_1, \dots, X_n is

$$P : \mathbb{R}^n \rightarrow [0, 1], (x_1, \dots, x_n) \mapsto \mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n). \quad (71)$$

Remarks

- It is convenient to use the vector notation $X = (X_1, \dots, X_n)$.
- An X thus defined is referred to as *random vector*.
- It is also convenient to use the vector notation $x = (x_1, \dots, x_n) \in \mathbb{R}^n$.
- We are concerned with distributions of X with joint CDF values $P(x)$.

Definition (Joint discrete and continuous multivariate distributions)

n random variables X_1, \dots, X_n have a *discrete multivariate distribution*, if the random vector (X_1, \dots, X_n) takes on only a finite number or countably many values $(x_1, \dots, x_n) \in \mathbb{R}^n$. The joint PMF of X_1, \dots, X_n is then defined as

$$p : \mathbb{R}^n \rightarrow [0, 1], (x_1, \dots, x_n) \mapsto p(x_1, \dots, x_n) := \mathbb{P}(X_1 = x_1, \dots, X_n = x_n), \quad (72)$$

or in vector notation as

$$p : \mathbb{R}^n \rightarrow [0, 1], x \mapsto p(x) := \mathbb{P}(X = x). \quad (73)$$

Similarly, n continuous random variables X_1, \dots, X_n have a *continuous multivariate distribution*, if there is a non-negative function $p : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$, such that for every $A \subseteq \mathbb{R}^n$

$$\mathbb{P}((X_1, \dots, X_n) \in A) = \int \cdots \int_A p(x_1, \dots, x_n) dx_1 \dots dx_n, \quad (74)$$

or in vector notation

$$\mathbb{P}(X \in A) = \int \cdots \int p(x) dx. \quad (75)$$

p is called the joint PDF of X_1, \dots, X_n .

Multivariate Gaussian distributions

Example (Multivariate Gaussian distributions)

Let X be an n -dimensional random vector with outcome set \mathbb{R}^n and (joint) PDF

$$p : \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto p(x) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right). \quad (76)$$

Then X is said to be distributed according to a *multivariate or n-dimensional Gaussian distribution* with *expectation parameter* $\mu \in \mathbb{R}^n$ and *positive-definite covariance matrix parameter* $\Sigma \in \mathbb{R}^{n \times n}$, for which we write $X \sim N(\mu, \Sigma)$. We abbreviate the PDF of a multivariate Gaussian distribution by

$$N(x; \mu, \Sigma) := (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right). \quad (77)$$

Remarks

- The parameter $\mu \in \mathbb{R}^n$ specifies the location of highest probability density in \mathbb{R}^n .
- The diagonal elements of Σ specify the width of the distribution w.r.t. X_1, \dots, X_n .
- The i, j th off-diagonal element of Σ specifies the covariation of X_i and X_j .
- The term $(2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}}$ is the normalization constant for the exponential term.

Marginal distributions

The marginal distribution of each $X_i, i = 1, \dots, n$ can be derived from the joint distribution of X_1, \dots, X_n by summation/integration over the remaining $n - 1$ variables for PMFs/PDFs, respectively.

For example, if an $X = (X_1, \dots, X_n)$ has PMF p_X , then the marginal PMF of X_i is evaluated as

$$p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p_X(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n). \quad (78)$$

Similarly, if an $X = (X_1, \dots, X_n)$ has PDF p_X , then the marginal PDF of X_i is evaluated as

$$p_{X_i}(x_i) = \int \cdots \iint \cdots \int p_X(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) dx_1 \dots dx_i dx_{i+1} \dots dx_n. \quad (79)$$

Marginal distributions

More generally, the marginal joint distribution of any k of the $X_i, i = 1, \dots, n$ can be derived from the joint distribution of X_1, \dots, X_n by summation/integration over the remaining $n - k$ variables for PMFs/PDFs, respectively.

For example, if $X = (X_1, X_2, X_3, X_4)$ has PMF p_X , then the marginal PMF of X_2, X_3 is evaluated as

$$p_{X_2, X_3}(x_2, x_3) = \sum_{x_1} \sum_{x_4} p_X(x_1, x_2, x_3, x_4). \quad (80)$$

Similarly, if an $X = (X_1, X_2, X_3, X_4)$ has PDF p_X , then the marginal PDF of X_1, X_2 is evaluated as

$$p_{X_1, X_2}(x_1, x_2) = \iint p_X(x_1, x_2, x_3, x_4) dx_3 dx_4. \quad (81)$$

Definition (Multivariate conditional PMFs/PDFs)

Suppose that the random vector $X = (X_1, \dots, X_n)$ is divided into two subvectors Y and Z , where Y is a k -dimensional random vector comprising k of the n random variables in X and Z is an $(n - k)$ -dimensional random vector comprising the remaining $(n - k)$ random variables in X . Suppose also that the n -dimensional joint PMF/PDF of (Y, Z) is $p_{Y,Z}$ and that the marginal $(n - k)$ -dimensional PMF/PDF of Z is p_Z . Then for every $z \in \mathbb{R}^{n-k}$ such that $p_Z(z) > 0$, the conditional k -dimensional PMF/PDF of Y given $Z = z$ is defined as

$$p_{Y|Z} : \mathbb{R}^k \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_{Y|Z}(y|z) := \frac{p_{Y,Z}(y,z)}{p_Z(z)}. \quad (82)$$

Remarks

- The definition of $p_{Y|Z}(y|z)$ may be rewritten as $p_{Y,Z}(y,z) = p_{Y|Z}(y|z)p_Z(z)$
- This allows for constructing joint from conditional and marginal distributions.
- The multivariate law of total probability and Bayes theorem follow directly.

Theorem (Multivariate Law of Total Probability and Bayes Theorem)

With the conditions and notations used in the definition of multivariate conditional PMFs/PDFs, and a continuous X , the marginal PDF of Y is given by

$$p_Y : \mathbb{R}^k \rightarrow \mathbb{R}_{\geq 0}, y \mapsto \int_{\mathbb{R}^k} p_{Y|Z}(y|z)p_Z(z) dz \quad (83)$$

and the conditional PDF of Z given Y is given by

$$p_{Z|Y} : \mathbb{R}^{n-k} \rightarrow \mathbb{R}_{\geq 0}, z \mapsto p_{Z|Y}(z|y) = \frac{p_Y(y|z)p_Z(z)}{p_Y(y)}. \quad (84)$$

For discrete X , integration is replaced by summation.

Definition (n independent random variables)

The random variables X_1, \dots, X_n are independent, if for every A_1, \dots, A_n

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n \mathbb{P}(X_i \in A_i). \quad (85)$$

If the random variables have a joint PMF/PDF p_X , then independence holds if

$$p_X(x_1, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i). \quad (86)$$

Remarks

- The PMF/PDF p_X is said to factorize into the PMFs/PDFs $p_{X_i}, i = 1, \dots, n$.
- Independence of n random variables is a common assumption in statistics.
- n independent random variables X_1, \dots, X_n
 \Leftrightarrow n -dimensional random vector $X = (X_1, \dots, X_n)$ with factorized PMF/PDF.

Definition (Independent and identically distributed (i.i.d.) random variables)

The random variables X_1, \dots, X_n are called independent and identically distributed, if

- (1) X_1, \dots, X_n are independent random variables, and
- (2) each X_i has the same marginal distribution.

Remarks

- If the marginal distributions have CDF P , we write $X_1, \dots, X_n \sim P$.
- If P has an associated PMF/PDF p , we write $X_1, \dots, X_n \sim p$.
- X_1, \dots, X_n is also called a *random sample* of size n from p .
- n i.i.d. random variables $X_1, \dots, X_n \Leftrightarrow n$ -dimensional random vector X with
 - factorized PMF/PDF $p_X(x) = \prod_{i=1}^n p_{X_i}(x_i)$
 - identical marginal PMFs/PDFs $p_{X_i}(x_i) = p_{X_j}(x_j), 1 \leq i, j \leq n$.

Joint distributions

- Joint distributions
- Marginal distributions
- Independent random variables
- Conditional distributions
- Multivariate distributions
- **Exercises**

Study Questions

1. Write down the definition of a joint PMF of two discrete random variables X and Y with finite outcome set.
2. Write down the definition of the joint PDF of two continuous random variables X and Y each taking values in \mathbb{R} .
3. Write down the definitions of the marginal PMFs and PDFs of a joint distribution of two random variables X and Y with joint PMF or PDF $p_{X,Y}$.
4. Write down the definition of the independence of two random variables X and Y .
5. Write down the necessary and sufficient condition for the independence of two random variables X and Y with joint PMF/PDF $p_{X,Y}$.
6. Write down the definitions of conditional PMFs and PDFs.
7. Write down the definition of a multivariate Gaussian PDF and comment on the meaning of its parameters.
8. Write down the definition of the independence of n random variables X_i .
9. What does it mean for n random variables X_1, \dots, X_n to be i.i.d.?
10. What does it mean for X_1, \dots, X_n to be a random sample of size n from p ?

Theoretical Exercises

1. Construct a mixed joint distribution comprising a marginal uniform distribution and a conditional binomial distribution DeGroot and Schervish (2012, Example 3.6.7).
2. Review the theory of multivariate Gaussian distributions.
3. Show that the PDF of n independent univariate Gaussian random variables corresponds to the PDF of an n -variate Gaussian PDF with diagonal covariance matrix parameter. In addition, consider the case of n i.i.d. univariate Gaussian random variables and its multivariate Gaussian analogon.

Theoretical Exercise 1

We first note that with $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$, we have the following multiplication rule for distributions (DeGroot and Schervish, 2012, Theorem 3.6.2):

Let X and Y be random variables such that X has PMF or PDF $p_X(x)$ and Y has PMF or PDF $p_Y(y)$. Also, assume that the conditional PMF or PDF of X given $Y = y$ is $p_{X|Y}(x|y)$, while the conditional PMF or PDF of Y given $X = x$ is $p_{Y|X}(y|x)$. Then for each y , such that $p_Y(y) > 0$ and each x ,

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y), \quad (87)$$

where $p_{X,Y}(x,y)$ is the joint PMF, PDF, or mixed PMF/PDF of X and Y . Similarly, for each x such that $p_X(x) > 0$ and each y ,

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x). \quad (88)$$

Note that if $p_Y(y_0) = 0$ for some value y_0 , then it can be assumed without loss of generality that $p_{X,Y}(x,y_0) = 0$ for all values of x . In this case, both sides of eq. (87) are 0, and the fact that $p_{X|Y}(x|y_0)$ is not defined becomes irrelevant. A similar statement applies to eq. (88).

Theoretical Exercise 1 (cont.)

We next consider an exemplary scenario (DeGroot and Schervish, 2012, Example 3.6.7)

- Assume that the conditional distribution of a discrete random variable X with outcome space \mathbb{N}_0^n given a random variable M is defined by a Binomial PMF with parameters M and n , i.e.,

$$p_{X|M}(x|\mu) := \text{Bin}(x; M, n) = \binom{n}{x} \mu^x (1 - \mu)^{n-x}. \quad (89)$$

- Also assume that the marginal distribution of M is given by the continuous uniform distribution of $[0, 1]$, i.e.,

$$p_M(\mu) := U(\mu, 0, 1) = 1 \quad (90)$$

- Then the joint distribution of X and M is given by the mixed PMF/PDF

$$p_{X,M}(x, \mu) = \binom{n}{x} \mu^x (1 - \mu)^{n-x} \cdot 1 = \binom{n}{x} \mu^x (1 - \mu)^{n-x}. \quad (91)$$

for $x = 0, \dots, n$ and $0 \leq \mu \leq 1$.

Exercises

Theoretical Exercise 2

In multivariate Gaussian distributions

- all marginal distributions are also Gaussian,
- all conditional distributions are also Gaussian,
- the parameters of marginal and conditional distributions can be computed based on the parameters of the joint distribution, and
- joint Gaussian distributions and their parameters can be composed from marginal and conditional Gaussian distributions and their parameters.

These properties of Gaussian distributions are essential in many applications, such as

- Mixed linear models and covariance component estimation
- Conjugate Bayesian inference for the Gaussian
- Bayesian filtering, especially Kalman filters

Proofs of these properties can be found e.g. in Anderson (2003).

Exercises

Theoretical Exercise 2 - Marginal Gaussian distributions

Let $X = (X_1, \dots, X_n)$ be distributed according to an n -dimensional Gaussian distribution, the expectation and covariance matrix parameters of which partition for $n = k + m$ according to

$$\mu = \begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix} \in \mathbb{R}^n, \quad (92)$$

where $\mu_y \in \mathbb{R}^k$ and $\mu_z \in \mathbb{R}^m$ and

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zy} & \Sigma_{zz} \end{pmatrix} \in \mathbb{R}^{n \times n}, \quad (93)$$

where $\Sigma_{yy} \in \mathbb{R}^{k \times k}$, $\Sigma_{yz} \in \mathbb{R}^{k \times m}$, $\Sigma_{zy} \in \mathbb{R}^{m \times k}$, and $\Sigma_{zz} \in \mathbb{R}^{m \times m}$.

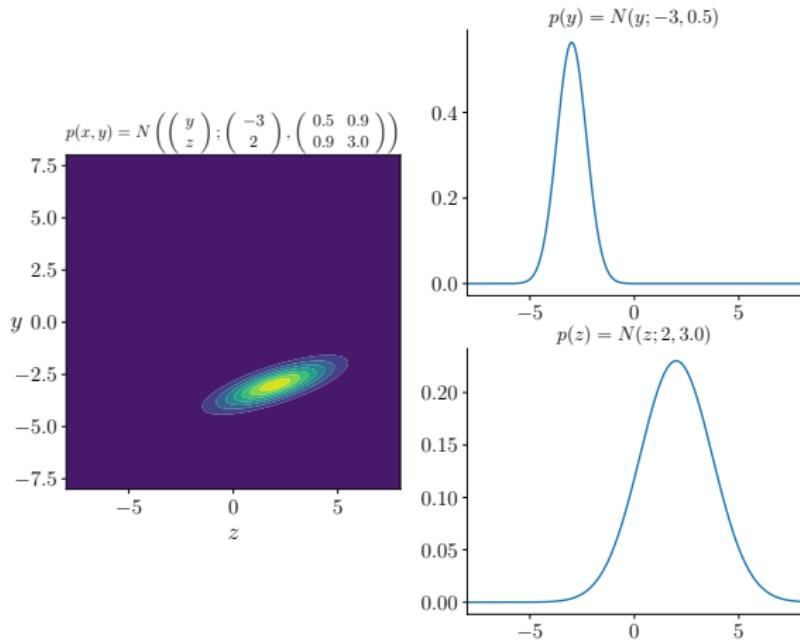
Then $Y := (X_1, \dots, X_k)$ and $Z := (X_{k+1}, \dots, X_n)$ are distributed as

$$Y \sim N(\mu_y, \Sigma_{yy}) \text{ and } Z \sim N(\mu_z, \Sigma_{zz}), \quad (94)$$

respectively.

Exercises

Theoretical Exercise 2 - Marginal Gaussian distributions



Theoretical Exercise 2 - Conditional Gaussian distributions

Given an $m+n$ -dimensional random vector (X, Y) distributed according to a Gaussian distribution with PDF

$$p_{X,Y} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}_{>0}, (x, y) \mapsto p_{X,Y}(x, y) := N((x, y); \mu_{x,y}, \Sigma_{x,y}), \quad (95)$$

where

$$\mu_{x,y} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma_{x,y} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \quad (96)$$

for $x, \mu_x \in \mathbb{R}^m, y, \mu_y \in \mathbb{R}^n$ and $\Sigma_{xx} \in \mathbb{R}^{m \times m}, \Sigma_{xy} \in \mathbb{R}^{m \times n}, \Sigma_{yy} \in \mathbb{R}^{n \times n}$, the distribution of X given Y has an m -dimensional conditional PDF

$$p_{X|Y}(\cdot|y) : \mathbb{R}^m \rightarrow \mathbb{R}_{>0}, x \mapsto p_{X|Y}(x|y) := N(x; \mu_{x|y}, \Sigma_{x|y}), \quad (97)$$

where

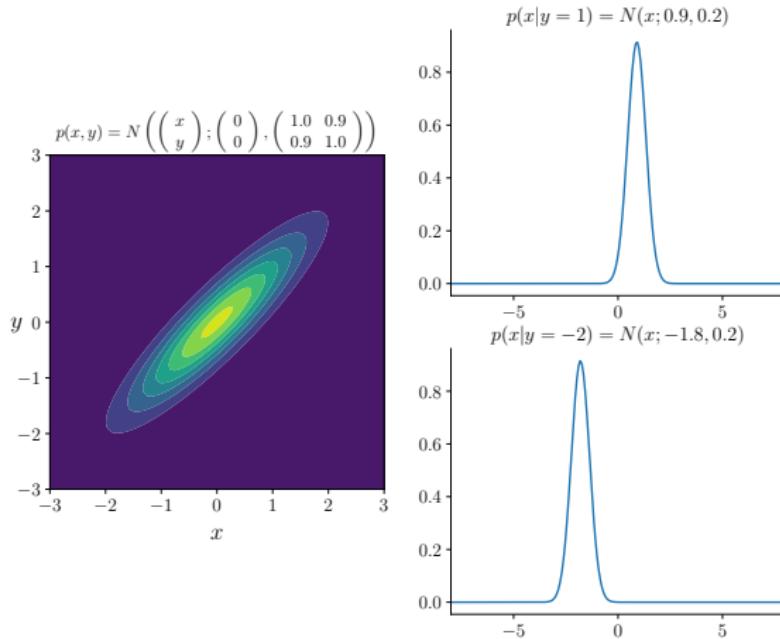
$$\mu_{x|y} = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (Y - \mu_y) \in \mathbb{R}^m \quad (98)$$

and

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \in \mathbb{R}^{m \times m}. \quad (99)$$

Exercises

Theoretical Exercise 2 - Conditional Gaussian distributions



Exercises

Theoretical Exercise 2 - Joint Gaussian distributions

Given an m -dimensional random vector X distributed according to a Gaussian distribution with PDF

$$p_X : \mathbb{R}^m \rightarrow \mathbb{R}_{>0}, x \mapsto p_X(x) := N(x; \mu_x, \Sigma_{xx}) \text{ for } \mu_x \in \mathbb{R}^m, \Sigma_{xx} \in \mathbb{R}^{m \times m}, \quad (100)$$

a matrix $A \in \mathbb{R}^{n \times m}$, a vector $b \in \mathbb{R}^n$, and a n -dimensional random vector Y conditionally distributed according to a Gaussian distribution with conditional PDF

$$p_{Y|X}(\cdot|x) : \mathbb{R}^n \rightarrow \mathbb{R}_{>0}, y \mapsto p_{Y|X}(y|x) := N(y; AX + b, \Sigma_{yy}) \text{ for } \Sigma_{yy} \in \mathbb{R}^{n \times n} \quad (101)$$

the $m + n$ -dimensional random vector (X, Y) is distributed according to a Gaussian distribution with joint PDF

$$p_{X,Y} : \mathbb{R}^{m+n} \rightarrow \mathbb{R}_{>0}, (x, y) \mapsto p_{X,Y}(x, y) = N((x, y); \mu_{x,y}, \Sigma_{x,y}), \quad (102)$$

where $\mu_{x,y} \in \mathbb{R}^{m+n}$ and $\Sigma_{x,y} \in \mathbb{R}^{m+n \times m+n}$, and in particular

$$\mu_{x,y} = \begin{pmatrix} \mu_x \\ A\mu_x + b \end{pmatrix} \text{ and } \Sigma_{x,y} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xx}A^T \\ A\Sigma_{xx} & \Sigma_{yy} + A\Sigma_{xx}A^T \end{pmatrix}. \quad (103)$$

Note that the parameters of the Gaussian joint distribution can be computed from the parameters of the Gaussian marginal and Gaussian conditional distributions.

Theoretical Exercise 3

Theorem (n independent Gaussian random variables)

For $i = 1, \dots, n$, let $N(x_i; \mu_i, \sigma_i^2)$ denote the PDF of n independent univariate Gaussian random variables X_1, \dots, X_n with $\mu_1, \dots, \mu_n \in \mathbb{R}$ and $\sigma_1^2, \dots, \sigma_n^2 > 0$. Further, let $N(x; \mu, \Lambda)$ denote the probability density function of an n -variate random vector X with $\mu := (\mu_1, \dots, \mu_n)$ and diagonal covariance matrix $\Lambda \in \mathbb{R}^{n \times n}$ with diagonal elements $\sigma_1^2, \dots, \sigma_n^2 > 0$. Then it holds that

$$p_X(x) = p_{X_1, \dots, X_n}(x_1, \dots, x_n) \quad (104)$$

and in particular that

$$N(x; \mu, \Lambda) = \prod_{i=1}^n N(x_i; \mu_i, \sigma_i^2). \quad (105)$$

Theoretical Exercise 3 (cont.)

Proof

$$\begin{aligned}
 N(x; \mu, \Lambda) &= (2\pi)^{-\frac{n}{2}} |\Lambda|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Lambda^{-1} (x - \mu)\right) \\
 &= \left(\prod_{i=1}^n 2\pi^{-\frac{1}{2}}\right) \left(\prod_{i=1}^n \sigma_i^2\right)^{-\frac{1}{2}} \exp\left(\left(\sum_{i=1}^n -\frac{1}{2\sigma_i^2} (x_i - \mu_i)^2\right)\right) \\
 &= \prod_{i=1}^n \left(2\pi\sigma_i^2\right)^{-\frac{1}{2}} \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma_i^2} (x_i - \mu_i)^2\right) \tag{106} \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2} (x_i - \mu_i)^2\right) \\
 &= \prod_{i=1}^n N(x_i; \mu_i, \sigma_i^2).
 \end{aligned}$$

□

Theoretical Exercise 3 (cont.)

Theorem (n i.i.d. Gaussian random variables)

For $i = 1, \dots, n$ let $N(x_i; \mu, \sigma^2)$ denote the PDF of n i.i.d. univariate Gaussian random variables X_1, \dots, X_n with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Further, let $N(x; 1_n\mu, \sigma^2 I_n)$ denote the PDF of an n -variate random vector X , where $1_n \in \mathbb{R}^n$ denotes a vector of all ones and $I_n \in \mathbb{R}^{n \times n}$ denotes the n -dimensional identity matrix. Then it holds that

$$p_X(x) = p_{X_1, \dots, X_n}(x_1, \dots, x_n) \quad (107)$$

and in particular that

$$N(x; 1_n\mu, \sigma^2 I_n) = \prod_{i=1}^n N(x_i; \mu, \sigma^2). \quad (108)$$

Proof

The statement follows directly from the proof of the multivariate/univariate identity of n independent Gaussian random variables.

□

Programming Exercises

1. Write a simulation that demonstrates that the marginal distributions of a bivariate Gaussian distribution with expectation parameter and covariance parameters

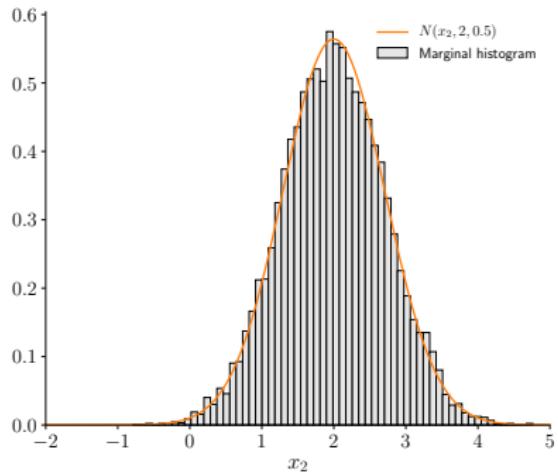
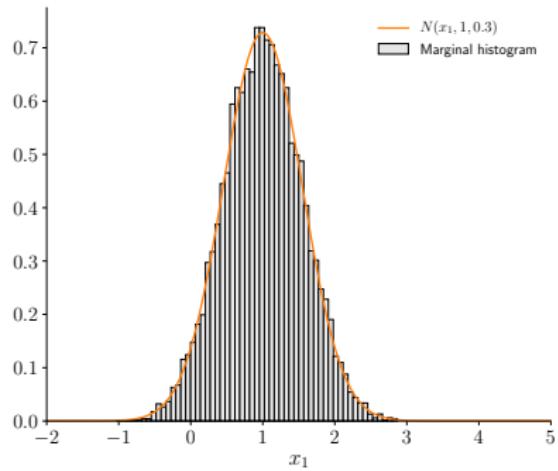
$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.5 \end{pmatrix}, \quad (109)$$

respectively, are given by univariate Gaussian distributions with expectation parameters $\mu_1 = 1, \mu_2 = 2$ and variance parameters $\sigma^2 = 0.3$ and $\sigma^2 = 0.5$, respectively.

2. Write a simulation that verifies that obtaining samples from 2 independent univariate Gaussian distributions with parameters $\mu_i, \sigma_i^2 > 0, i = 1, 2$ is equivalent to obtaining samples from a two-dimensional Gaussian distribution with the appropriately specified parameters $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2 \times 2}$.
3. Write a simulation that exemplary verifies the analytical results on conditional Gaussian distributions for the case of a bivariate Gaussian distribution.

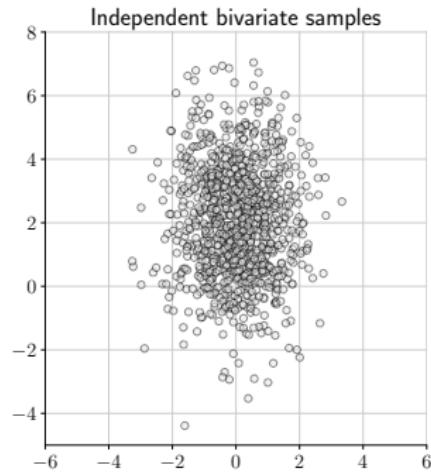
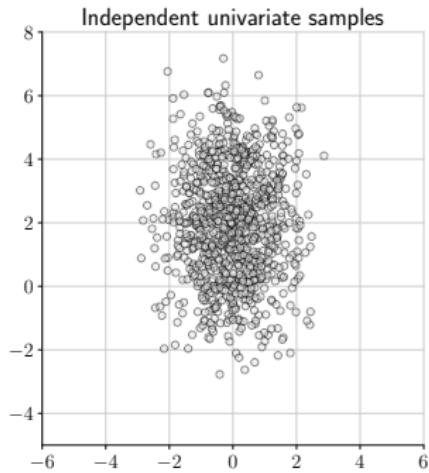
Exercises

Programming Exercise 1



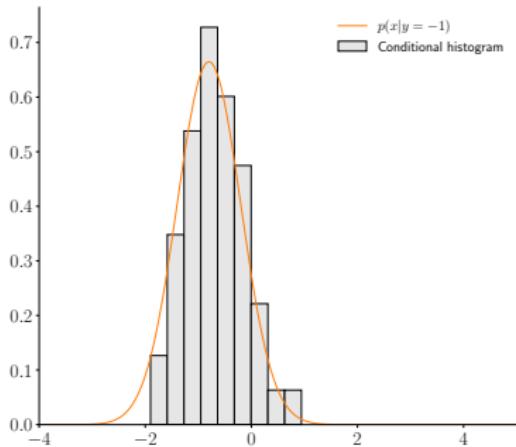
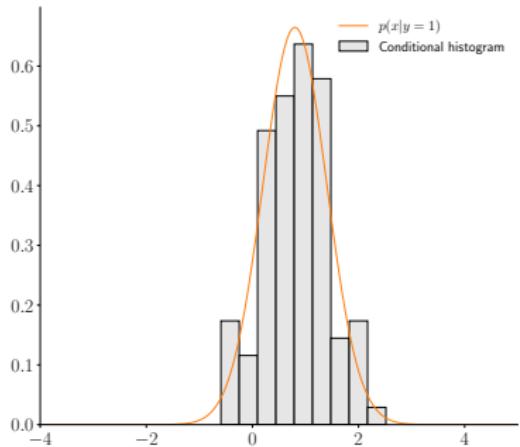
Exercises

Programming Exercise 2



Exercises

Programming Exercise 3



(5) Transformations

Bibliographic remarks

The majority of the presented material follows DeGroot and Schervish (2012, Sections 3.8 - 3.9). The results on the combinations and transformations of Gaussian variables follows Casella and Berger (2012, Section 5.3)

Fundamental question

- Assume a random entity X with outcome space \mathcal{X} has distribution \mathbb{P}_X .
- Consider a function $f : \mathcal{X} \rightarrow \mathcal{Y}, x \mapsto f(x)$ of known functional form.
- What is the distribution of the random variable $Y := f(X)$?

Essential applications

- Estimators and statistics are functions of random variables.
- Frequentist theory concerns the distributions of estimators and statistics.
- Probabilistic models are functional transformations of parameter priors.

Transformations

- The probability integral transform
- The univariate PDF transform
- The multivariate PDF transform
- Linear combinations
- Gaussian transformations
- Exercises

Transformations

- **The probability integral transform**
- The univariate PDF transform
- The multivariate PDF transform
- Linear combinations
- Gaussian transformations
- Exercises

Theorem (The probability integral transform)

Let X be a continuous random variable with CDF P_X and let $Y := P_X(X)$ be the *probability integral transform*. Then $Y \sim U(0, 1)$. Moreover, let $Y \sim U(0, 1)$ and let P_X^{-1} be the inverse of a continuous CDF P_X . Then $X := P_X^{-1}(Y)$ has CDF P_X .

Remarks

- Pseudo-random number generators of samples from the uniform distribution can be used to generate samples from arbitrary distributions.
- Let Y have the uniform distribution on $[0, 1]$ and let P_X^{-1} denote an inverse CDF. Then $X := P_X^{-1}(Y)$ has CDF P_X .
- Equivalently, let $Y_1, \dots, Y_n \sim U(0, 1)$. Then $P_X^{-1}(Y_1), \dots, P_X^{-1}(Y_n)$ will appear to form an i.i.d. sample from X .

The probability integral transform

Proof

We first consider the CDF P_Y of Y and show that it corresponds to the CDF of the uniform distribution,

$$P_Y : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto P_Y(y) := \begin{cases} 0, & y < 0 \\ y, & 0 \leq y \leq 1 \\ 1, & y > 1. \end{cases} \quad (110)$$

To see this, we first note that because P_X is a CDF, we have $0 \leq P_X(x) \leq 1$ for $x \in \mathbb{R}$ and thus $\mathbb{P}(Y < 0) = 0$. Further, with $\mathbb{P}(Y > 1) = 0$ it follows that $\mathbb{P}(Y \leq 1) = 1 - \mathbb{P}(Y > 1) = 1$. Moreover, we have

$$P_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(P_X(X) \leq y) = \mathbb{P}\left(P_X^{-1}(P_X(X)) \leq P_X^{-1}(y)\right) \quad (111)$$

and thus

$$P_Y(y) = \mathbb{P}\left(X \leq P_X^{-1}(y)\right) = P_X\left(P_X^{-1}(y)\right) = y. \quad (112)$$

Finally, with

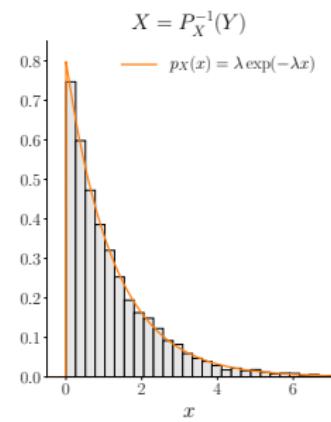
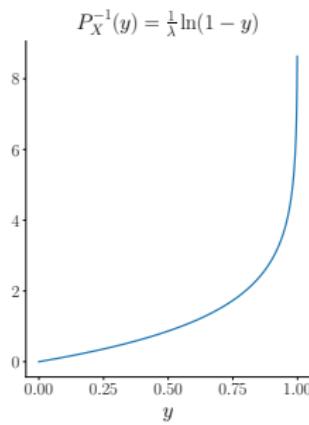
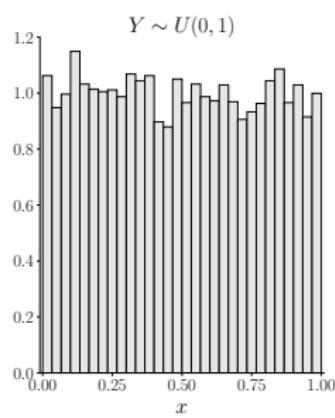
$$Y = P_X(X) \Leftrightarrow P_X^{-1}(Y) = P_X^{-1}(P_X(X)) \Leftrightarrow X = P_X^{-1}(Y) \quad (113)$$

the second part of the theorem follows immediately, because we assumed that X has CDF P_X .

□

The probability integral transform

Example: Generating samples from an exponential distribution



Transformations

- The probability integral transform
- **The univariate PDF transform**
- The multivariate PDF transform
- Linear combinations
- Gaussian transformations
- Exercises

Theorem (The univariate PDF transform for bijective transformations)

Let X be a random variable with PDF p_X and for which $\mathbb{P}(]a, b[) = 1$, where a and/or b are either finite or infinite. Let $Y = f(X)$, where f is differentiable and bijective for $]a, b[$. Let $f(]a, b[)$ be the image of $]a, b[$ under f . Finally, let $f^{-1}(y)$ denote the inverse of $f(x)$ for $y \in f(]a, b[)$ and let $f'(x)$ denote the first derivative of f at x . Then the PDF of Y is given by

$$p_Y : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) := \begin{cases} \frac{1}{|f'(f^{-1}(y))|} p_X(f^{-1}(y)) & \text{for } y \in f(]a, b[) \\ 0 & \text{for } y \in \mathbb{R} \setminus f(]a, b[). \end{cases} \quad (114)$$

Remarks

- Linear-affine transformations of Gaussians are an important application.
- The Z-transformation is an important application.

The univariate PDF transform

Proof

We first note that because f is a differentiable bijective function on the open interval $]a, b[$ it is either strictly increasing or strictly decreasing. Assume first that f is increasing on $]a, b[$. Then f^{-1} is also increasing for all $y \in f(]a, b[)$ and

$$P_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(f(X) \leq y) = \mathbb{P}\left(f^{-1}(f(X)) \leq f^{-1}(y)\right) = \mathbb{P}\left(X \leq f^{-1}(y)\right) = P_X\left(f^{-1}(y)\right). \quad (115)$$

P_Y is thus differentiable at all y where both f^{-1} and P_X is differentiable at $f^{-1}(y)$. With the chain rule of differentiation and the inverse function theorem $(f^{-1}(x))' = 1/f'(f^{-1}(x))$, it thus follows that the PDF p_Y evaluates to

$$p_Y(y) = \frac{d}{dy} P_Y(y) = \frac{d}{dy} P_X\left(f^{-1}(y)\right) = p_X\left(f^{-1}(y)\right) \frac{d}{dy} f^{-1}(y) = \frac{1}{f'\left(f^{-1}(y)\right)} p_X\left(f^{-1}(y)\right), \quad (116)$$

Because f^{-1} is increasing, $d/dy(f^{-1}(y))$ is positive, and the theorem holds. Similarly, if f is decreasing on $]a, b[$, then f^{-1} is also decreasing for all $y \in f(]a, b[)$ and for each $y \in f(]a, b[)$ and

$$P_Y(y) = \mathbb{P}(f(X) \leq y) = \mathbb{P}\left(f^{-1}(f(X)) \geq f^{-1}(y)\right) = \mathbb{P}\left(X \geq f^{-1}(y)\right) = 1 - P_X\left(f^{-1}(y)\right), \quad (117)$$

With the chain rule of differentiation and the inverse function theorem, it thus follows that the PDF p_Y evaluates to

$$p_Y(y) = \frac{d}{dy} (1 - P_Y(y)) = -\frac{d}{dy} P_X\left(f^{-1}(y)\right) = -p_X\left(f^{-1}(y)\right) \frac{d}{dy} f^{-1}(y) = -\frac{1}{f'\left(f^{-1}(y)\right)} p_X\left(f^{-1}(y)\right). \quad (118)$$

Since f^{-1} is strictly decreasing, $d/dy(f^{-1}(y))$ is negative, such that $-d/dy(f^{-1}(y))$ equals $|d/dy(f^{-1}(y))|$ and the theorem holds.

□

The univariate PDF transform

Theorem (The univariate PDF transform for linear-affine functions)

Let X be a random variable with PDF p_X and let $Y = f(X)$ with $f(X) := aX + b$ for $a \neq 0$. Then the PDF of Y is given by

$$p_Y : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) := \frac{1}{|a|} p_X \left(\frac{y - b}{a} \right). \quad (119)$$

Proof

We first note that

$$f^{-1} : \mathbb{R} \rightarrow \mathbb{R}, y \mapsto f^{-1}(y) = \frac{y - b}{a} \quad (120)$$

because then $f \circ f^{-1} = \text{id}_{\mathbb{R}}$ as

$$f(f^{-1}(x)) = a \left(\frac{x - b}{a} \right) + b = x - b + b = x \text{ for all } x \in \mathbb{R}. \quad (121)$$

We next note that

$$f' : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f'(x) = \frac{d}{dx}(ax + b) = a. \quad (122)$$

We thus have

$$p_Y : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) = \frac{1}{|f'(f^{-1}(y))|} p_X(f^{-1}(y)) = \frac{1}{|a|} p_X \left(\frac{y - b}{a} \right). \quad (123)$$

□

Theorem (The PDF transform for piecewise bijective transformations)

Let X be a random variable with outcome set \mathcal{X} and PDF p_X . Assume further that $Y = f(X)$, where f is such that the outcome set of X can be partitioned into a finite number of sets $\mathcal{X}_1, \dots, \mathcal{X}_k$ with a corresponding number of sets $\mathcal{Y}_1 := f(\mathcal{X}_1), \dots, \mathcal{Y}_k := f(\mathcal{X}_k)$ in the outcome set \mathcal{Y} of Y (which may not be mutually exclusive) such that the transformation f is bijective for all $\mathcal{X}_1, \dots, \mathcal{X}_k$. Let further f_i^{-1} denote the inverse of f on \mathcal{X}_i and assume that f'_i exists and is continuous for all $i = 1, \dots, k$. Then the PDF of Y is given by

$$p_Y : \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) := \sum_{i=1}^k 1_{\mathcal{Y}_i}(y) \frac{1}{|f'_i(f_i^{-1}(y))|} p_X(f_i^{-1}(y)). \quad (124)$$

Remarks

- Proof omitted.
- The derivation of the χ^2 distribution is an important application.

Transformations

- The probability integral transform
- The univariate PDF transform
- **The multivariate PDF transform**
- Linear combinations
- Gaussian transformations
- Exercises

Theorem (The multivariate probability density function transform)

Let X be an n -dimensional random vector with PDF p_X and let $Y = f(X)$ be an n -dimensional random vector, where $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is differentiable and bijective. Let $f^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ denote the inverse of f . Let further

$$J^f(x) = \left(\frac{\partial}{\partial x_j} f_i(x) \right)_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n} \quad (125)$$

denote the Jacobian matrix of f at $x \in \mathbb{R}$, let $|J^f(x)|$ denote its determinant, and assume that $|J^f(x)| \neq 0$ for all $x \in \mathbb{R}^n$. Then the PDF of Y is given by

$$p_Y : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) := \begin{cases} \frac{1}{|J^f(f^{-1}(y))|} p_X(f^{-1}(y)) & \text{for } y \in f(\mathbb{R}^n) \\ 0 & \text{for } y \in \mathbb{R}^n \setminus f(\mathbb{R}^n). \end{cases}$$

Remarks

- A straight-forward generalization of the univariate case.
- Proof omitted.

Theorem (The multivariate PDF transform for linear functions)

Let X be a random vector with PDF p_X . Let $Y = f(X)$ with

$$f(x) = Ax \text{ with nonsingular } A \in \mathbb{R}^{n \times n}. \quad (126)$$

Then the PDF of Y is given by

$$p_Y : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) = \frac{1}{|A|} p_X(A^{-1}y) \quad (127)$$

where $|A|$ and A^{-1} denote the determinant and the inverse of A , respectively.

Remarks

- A straight-forward application the multivariate PDF transform theorem.
- Central for classical and Bayesian linear Gaussian models.

The multivariate PDF transform

Proof

We first show that

$$f^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^n, y \mapsto f^{-1}(y) := A^{-1}y. \quad (128)$$

To this end, we note that

$$f^{-1}(f(x)) = A^{-1}Ax = x = \text{id}_{\mathbb{R}^n}(x). \quad (129)$$

We next show that

$$J^f(f^{-1}(y)) = A. \quad (130)$$

To this end, we first note that

$$f_i(x) = \sum_{j=1}^n a_{ij} x_j. \quad (131)$$

Thus

$$J^f(x) = \left(\frac{\partial}{\partial x_j} f_i(x) \right)_{i,j=1,\dots,n} \quad (132)$$

$$= \left(\sum_{j=1}^n \frac{\partial}{\partial x_j} a_{ij} x_j \right)_{i,j=1,\dots,n} \quad (133)$$

$$= (a_{ij})_{i,j=1,\dots,n} \quad (134)$$

$$= A \in \mathbb{R}^{n \times n}. \quad (135)$$

□

Transformations

- The probability integral transform
- The univariate PDF transform
- The multivariate PDF transform
- **Linear combinations**
- Gaussian transformations
- Exercises

Theorem (Linear combination of two continuous random variables)

Let X_1 and X_2 be two continuous random variables with joint PDF $p_{X_1, X_2}(x_1, x_2)$, and let

$$Y = a_1 X_1 + a_2 X_2 + b \text{ with } a_1 \neq 0. \quad (136)$$

Then Y has a continuous distribution with PDF

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X_1, X_2} \left(\frac{y - b - a_2 x_2}{a_1}, x_2 \right) \frac{1}{a_1} dx_2 \quad (137)$$

Remark

- The case $a_1 = a_2 = 1, b = 0$ is of special interest.

Linear combinations

Proof

(1) We first note that for any joint PDF p_X of a random vector X and any multivariate real-valued function f such that $Y := f(X)$, the CDF of Y takes on values

$$P_Y(y) = \int_{A_y} p_X(x) dx, \text{ where } A_y := \{x | f(x) \leq y\}, \quad (138)$$

because

$$P_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(f(X) \leq y) = \mathbb{P}(X \in \{x | f(x) \leq y\}) = \mathbb{P}(X \in A_y) = \int_{A_y} p_X(x) dx$$

(2) We next evaluate the CDF P_Y of Y of the linear combinations theorem in the form

$$P_Y(y) = \int_{-\infty}^y p_Y(v) dv, \quad (139)$$

from which the form of p_Y then follows directly. To this end, we define

$$A_y := \{(x_1, x_2) | a_1 x_1 + a_2 x_2 + b \leq y\} \text{ for all } y \in \mathbb{R}. \quad (140)$$

Then from (1), we have

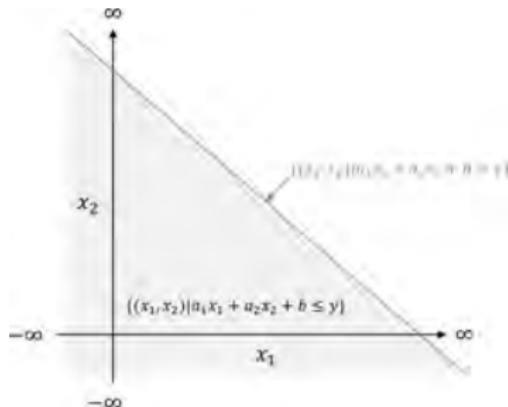
$$P_Y(y) = \iint_{A_y} p_{X_1, X_2}(x_1, x_2) dx_1 dx_2. \quad (141)$$

Linear combinations

Proof (cont.)

(3) To evaluate this integral, visualized below, we consider $-\infty < x_2 < \infty$ and for each x_2 integrate x_1 from $-\infty$ to

$$x_1 = \frac{y - a_2 x_2 - b}{a_1} \Leftrightarrow a_1 x_1 + a_2 x_2 + b = y. \quad (142)$$



We thus consider the integral

$$P_Y(y) = \iint_{A_y} p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{y-a_2 x_2 - b}{a_1}} p_{X_1, X_2}(x_1, x_2) dx_1 dx_2. \quad (143)$$

Linear combinations

Proof (cont.)

The inner integral on the right-hand side of the above can then be rewritten by means of the integration by substitution rule (also known as a “change of variable”) as

$$\int_{-\infty}^{\frac{y-a_2x_2-b}{a_1}} p_{X_1, X_2}(x_1, x_2) dx_1 = \int_{-\infty}^y p_{X_1, X_2}\left(\frac{\xi - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} d\xi \quad (144)$$

(see below for a detailed derivation). Substitution in the above then yields

$$\begin{aligned} P_Y(y) &= \int_{-\infty}^{\infty} \int_{-\infty}^y p_{X_1, X_2}\left(\frac{\xi - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} d\xi dx_2 \\ &= \int_{-\infty}^y \int_{-\infty}^{\infty} p_{X_1, X_2}\left(\frac{\xi - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} dx_2 d\xi \end{aligned} \quad (145)$$

But then it follows from basic calculus that

$$\begin{aligned} p_Y(y) &= \frac{d}{dy} P_Y(y) \\ &= \frac{d}{dy} \int_{-\infty}^y \int_{-\infty}^{\infty} p_{X_1, X_2}\left(\frac{\xi - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} dx_2 d\xi \\ &= \int_{-\infty}^{\infty} p_{X_1, X_2}\left(\frac{y - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} dx_2 \end{aligned} \quad (146)$$

Linear combinations

Proof (cont.)

(4) Finally, we show that

$$\int_{-\infty}^{\frac{y-a_2x_2-b}{a_1}} p_{X_1, X_2}(x_1, x_2) dx_1 = \int_{-\infty}^y p_{X_1, X_2}\left(\frac{\xi - b - a_2x_2}{a_1}, x_2\right) \frac{1}{a_1} d\xi \quad (147)$$

by means of the integration by substitution rule. To this end, we first recall that the integration by substitution rule state that for univariate real-valued functions g and h it holds that

$$\int_{h(a)}^{h(b)} g(x) dx = \int_a^b g(h(x))h'(x) dx. \quad (148)$$

For constant $x_2 \in \mathbb{R}$, we next define

$$g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto g(x) := p_{X_1, X_2}(x, x_2) \quad (149)$$

and

$$h : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto h(x) := \frac{1}{a_1}(x - a_2x_2 - b). \quad (150)$$

We note that the derivative of h at x evaluates to

$$h'(x) = \frac{1}{a_1}. \quad (151)$$

Finally, we set $b := y$ and $a := -\infty$.

Proof (cont.)

Substitution in (148) then yields

$$\begin{aligned}
 & \int_{h(a)}^{h(b)} g(x) dx = \int_a^b g(h(x))h'(x) dx \\
 \Leftrightarrow & \int_{h(-\infty)}^{h(y)} p_{X_1, X_2}(x, x_2) dx = \int_{-\infty}^y p_{X_1, X_2}(h(x), x_2) \frac{1}{a_1} dx. \\
 \Leftrightarrow & \int_{-\infty}^{\frac{y-a_2x_2-b}{a_1}} p_{X_1, X_2}(x, x_2) dx = \int_{-\infty}^y p_{X_1, X_2}\left(\frac{x-a_2x_2-b}{a_1}, x_2\right) \frac{1}{a_1} dx \\
 \Leftrightarrow & \int_{-\infty}^{\frac{y-a_2x_2-b}{a_1}} p_{X_1, X_2}(x_1, x_2) dx_1 = \int_{-\infty}^y p_{X_1, X_2}\left(\frac{\xi-a_2x_2-b}{a_1}, x_2\right) \frac{1}{a_1} d\xi.
 \end{aligned} \tag{152}$$

□

Theorem (Convolution of random variables)

Let X_1 and X_2 be two independent continuous random variables with marginal PDFs p_{X_1} and p_{X_2} , respectively, and let $Y := X_1 + X_2$. Then a PDF of the distribution of Y is given by the *convolution* of p_{X_1} and p_{X_2} , i.e.,

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X_1}(y - x_2)p_{X_2}(x_2) dx_2 = \int_{-\infty}^{\infty} p_{X_1}(x_1)p_{X_2}(y - x_1) dx_1 \quad (153)$$

Proof

We first note that for independent X_1, X_2 , p_{X_1, X_2} factorizes. Setting $a_1 = a_2 = 1$ and b in the Theorem on linear combinations of two continuous random variables then yields

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X_1}\left(\frac{y - 0 - 1x_2}{1}\right)p_{X_2}(x_2) dx_2 = \int_{-\infty}^{\infty} p_{X_1}(y - x_2)p_{X_2}(x_2) dx_2. \quad (154)$$

Finally, by setting $X_1 := X_2$ and $X_2 := X_1$, we obtain

$$p_Y(y) = \int_{-\infty}^{\infty} p_{X_2}(y - x_1)p_{X_1}(x_1) dx_1 = \int_{-\infty}^{\infty} p_{X_1}(x_1)p_{X_2}(y - x_1) dx_1. \quad (155)$$

A direct proof of the convolution formula can be given by considering the transformation $(X_1, X_2) \mapsto (X_1 + X_2, X_1)$ and marginalization (e.g. Casella and Berger (2012, Theorem 5.2.9)).

□

Transformations

- The probability integral transform
- The univariate PDF transform
- The multivariate PDF transform
- Linear combinations
- **Gaussian transformations**
- Exercises

A selection of Gaussian transformations

- The Z transform
- Bijective linear transformations
- Linear-affine transformations
- The General Linear Model
- The χ^2 distribution
- The t distribution

A selection of Gaussian transformations

- **The Z transform**
- Bijective linear transformations
- Linear-affine transformations
- The General Linear Model
- The χ^2 distribution
- The t distribution

Theorem (The Z transform)

Let $X \sim N(\mu, \sigma^2)$ and $Y = f(X)$ with $f(x) := \frac{x-\mu}{\sigma}$. Then $Y \sim N(0, 1)$.

Proof

We first note that $f^{-1}(y) = \sigma y + \mu$ and $f'(x) = \frac{1}{\sigma}$. With the univariate PDF transform for linear functions, we then have for the PDF of Y

$$\begin{aligned} p_Y(y) &= \frac{1}{|1/\sigma|} N\left(\sigma y + \mu; \mu, \sigma^2\right) \\ &= \frac{1}{1/\sqrt{\sigma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (\sigma y + \mu - \mu)^2\right) \\ &= \frac{\sqrt{\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sigma^2 y^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y^2\right) \\ &= N(y; 0, 1) \end{aligned} \tag{156}$$

and thus $Y \sim N(0, 1)$.

□

A selection of Gaussian transformations

- The Z transform
- **Bijective linear transformations**
- The General Linear Model
- The χ^2 distribution
- The t distribution

Theorem (Bijective linear transformations)

Let X denote random vector distributed according to an n -variate Gaussian distribution with probability density function $N(x; \mu_x, \Sigma_x)$ for $x, \mu_x \in \mathbb{R}^n, \Sigma_x \in \mathbb{R}^{n \times n}$ p.d., and let $A \in \mathbb{R}^{n \times n}$ be a matrix of full column-rank. Then the random vector

$$Y := AX \tag{157}$$

is distributed according to an n -variate Gaussian distribution with probability density function $N(y; \mu_y, \Sigma_y)$ with $y, \mu_y \in \mathbb{R}^m$ and $\Sigma_y \in \mathbb{R}^{n \times n}$, where

$$\mu_y := A\mu_x \text{ and } \Sigma_y := A\Sigma_x A^T. \tag{158}$$

The multivariate PDF transform

Proof

With the multivariate PDF transform theorem for linear functions, we first note that we have for the normalization term

$$|A|^{-1} (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} = (2\pi)^{-\frac{n}{2}} |A|^{-\frac{1}{2}} |\Sigma|^{-\frac{1}{2}} |A^T|^{-\frac{1}{2}} = (2\pi)^{-\frac{n}{2}} |A\Sigma A^T|^{-\frac{1}{2}} \quad (159)$$

We next note that for the argument of the exponential term, we have with $(A^{-1})^T = (A^T)^{-1}$

$$\begin{aligned} & -\frac{1}{2}(A^{-1}y - \mu)^T \Sigma^{-1}(A^{-1}y - \mu) \\ &= \dots \\ &= \dots \\ &= \dots \\ &= \dots \\ &= -\frac{1}{2}(y - A\mu)^T (A\Sigma A^T)^{-1}(y - A\mu) \end{aligned} \quad (160)$$

which completes the proof. \square

A selection of Gaussian transformations

- The Z transform
- Bijective linear transformations
- **Linear-affine transformations**
- The General Linear Model
- The χ^2 distribution
- The t distribution

Theorem (Linear transformations of multivariate Gaussian distributions)

Let $X \sim N(\mu, \Sigma)$ denote an n -dimensional Gaussian random vector and let

$$Y := AX + b \text{ with } A \in \mathbb{R}^{m \times n} \text{ and } \mathbb{R}^m. \quad (161)$$

Then

$$Y \sim N(A\mu + b, A\Sigma A^T) \quad (162)$$

Remark

- For a proof, see Anderson (2003, Section 2.4).

A selection of Gaussian transformations

- The Z transform
- Bijective linear transformations
- Linear-affine transformations
- **The General Linear Model**
- The χ^2 distribution
- The t distribution

The General Linear Model

- n data variables y_i , np predictor variables x_{ij} , p effect parameters, $\sigma^2 > 0$

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \text{ for } i = 1, \dots, n. \quad (163)$$

- $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p, \sigma^2 > 0$

$$y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n). \quad (164)$$

- Thus

$$y = f(\varepsilon) \text{ with } f : \mathbb{R}^n \rightarrow \mathbb{R}^n, \varepsilon \mapsto f(\varepsilon) := I_n \varepsilon + X\beta. \quad (165)$$

- Hence

$$y \sim N(I_n 0_n + X\beta, I_n \sigma^2 I_n I_n^T) \Rightarrow y \sim N(X\beta, \sigma^2 I_n). \quad (166)$$

A selection of Gaussian transformations

- The Z transform
- Bijective linear transformations
- Linear-affine transformations
- Linear-affine transformations
- The General Linear Model
- **The χ^2 distribution**
- The t distribution

Example (χ^2 random variable)

Let X be a continuous random variable with outcome set $\mathbb{R}_{>0}$ and PDF

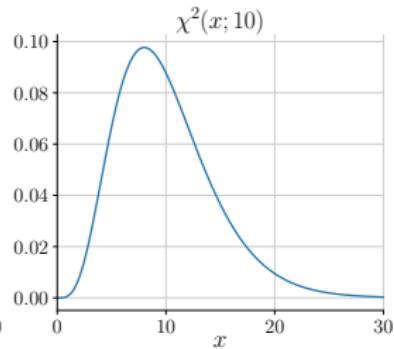
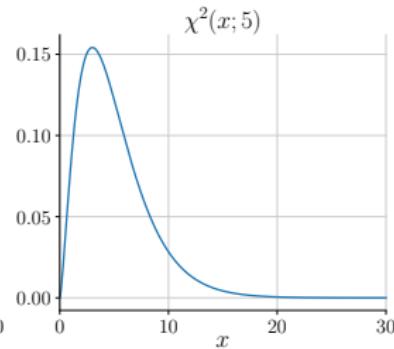
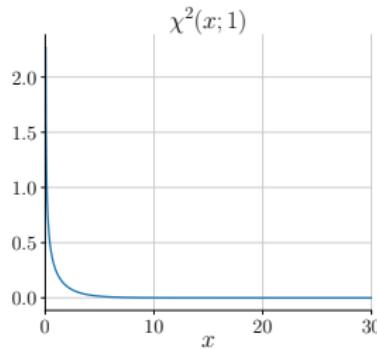
$$p : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, x \mapsto p(x) := \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} x^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}x\right), \quad (167)$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a χ^2 distribution with n degrees of freedom, for which we write $X \sim \chi^2(n)$. We abbreviate the PDF of a χ^2 random variable by

$$\chi^2(x; n) := \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} x^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}x\right). \quad (168)$$

Gaussian transformations

Example (χ^2 distributions)



Theorem (Sum of squared standard normal random variables)

For $i = 1, \dots, n$, let $X_i \sim N(0, 1)$ be independent standard normal random variables and let $Y := \sum_{i=1}^n X_i^2$. Then Y is a χ^2 random variable with n degrees of freedom.

Proof

We content with the case $n := 1$. To this end, we note that with the univariate PDF theorem for piecewise bijective transformations, the PDF of a random variable $Y := f(X)$ resulting from the transformation of a random variable X with PDF p_X by a piecewise differentiable and invertible function is given by

$$p_Y(y) = \sum_{i=1}^k 1_{\mathcal{Y}_i} \frac{1}{|f'_i(f_i^{-1}(y))|} p_X(f_i^{-1}(y)). \quad (169)$$

We next define

$$\mathcal{X}_1 :=]-\infty, 0[, \mathcal{X}_2 :=]0, \infty[, \text{ and } \mathcal{Y}_i := \mathbb{R}_{>0} \text{ for } i = 1, 2, \quad (170)$$

as well as

$$f_i : \mathcal{X}_i \rightarrow \mathcal{Y}_i, x \mapsto f_i(x) := x^2 =: y \text{ for } i = 1, 2. \quad (171)$$

The derivatives and inverse functions of f_i are then given by

Gaussian transformations

Proof (cont.)

$$f'_i : \mathcal{X}_i \rightarrow \mathcal{X}_i, x \mapsto f'_i(x) = 2x \text{ for } i = 1, 2, \quad (172)$$

and

$$f_1^{-1} : \mathcal{Y}_1 \rightarrow \mathcal{X}_1, y \mapsto f_1^{-1}(y) = -\sqrt{y} \text{ and } f_2^{-1} : \mathcal{Y}_2 \rightarrow \mathcal{X}_2, y \mapsto f_2^{-1}(y) = \sqrt{y}, \quad (173)$$

respectively. Substitution in eq. (169) then yields

$$\begin{aligned} p_Y(y) &= 1_{\mathcal{Y}_1}(y) \frac{1}{|f'_1(f_1^{-1}(y))|} p_X(f_1^{-1}(y)) + 1_{\mathcal{Y}_2}(y) \frac{1}{|f'_2(f_2^{-1}(y))|} p_X(f_2^{-1}(y)) \\ &= \frac{1}{|2(-\sqrt{y})|} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(-\sqrt{y})^2\right) + \frac{1}{|2(\sqrt{y})|} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{y})^2\right) \\ &= \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y\right) + \frac{1}{2\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} \exp\left(-\frac{1}{2}y\right). \end{aligned} \quad (174)$$

On the other hand, with $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, the PDF of a χ^2 random variable Y with $n = 1$ is given by

$$\frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{\frac{1}{2}}} y^{\frac{1}{2}-1} \exp\left(-\frac{1}{2}y\right) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} \exp\left(-\frac{1}{2}y\right) \quad (175)$$

Thus, if $X_i \sim N(0, 1)$, then $Y := X_i^2 \sim \chi^2(1)$. □

A selection of Gaussian transformations

- The Z transform
- Bijective linear transformations
- The General Linear Model
- The χ^2 distribution
- **The t distribution**

Example (t random variable)

Let X be random variable with outcome set \mathbb{R} and PDF

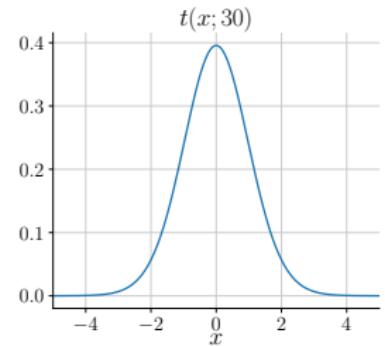
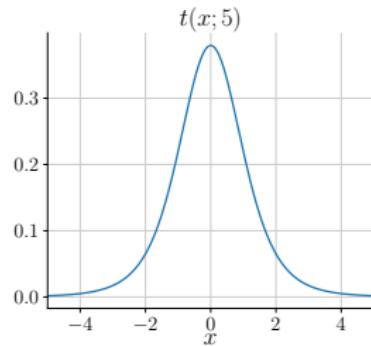
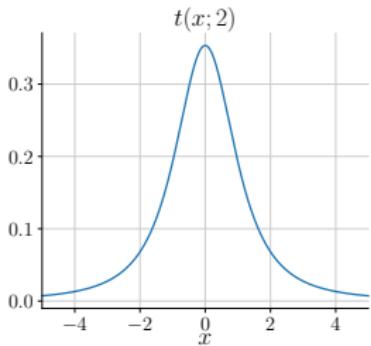
$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, t \mapsto p(t) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \quad (176)$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a t -distribution with n degrees of freedom, for which we write $X \sim t(n)$. We abbreviate the PDF of a t random variable by

$$t(x; n) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \quad (177)$$

Probability density functions

Example (t distributions)



Theorem (t distribution)

Let $Z \sim N(0, 1)$ be a standard normal random variable, let $V \sim \chi^2(n)$ be a χ^2 random variable with n degrees of freedom, and assume that Z and V are independent random variables. Then the random variable

$$T := \frac{Z}{\sqrt{V/n}} \tag{178}$$

is a t random variable with n degrees of freedom.

Remarks

- Maybe the most classical result of Frequentist statistics (Student, 1908).
- A historical perspective is given by (Zabell, 2008).

Gaussian transformations

Proof

We first note that the joint distribution of Z and V has PDF

$$p_{Z,V}(z, v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} v^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}v\right). \quad (179)$$

We next consider the transformation

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}^2, (z, v) \mapsto f(z, v) := \left(\frac{z}{\sqrt{v/n}}, v \right) =: (t, w) \quad (180)$$

and use the multivariate PDF transform theorem to derive the PDF of (t, w) . To this end, we first recall that if X is an n -dimensional random vector with PDF p_X and $Y := f(X)$ for differentiable and bijective $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, then the PDF of Y is given by

$$p_Y : \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}, y \mapsto p_Y(y) := \frac{1}{|J_f(f^{-1}(y))|} p_X(f^{-1}(y)) \quad (181)$$

For the current transformation f , we first note that

$$f^{-1} : \mathbb{R}^2 \rightarrow \mathbb{R}^2, (t, w) \mapsto f^{-1}(t, w) := \left(\sqrt{w/nt}, w \right), \quad (182)$$

because

$$f^{-1}(f(z, v)) = f^{-1}\left(\frac{z}{\sqrt{v/n}}, v\right) = \left(\frac{\sqrt{v/n}z}{\sqrt{v/n}}, v\right) = (z, v) \text{ for all } (z, v) \in \mathbb{R}^2. \quad (183)$$

Proof (cont.)

We next note that the determinant of the Jacobian matrix of f at (z, v) evaluates to

$$|J^f(z, v)| = \left| \begin{array}{cc} \frac{\partial}{\partial z} \left(\frac{z}{\sqrt{v/n}} \right) & \frac{\partial}{\partial v} \left(\frac{z}{\sqrt{v/n}} \right) \\ \frac{\partial}{\partial z} v & \frac{\partial}{\partial v} v \end{array} \right| = \left(\frac{v}{n} \right)^{-1/2}, \quad (184)$$

such that

$$\frac{1}{|J^f(f^{-1}(z, v))|} = \left(\frac{w}{n} \right)^{1/2}. \quad (185)$$

Substitution in (181) then yields

$$p_{T,W}(t, w) = \left(\frac{w}{n} \right)^{1/2} p_{Z,V} \left(\sqrt{w/nt}, w \right), \quad (186)$$

and thus

Gaussian transformations

Proof (cont.)

$$\begin{aligned} p_T(t) &= \int_0^\infty p_{T,W}(t,w) dw \\ &= \int_0^\infty \left(\frac{w}{n}\right)^{1/2} p_{Z,V}\left(\sqrt{w/nt}, w\right) dw \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\sqrt{w/nt})^2\right) \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} w^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}w\right) \left(\frac{w}{n}\right)^{1/2} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}} n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{w}{n}t^2\right) w^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}w\right) w^{1/2} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}} n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\frac{w}{n}t^2 - \frac{1}{2}w\right) w^{\frac{n}{2}-1} w^{\frac{1}{2}} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}} n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{w}{n}t^2 + w\right)\right) w^{\frac{n+1}{2}-1} dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}} n^{\frac{1}{2}}} \int_0^\infty \exp\left(-\frac{1}{2}\left(1 + \frac{t^2}{n}\right)\right) w^{\frac{n+1}{2}-1} dw \end{aligned} \tag{187}$$

Gaussian transformations

Proof (cont.)

We next note that the integrand on the left-hand side of the above corresponds to the kernel of a Gamma PDF with parameters $\alpha = \frac{n+1}{2}$ and $\beta = \frac{2}{1 + \frac{t^2}{n}}$.

Explicitly,

$$\begin{aligned}\Gamma(w; \alpha, \beta) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} w^{\alpha-1} \exp\left(-\frac{w}{\beta}\right) \\ &\Rightarrow \Gamma\left(w; \frac{n+1}{2}, \frac{2}{1 + \frac{t^2}{n}}\right) = \frac{1}{\Gamma\left(\frac{n+1}{2}\right) \left(\frac{2}{1 + \frac{t^2}{n}}\right)^{\frac{n+1}{2}}} w^{\frac{n+1}{2}-1} \exp\left(-\frac{w}{\frac{2}{1 + \frac{t^2}{n}}}\right) \\ &= \frac{1}{\Gamma\left(\frac{n+1}{2}\right) \left(\frac{2}{1 + \frac{t^2}{n}}\right)^{\frac{n+1}{2}}} \exp\left(-\frac{1}{2} \left(1 + \frac{t^2}{n}\right)\right) w^{\frac{n+1}{2}-1}.\end{aligned}$$

Gaussian transformations

Proof (cont.)

We thus have

$$p_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}} n^{\frac{1}{2}}} \int_0^\infty \Gamma\left(w; \frac{n+1}{2}, \frac{2}{1 + \frac{t^2}{n}}\right) dw \quad (188)$$

Finally, we note that the integral term of the above corresponds to the normalization term of the Gamma PDF. We thus have

$$p_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}} n^{\frac{1}{2}}} \Gamma\left(\frac{n+1}{2}\right) \left(\frac{2}{1 + \frac{t^2}{n}}\right)^{\frac{n+1}{2}} \quad (189)$$

which corresponds to the PDF of a T random variable.

A selection of Gaussian transformations

- The Z transform
- Bijective linear transformations
- The General Linear Model
- The χ^2 distribution
- **The t distribution**

Example (t random variable)

Let X be random variable with outcome set \mathbb{R} and PDF

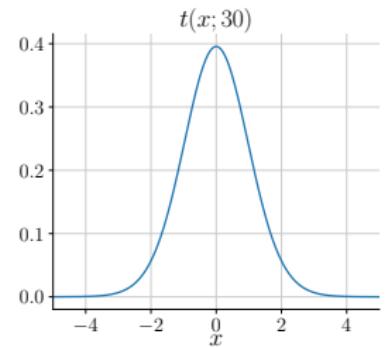
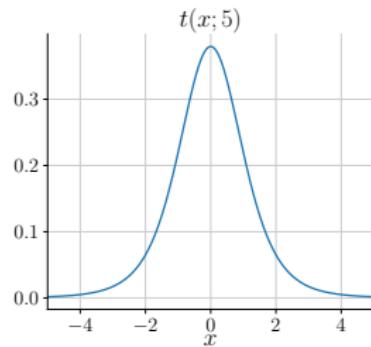
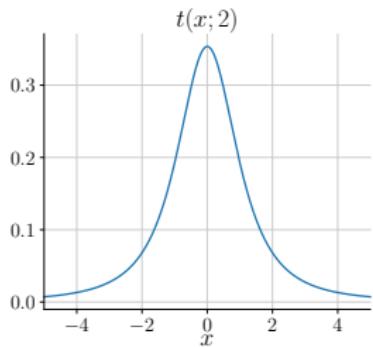
$$p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, t \mapsto p(t) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \quad (190)$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a t -distribution with n degrees of freedom, for which we write $X \sim t(n)$. We abbreviate the PDF of a t random variable by

$$t(x; n) := \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \quad (191)$$

Probability density functions

Example (t distributions)



Transformations

- The probability integral transform
- The univariate PDF transform
- The multivariate PDF transform
- Linear combinations
- Gaussian transformations
- **Exercises**

Study questions

1. Let X be a discrete random variable and $Y = f(X)$ a transformation of X . Express the PMF p_Y of Y in terms of the distribution of X .
2. Let X be a continuous random variable with PDF p_X and let $Y = f(X)$ be a transformation of X . Write down the direct calculation procedure to derive the PDF p_Y of Y .
3. Write down the probability integral transform theorem.
4. How can a uniform random number generator be used to create random numbers with arbitrary distribution?
5. Write down the univariate probability density function transform theorem.
6. Write down the univariate probability density function transform theorem for linear functions.
7. Write down the Z-transformation of a univariate Gaussian random variable.
8. Write down the multivariate probability density function transform theorem.
9. Write down the multivariate probability density function transform theorem for linear functions.
10. Let an n -dimensional random vector X be distributed according to a multivariate Gaussian distribution, $X \sim N(\mu_x, \Sigma_x)$. Let $A \in \mathbb{R}^{n \times n}$ be a matrix of full column-rank. How is the random vector $Y := AX$ distributed?

Theoretical exercises

1. Let two random variables X_1 and X_2 have a joint PDF p_{X_1, X_2} and let

$$Y := a_1 X_1 + a_2 X_2 + b \text{ with } a_1 \neq 0. \quad (192)$$

Show that the Y has a continuous distribution with PDF

$$p_Y(y) := \int_{-\infty}^{\infty} p_{X_1, X_2} \left(\frac{y - b - a_2 x_2}{a_1}, x_2 \right) \frac{1}{|a_1|} dx_2 \quad (193)$$

(DeGroot and Schervish, 2012, Theorem 3.9.4). In addition, write down p_Y for X_1, X_2 independent, $a_1 := a_2 := 1$ and $b := 0$ (c.f. DeGroot and Schervish, 2012, Definition 3.9.1).

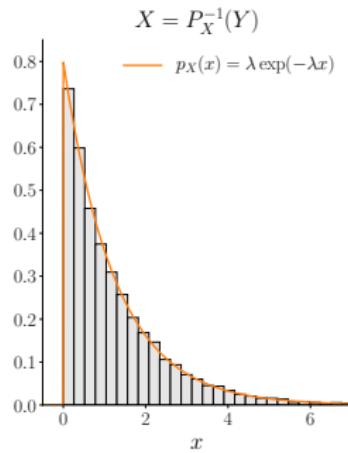
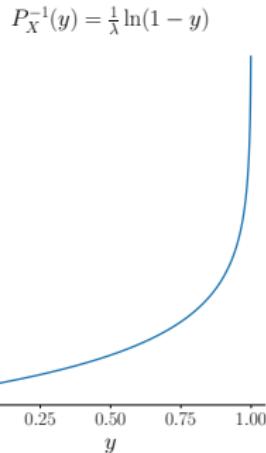
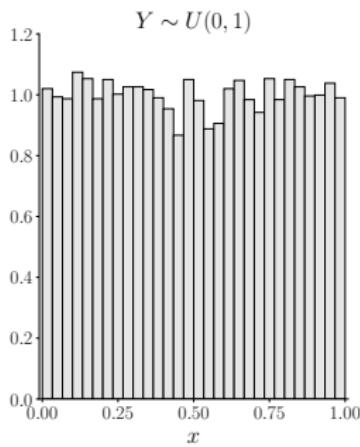
2. Derive the functional form of the χ^2 -distribution PDF (Casella and Berger, 2012, p. 52-53).
3. Derive the functional form of the t -distribution PDF (Casella and Berger, 2012, p. 223 - 224).

Programming exercises

1. Write a program that generates pseudo-random numbers from an exponential distribution using a uniform pseudo-random number generator and the probability integral transform theorem.
2. Let $X \sim N(0, 1)$ and let $Y = \exp(X)$. Evaluate the PDF of Y analytically and verify your evaluation using a simulation based on drawing random numbers from $N(0, 1)$.
3. Let $X \sim N(0, 1)$ and let $Y = X^2$. By simulation, validate that Y is distributed according to a chi-squared distribution with one degree of freedom. Next, let $X_1, \dots, X_{10} \sim N(0, 1)$ and let $Y = \sum_{i=1}^{10} X_i^2$. By simulation, validate that Y is distributed according to a chi-squared distribution with ten degrees of freedom.

Exercises

Programming Exercise 1



Programming Exercise 2

We first note that for

$$f : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto f(x) := \exp(x) \quad (194)$$

we have

$$f' : \mathbb{R} \rightarrow \mathbb{R}_{>0}, x \mapsto f'(x) := \exp(x) \quad (195)$$

and

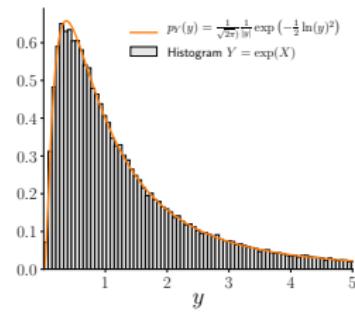
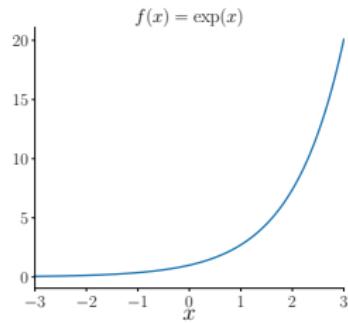
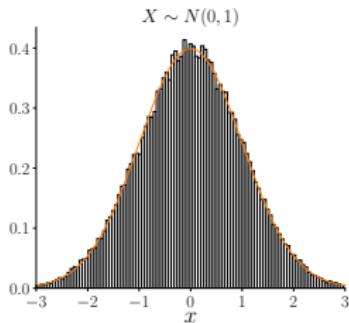
$$f^{-1} : \mathbb{R}_{>0} \rightarrow \mathbb{R}, y \mapsto f^{-1}(y) := \ln(y). \quad (196)$$

Substitution in the univariate PDF transform for bijective transformations then yields

$$p_Y(y) = \frac{1}{|\exp(\ln(y))|} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \ln(y)^2\right) = \frac{1}{\sqrt{2\pi}} \frac{1}{|y|} \exp\left(-\frac{1}{2} \ln(y)^2\right) \quad (197)$$

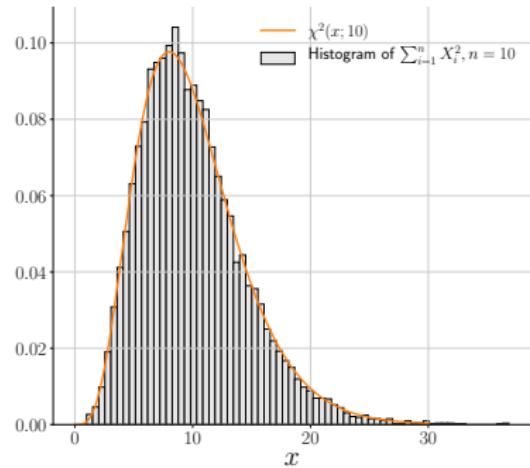
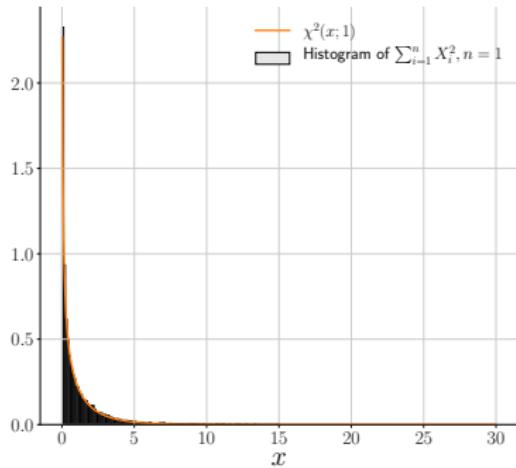
Exercises

Programming Exercise 2



Exercises

Programming Exercise 3



(6) Expectation and covariance

Bibliographic remarks

The treatment follows Wasserman (2004, Sections 3.1 - 3.3) and DeGroot and Schervish (2012, Sections 4.1 - 4.3, 4.6).

Expectation and covariance

- Expectation
- Variance and standard deviation
- Sample mean, sample variance, sample standard deviation
- Covariance and correlation
- Sample covariance and sample correlation
- Exercises

Expectation and covariance

- **Expectation**
- Variance and standard deviation
- Sample mean, sample variance, sample standard deviation
- Covariance and correlation
- Sample covariance and sample correlation
- Exercises

Definition (Expectation)

Let $(\Omega, \mathcal{A}, \mathbb{P})$ denote a probability space and let X denote a random variable. The *expectation* (or *expected value*) of X is defined as

- $\mathbb{E}(X) := \sum_{x \in \mathcal{X}} x p_X(x)$, if $X : \Omega \rightarrow \mathcal{X}$ is discrete with PMF p_X , and as
- $\mathbb{E}(X) := \int_{-\infty}^{\infty} x p_X(x) dx$, if $X : \Omega \rightarrow \mathbb{R}$ is continuous with PDF p_X .

The expectation of a random variable is said to exist, if it is finite.

Remarks

- The expectation is a one-number summary of a distribution.
- Intuitively, $\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^n X_i$ for a large number n of i.i.d. draws X_i .

Example (Bernoulli variable expectation)

Let $X \sim \text{Bern}(\mu)$. Then $\mathbb{E}(X) = \mu$.

Proof

X is discrete with $\mathcal{X} = \{0, 1\}$. Thus,

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x \in \{0, 1\}} x \text{Bern}(x; \mu) \\ &= 0 \cdot \mu^0 (1 - \mu)^{1-0} + 1 \cdot \mu^1 (1 - \mu)^{1-1} \\ &= 1 \cdot \mu^1 (1 - \mu)^0 \\ &= \mu.\end{aligned}\tag{198}$$

□

Expectation

Example (Gaussian variable expectation)

Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{E}(X) = \mu$.

Proof

We first note without proof that

$$\int_{-\infty}^{\infty} \exp(-x^2) dx = \sqrt{\pi}. \quad (199)$$

From the definition of the expectation for continuous random variables, have

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx. \quad (200)$$

With the integration by substitution rule

$$\int_{g(a)}^{g(b)} f(x) dx = \int_a^b f(g(x))g'(x) dx \quad (201)$$

and the definition of

$$g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto g(x) := \sqrt{2\sigma^2}x + \mu \text{ with } g'(x) = \sqrt{2\sigma^2}, \quad (202)$$

we then have

Example (Gaussian variable expectation)

Proof (cont.)

$$\begin{aligned}
 \mathbb{E}(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu) \exp\left(-\frac{1}{2\sigma^2}((\sqrt{2\sigma^2}x + \mu) - \mu)^2\right) \sqrt{2\sigma^2} dx \\
 &= \frac{\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu) \exp\left(-x^2\right) dx \\
 &= \frac{1}{\sqrt{\pi}} \left(\sqrt{2\sigma^2} \int_{-\infty}^{\infty} x \exp\left(-x^2\right) dx + \mu \int_{-\infty}^{\infty} \exp\left(-x^2\right) dx \right) \\
 &= \frac{1}{\sqrt{\pi}} \left(\sqrt{2\sigma^2} \int_{-\infty}^{\infty} x \exp\left(-x^2\right) dx + \mu\sqrt{\pi} \right)
 \end{aligned} \tag{203}$$

An anti-derivative of $x \exp(-x^2)$ is given by $-\frac{1}{2} \exp(-x^2)$. With $\lim_{x \rightarrow -\infty} \exp(-x^2) = 0$ and $\lim_{x \rightarrow \infty} \exp(-x^2) = 0$ the remaining integral term thus vanishes and we obtain

$$\mathbb{E}(X) = \frac{1}{\sqrt{\pi}} (\mu\sqrt{\pi}) = \mu. \tag{204}$$

□

Expectation

Theorem (Properties of expectations)

- (1) (Linear-affine transformations) Let X denote a random variable, let $a, b \in \mathbb{R}$, and let $Y := aX + b$. Then

$$\mathbb{E}(Y) = a\mathbb{E}(X) + b. \quad (205)$$

- (2) (Linear combinations). Let X_1, \dots, X_n denote random variables and let $a_1, \dots, a_n \in \mathbb{R}$. Then

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i). \quad (206)$$

- (3) (Factorization under independence). Let X_1, \dots, X_n denote independent random variables. Then

$$\mathbb{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbb{E}(X_i). \quad (207)$$

Remark

- These properties are often helpful when evaluating expectations.

Expectation

Proof (\rightarrow Übung)

(1) follows directly with the linearity properties of sums and expectations. We consider the case of a continuous random variable X with PDF p_X in more detail. In this case, we have

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(aX + b) \\ &= \int (ax + b)p_X(x) dx \\ &= \int ap_X(x)x + bp_X(x) dx \\ &= a \int p_X(x)x dx + b \int p_X(x) dx \\ &= a\mathbb{E}(X) + b.\end{aligned}\tag{208}$$

(2) follows directly with the linearity properties of sums and expectations. We consider the case of two continuous random variables X_1 and X_2 with bivariate PDF p_{X_1, X_2} in more detail. In this case, we have

Expectation

Proof (cont.)

$$\begin{aligned} & \mathbb{E} \left(\sum_{i=1}^2 a_i X_i \right) \\ &= \mathbb{E}(a_1 X_1 + a_2 X_2) \\ &= \iint (a_1 x_1 + a_2 x_2) p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \\ &= \iint a_1 x_1 p_{X_1, X_2}(x_1, x_2) + a_2 x_2 p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \\ &= a_1 \iint x_1 p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 + a_2 \iint x_2 p_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \quad (209) \\ &= a_1 \int x_1 \left(\int p_{X_1, X_2}(x_1, x_2) dx_2 \right) dx_1 + a_2 \int x_2 \left(\int p_{X_1, X_2}(x_1, x_2) dx_1 \right) dx_2 \\ &= a_1 \int x_1 p_{X_1}(x_1) dx_1 + a_2 \int x_2 p_{X_2}(x_2) dx_2 \\ &= a_1 \mathbb{E}(X_1) + a_2 \mathbb{E}(X_2) \\ &= \sum_{i=1}^2 a_i \mathbb{E}(X_i). \end{aligned}$$

Finally, an induction argument can be used to generalize the bivariate to the n -variate case.

Expectation

Proof (cont.)

(3) We consider the case of n continuous random variables with joint PDF p_{X_1, \dots, X_n} . Because X_1, \dots, X_n are independent, it holds that

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i). \quad (210)$$

We thus have

$$\begin{aligned} \mathbb{E}\left(\prod_{i=1}^n x_i\right) &= \int \cdots \int \left(\prod_{i=1}^n X_i\right) p_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \dots dx_n \\ &= \int \cdots \int \prod_{i=1}^n x_i \prod_{i=1}^n p_{X_i}(x_i) dx_1 \dots dx_n \\ &= \int \cdots \int \prod_{i=1}^n x_i p_{X_i}(x_i) dx_1 \dots dx_n \\ &= \prod_{i=1}^n \int x_i p_{X_i}(x_i) dx_i \\ &= \prod_{i=1}^n \mathbb{E}(X_i). \end{aligned} \quad (211)$$

□

Expectation and covariance

- Expectation
- **Variance and standard deviation**
- Sample mean, sample variance, sample standard deviation
- Covariance and correlation
- Sample covariance and sample correlation
- Exercises

Definition (Variance and standard deviation)

Let X be a random variable with expectation $\mathbb{E}(X)$. The variance of X is defined as

$$\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}(X))^2), \quad (212)$$

assuming that this expectation exists. The standard deviation is defined as

$$\mathbb{S}(X) := \sqrt{\mathbb{V}(X)}. \quad (213)$$

Remarks

- The variance measures the spread of a distribution.
- The square is necessitated by $\mathbb{E}(X - \mathbb{E}(X)) = \mathbb{E}(X) - \mathbb{E}(X) = 0$.
- An alternative measure of distribution spread is $\mathbb{E}(|X - \mathbb{E}(X)|)$.
- Another alternative measure is the entropy of a distribution.

Variance and standard deviation

Example (Bernoulli variable variance)

Let $X \sim \text{Bern}(\mu)$. Then the variance of X is

$$\mathbb{V}(X) = \mu(1 - \mu). \quad (214)$$

Proof

X is discrete and we have $\mathbb{E}(X) = \mu$. Thus

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}((X - \mu)^2) \\&= \sum_{x \in \{0,1\}} (x - \mu)^2 \text{Bern}(x; \mu) \\&= (0 - \mu)^2 \mu^0 (1 - \mu)^{1-0} + (1 - \mu)^2 \mu^1 (1 - \mu)^{1-1} \\&= \mu^2 (1 - \mu) + (1 - \mu)^2 \mu \\&= (\mu^2 + (1 - \mu)\mu) (1 - \mu) \\&= (\mu^2 + \mu - \mu^2) (1 - \mu) \\&= \mu(1 - \mu).\end{aligned} \quad (215)$$

□

Theorem (Variance translation theorem)

Let X be a random variable. Then

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2. \quad (216)$$

Proof

With the definition of the variance and the linearity of expectations, we have

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2.\end{aligned} \quad (217)$$

□

Remark

- The theorem is useful, if computing $\mathbb{E}(X^2)$ and $\mathbb{E}(X)$ is easy.

Variance and standard deviation

Example (Gaussian variable variance)

Let $X \sim N(\mu, \sigma^2)$. Then $\mathbb{V}(X) = \sigma^2$.

Proof

We first note that the with the variance translation theorem

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx - \mu^2 \quad (218)$$

With the integration by substitution rule

$$\int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(x) dx \quad (219)$$

and the definition of

$$g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \sqrt{2\sigma^2}x + \mu, g(-\infty) := -\infty, g(\infty) := \infty, \text{ with } g'(x) = \sqrt{2\sigma^2}, \quad (220)$$

the integral term on the right-hand side of eq. (218) can be rewritten as

Variance and standard deviation

Proof (cont.)

$$\begin{aligned} & \int_{-\infty}^{\infty} x^2 \exp \left(-\frac{1}{2\sigma^2} (x - \mu)^2 \right) dx \\ &= \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu)^2 \exp \left(-\frac{1}{2\sigma^2} ((\sqrt{2\sigma^2}x + \mu) - \mu)^2 \right) \sqrt{2\sigma^2} dx \\ &= \sqrt{2\sigma^2} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu)^2 \exp \left(-\frac{2\sigma^2 x^2}{2\sigma^2} \right) dx \\ &= \sqrt{2\sigma^2} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu)^2 \exp (-x^2) dx. \end{aligned} \tag{221}$$

We thus have

$$\begin{aligned} \mathbb{V}(X) &= \frac{\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x + \mu)^2 \exp (-x^2) dx - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\sqrt{2\sigma^2}x)^2 + 2\sqrt{2\sigma^2}x\mu + \mu^2 \exp (-x^2) dx - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} x^2 \exp (-x^2) dx + 2\sqrt{2\sigma^2}\mu \int_{-\infty}^{\infty} x \exp (-x^2) dx + \mu^2 \int_{-\infty}^{\infty} \exp (-x^2) dx \right) - \mu^2 \end{aligned} \tag{222}$$

Variance and standard deviation

Proof (cont.)

Taking

$$\int_{-\infty}^{\infty} x \exp(-x^2) dx = 0 \text{ and } \int_{-\infty}^{\infty} \exp(-x^2) dx = \sqrt{\pi} \quad (223)$$

as given, we then obtain

$$\begin{aligned} \mathbb{V}(X) &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} x^2 \exp(-x^2) dx + \mu^2 \sqrt{\pi} \right) - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2 \exp(-x^2) dx + \mu^2 - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2 \exp(-x^2) dx \end{aligned} \quad (224)$$

With the integration by parts rule

$$\int_a^b f'(x)g(x) dx = f(x)g(x)|_a^b - \int_a^b f(x)g'(x) dx \quad (225)$$

and the definition of

$$f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto f(x) := \exp(-x^2) \text{ with } f'(x) = -2 \exp(-x^2) \quad (226)$$

and

$$g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto g(x) := -\frac{1}{2}x \text{ with } g'(x) = -\frac{1}{2} \quad (227)$$

□

Variance and standard deviation

Proof (cont.)

such that

$$f'(x)g(x) = -2 \exp(-x^2) \left(-\frac{1}{2}x \right) = x^2 \exp(-x^2), \quad (228)$$

we then have

$$\begin{aligned} \mathbb{V}(X) &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} x^2 \exp(-x^2) dx \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \left(-\frac{1}{2}x \exp(-x^2) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \exp(-x^2) \left(-\frac{1}{2} \right) dx \right) \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \left(-\frac{1}{2}x \exp(-x^2) \Big|_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \exp(-x^2) dx \right), \end{aligned} \quad (229)$$

From $\lim_{x \rightarrow \pm\infty} \exp(-x^2) = 0$, we infer that the first term in the bracketed term on the right-hand side of the above evaluates to 0, such that we obtain

$$\mathbb{V}(X) = \frac{2\sigma^2}{\sqrt{\pi}} \left(\frac{1}{2} \int_{-\infty}^{\infty} \exp(-x^2) dx \right) = \frac{\sigma^2}{\sqrt{\pi}} \sqrt{\pi} = \sigma^2. \quad (230)$$

□

Theorem (Variance properties)

(1) (Linear-affine transformations.) Let X be a random variable, let $a, b \in \mathbb{R}$, and let $Y := aX + b$. Then

$$\mathbb{V}(Y) = a^2\mathbb{V}(X) \text{ and } \mathbb{S}(Y) = |a|\mathbb{S}(X) \quad (231)$$

(2) (Linear combinations of independent random variables.) Let X_1, \dots, X_n denote independent random variables and let $a_1, \dots, a_n \in \mathbb{R}$. Then

$$\mathbb{V} \left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i). \quad (232)$$

Variance and standard deviation

Proof (\rightarrow Übung)

(1) We first note that $\mathbb{E}(Y) = a\mathbb{E}(X) + b$. For the variance of Y , we thus have

$$\begin{aligned}\mathbb{V}(Y) &= \mathbb{E}((Y - \mathbb{E}(Y))^2) \\ &= \mathbb{E}((aX + b - a\mathbb{E}(X) - b)^2) \\ &= \mathbb{E}((aX - a\mathbb{E}(X))^2) \\ &= \mathbb{E}((a(X - \mathbb{E}(X)))^2) \\ &= \mathbb{E}(a^2(X - \mathbb{E}(X))^2) \\ &= a^2\mathbb{E}((X - \mathbb{E}(X))^2) \\ &= a^2\mathbb{V}(X)\end{aligned}\tag{233}$$

Taking the square root then yields the result for the standard deviation.

(2) We consider the case of two independent random variables X_1 and X_2 in more detail. We first note that in this case, we have

$$\mathbb{E}(a_1X_1 + a_2X_2) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2).\tag{234}$$

We thus have

Proof (cont.)

$$\begin{aligned} & \mathbb{V} \left(\sum_{i=1}^2 a_i X_i \right) \\ &= \mathbb{V}(a_1 X_1 + a_2 X_2) \\ &= \mathbb{E} \left((a_1 X_1 + a_2 X_2 - \mathbb{E}(a_1 X_1 + a_2 X_2))^2 \right) \\ &= \mathbb{E} \left((a_1 X_1 + a_2 X_2 - a_1 \mathbb{E}(X_1) - a_2 \mathbb{E}(X_2))^2 \right) \\ &= \mathbb{E} \left((a_1 X_1 - a_1 \mathbb{E}(X_1) + a_2 X_2 - a_2 \mathbb{E}(X_2))^2 \right) \\ &= \mathbb{E} \left(((a_1(X_1 - \mathbb{E}(X_1)) + (a_2(X_2 - \mathbb{E}(X_2)))^2 \right) \\ &= \mathbb{E} \left((a_1(X_1 - \mathbb{E}(X_1)))^2 - 2(a_1(X_1 - \mathbb{E}(X_1))(a_2(X_2 - \mathbb{E}(X_2))) + (a_2(X_2 - \mathbb{E}(X_2)))^2 \right) \\ &= \mathbb{E} \left((a_1^2(X_1 - \mathbb{E}(X_1))^2 - 2a_1 a_2 (X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) + a_2^2(X_2 - \mathbb{E}(X_2))^2 \right) \\ &= a_1^2 \mathbb{E} \left((X_1 - \mathbb{E}(X_1))^2 \right) - 2a_1 a_2 \mathbb{E} \left((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) \right) + a_2^2 \mathbb{E} \left((X_2 - \mathbb{E}(X_2))^2 \right) \\ &= a_1^2 \mathbb{V}(X_1) - 2a_1 a_2 \mathbb{E} \left((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) \right) + a_2^2 \mathbb{V}(X_2) \\ &= \sum_{i=1}^2 a_i^2 \mathbb{V}(X_i) - 2a_1 a_2 \mathbb{E} \left((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)) \right) \end{aligned}$$

Variance and standard deviation

Proof (cont.)

Because X_1 and X_2 are independent, we have with the properties of expectations for independent random variables that

$$\begin{aligned}\mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) &= \mathbb{E}((X_1 - \mathbb{E}(X_1)))\mathbb{E}((X_2 - \mathbb{E}(X_2))) \\ &= (\mathbb{E}(X_1) - \mathbb{E}(X_1))(\mathbb{E}(X_2) - \mathbb{E}(X_2)) \\ &= 0\end{aligned}\tag{235}$$

and thus

$$\mathbb{V}\left(\sum_{i=1}^2 a_i X_i\right) = \sum_{i=1}^2 a_i^2 \mathbb{V}(X_i).\tag{236}$$

Finally, an induction argument can be used to generalize the bivariate to the n -variate case.

□

Expectation and covariance

- Expectation
- Variance and standard deviation
- **Sample mean, sample variance, sample standard deviation**
- Covariance and correlation
- Sample covariance and sample correlation
- Exercises

Sample mean, sample variance, sample standard deviation

Definition (Sample mean, sample variance, sample standard deviation)

Let X_1, \dots, X_n denote random variables. Then

- the *sample mean* of X_1, \dots, X_n is defined as the arithmetic average of X_1, \dots, X_n ,

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i, \quad (237)$$

- the *sample variance* of X_1, \dots, X_n is defined as

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad (238)$$

- the *sample standard deviation* is defined as

$$S_n := \sqrt{S_n^2}. \quad (239)$$

Remarks

- $\mathbb{E}(X)$, $\mathbb{V}(X)$, and $\mathbb{S}(X)$ are summaries of a random variable X .
- \bar{X}_n , S_n^2 , and S_n are summaries of the random sample X_1, \dots, X_n .
- \bar{X}_n , S_n^2 , and S_n are random variables, their realizations are denoted by \bar{x}_n , s_n^2 , and s_n .

Example (Sample mean, sample variance, sample standard deviation)

- Let $X_1, \dots, X_{10} \sim N(1, 2)$.
- Assume the following realizations

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.54	1.01	-3.28	0.35	2.75	-0.51	2.32	1.49	0.96	1.25

- Sample mean realization

$$\bar{x}_{10} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{6.88}{10} = 0.68. \quad (240)$$

- Sample variance realization

$$s_{10}^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x}_{10})^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - 0.68)^2 = \frac{25.37}{9} = 2.82. \quad (241)$$

- Sample standard deviation realization

$$s_{10} = \sqrt{s_{10}^2} = \sqrt{2.82} = 1.68. \quad (242)$$

Expectation and covariance

- Expectation
- Variance and standard deviation
- Sample mean, sample variance, sample standard deviation
- **Covariance and correlation**
- Sample covariance and sample correlation
- Exercises

Definition (Covariance and correlation coefficient)

The *covariance* of two random variables X and Y with finite expectations is defined as

$$\mathbb{C}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \quad (243)$$

if this expectation exists. The *correlation* of two random variables X and Y with finite expectations is defined as

$$\rho(X, Y) := \frac{\mathbb{C}(X, Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}} = \frac{\mathbb{C}(X, Y)}{\mathbb{S}(X)\mathbb{S}(Y)}. \quad (244)$$

Remarks

- The covariance of X with itself is the variance of X .
- $\rho(X, Y)$ is also called the *correlation coefficient* of X and Y .
- If $\rho(X, Y) = 0$, X and Y are called *uncorrelated*.
- We show $-1 \leq \rho(X, Y) \leq 1$ in Lecture 7 using the Cauchy-Schwarz inequality.

Example (Covariance and correlation of two discrete random variables)

Let X_1 and X_2 be two discrete random variables with joint PMF p_{X_1, X_2} given by

$p_{X_1, X_2}(x_1, x_2)$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$p_{X_1}(x_1)$
$x_1 = 1$	0.10	0.05	0.15	0.30
$x_1 = 2$	0.60	0.05	0.05	0.70
$p_{X_2}(x_2)$	0.70	0.10	0.20	

(cf. Lange and Mosler (2017, B 4.9)). To evaluate $\mathbb{C}(X_1, X_2)$ and $\rho(X_1, X_2)$, we first note that

$$\mathbb{E}(X_1) = \sum_{x_1=1}^2 x_1 p_{X_1}(x_1) = 1 \cdot 0.3 + 2 \cdot 0.7 = 1.7 \quad (245)$$

and

$$\mathbb{E}(X_2) = \sum_{x_2=1}^3 x_2 p_{X_2}(x_2) = 1 \cdot 0.7 + 2 \cdot 0.1 + 3 \cdot 0.2 = 1.5. \quad (246)$$

With the definition of the covariance of X_1 and X_2 , we then have

Covariance and correlation

$$\begin{aligned}\mathbb{C}(X_1, X_2) &= \mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))) \\ &= \sum_{x_1=1}^2 \sum_{x_2=1}^3 (x_1 - \mathbb{E}(X_1))(x_2 - \mathbb{E}(X_2)) p_{X_1, X_2}(x_1, x_2) \\ &= \sum_{x_1=1}^2 \sum_{x_2=1}^3 (x_1 - 1.7)(x_2 - 1.5) p_{X_1, X_2}(x_1, x_2) \\ &= \sum_{x_1=1}^2 (x_1 - 1.7)(1 - 1.5) p_{X_1, X_2}(x_1, 1) \\ &\quad + (x_1 - 1.7)(2 - 1.5) p_{X_1, X_2}(x_1, 2) \\ &\quad + (x_1 - 1.7)(3 - 1.5) p_{X_1, X_2}(x_1, 3) \\ &= (1 - 1.7)(1 - 1.5) p_{X_1, X_2}(1, 1) + (1 - 1.7)(2 - 1.5) p_{X_1, X_2}(1, 2) + (1 - 1.7)(3 - 1.5) p_{X_1, X_2}(1, 3) \\ &\quad + (2 - 1.7)(1 - 1.5) p_{X_1, X_2}(2, 1) + (2 - 1.7)(2 - 1.5) p_{X_1, X_2}(2, 2) + (2 - 1.7)(3 - 1.5) p_{X_1, X_2}(2, 3) \\ &= (-0.7) \cdot (-0.5) \cdot 0.10 &+ (-0.7) \cdot 0.5 \cdot 0.05 &+ (-0.7) \cdot 1.5 \cdot 0.15 \\ &\quad + 0.3 \cdot (-0.5) \cdot 0.60 &+ 0.3 \cdot 0.5 \cdot 0.05 &+ 0.3 \cdot 1.5 \cdot 0.05 \\ &= 0.035 - 0.0175 - 0.1575 - 0.09 + 0.0075 + 0.0225 \\ &= -0.2\end{aligned}$$

Theorem (Correlation of bivariate Gaussian random variables)

Let $(X, Y) \sim N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \sigma_{XX}^2 & \sigma_{XY}^2 \\ \sigma_{YX}^2 & \sigma_{YY}^2 \end{pmatrix}. \quad (247)$$

Then

$$\mathbb{C}(X, Y) = \sigma_{XY}^2 = \sigma_{YX}^2 \text{ and } \rho(X, Y) = \frac{\sigma_{XY}^2}{\sigma_{XX}\sigma_{YY}}. \quad (248)$$

Proof

For a proof, see Casella and Berger (2012, pp.175).

Theorem (Covariance translation theorem)

Let X and Y denote two random variables. Then

$$\mathbb{C}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \quad (249)$$

Proof

With the definition of the covariance, we have

$$\begin{aligned}\mathbb{C}(X, Y) &= \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\ &= \mathbb{E}(XY - X\mathbb{E}(Y) + \mathbb{E}(X)Y - \mathbb{E}(X)\mathbb{E}(Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).\end{aligned} \quad (250)$$

□

Remarks

- The theorem is useful, if computing $\mathbb{E}(XY)$, $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ is easy.
- For $Y = X$, we recover $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Theorem (Covariance, correlation and independence)

Let X and Y denote two random variables. If X and Y are independent random variables, then $\mathbb{C}(X, Y) = 0$ and X and Y are uncorrelated. Conversely, if $\mathbb{C}(X, Y) = 0$ and hence X and Y are uncorrelated, then X and Y are not necessarily independent.

Proof

(1) We first show that the independence of X and Y implies that their covariance is zero. To this end, we note that for independent random variables, we have

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y). \quad (251)$$

With the covariance translation theorem, it then follows that

$$\mathbb{C}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0. \quad (252)$$

With the definition of the correlation coefficient, it follows immediately that $\rho(X, Y) = 0$ and thus that X and Y are uncorrelated.

Covariance and correlation

Proof (cont.)

(2) We next show by example that the covariance of non-independent random variables X and Y can be zero. To this end, we consider the case of two discrete random variables X and Y with outcome spaces $\mathcal{X} = \{-1, 0, 1\}$ and $\mathcal{Y} = \{0, 1\}$, marginal PMF of X given by $p_X(X = x) = 1/3$ for $x \in \mathcal{X}$ and the definition $Y := X^2$. We first note that

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} x p_X(X = x) = -1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0 \quad (253)$$

and

$$\mathbb{E}(XY) = \mathbb{E}(XX^2) = \mathbb{E}(X^3) = \sum_{x \in \mathcal{X}} x^3 p_X(X = x) = -1^3 \cdot \frac{1}{3} + 0^3 \cdot \frac{1}{3} + 1^3 \cdot \frac{1}{3} = 0. \quad (254)$$

With the covariance translation theorem, we thus have

$$\mathbb{C}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) - \mathbb{E}(X)\mathbb{E}(Y) = 0 - 0 \cdot \mathbb{E}(Y) = 0. \quad (255)$$

The covariance of X and Y is thus zero. However, as shown below, the joint PMF of X and Y does not factorize, and thus X and Y are not independent.

Covariance and correlation

Proof (cont.)

The definition of $Y := X^2$ entails the following conditional PMF $p_{Y|X}$:

		$p_{Y X}(y x)$		
		$x = -1$	$x = 0$	$x = 1$
$y = 0$	0	1	0	
	1	0	1	

The marginal PMF of p_X and the conditional PMF $p_{Y|X}$ in turn entail the following joint PMF $p_{X,Y}$:

		$p_{X,Y}(x,y)$			$p_Y(y)$
		$x = -1$	$x = 0$	$x = 1$	
$y = 0$	0	1/3	0	1/3	1/3
	1/3	0	1/3	2/3	
$p_X(x)$	1/3	1/3	1/3		

But, for example

$$p_{X,Y}(x = -1, y = 0) = 0 \neq \frac{1}{9} = \frac{1}{3} \cdot \frac{1}{3} = p_X(x = -1)p_Y(y = 0) \quad (256)$$

and hence X and Y are not independent.

□

Theorem (Variances of sums and differences of random variables)

Let X and Y denote two random variables and let $a, b, c \in \mathbb{R}$. Then

$$\mathbb{V}(aX + bY + c) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathbb{C}(X, Y). \quad (257)$$

In particular,

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathbb{C}(X, Y) \quad (258)$$

and

$$\mathbb{V}(X - Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\mathbb{C}(X, Y) \quad (259)$$

Remarks

- Variances of random variable do not simply add.
- The variance of the sum of two random variables depends on their covariance.

Covariance and correlation

Proof (\rightarrow Übung)

We first note that

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c. \quad (260)$$

We thus have

$$\begin{aligned} & \mathbb{V}(aX + bY + c) \\ &= \mathbb{E} \left((aX + bY + c - a\mathbb{E}(X) - b\mathbb{E}(Y) - c)^2 \right) \\ &= \mathbb{E} \left((a(X - \mathbb{E}(X)) + b(Y - \mathbb{E}(Y)))^2 \right) \\ &= \mathbb{E} \left(a^2(X - \mathbb{E}(X))^2 + b^2(Y - \mathbb{E}(Y))^2 + 2ab(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right) \\ &= a^2\mathbb{E} \left((X - \mathbb{E}(X))^2 \right) + b^2\mathbb{E} \left((Y - \mathbb{E}(Y))^2 \right) + 2ab\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\ &= a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\mathbb{C}(X, Y) \end{aligned} \quad (261)$$

The special cases then follow directly with $a = b = 1$ and with $a = 1, b = -1$, respectively.

□

Theorem (Correlation and linear-affine transformations)

Let X and Y denote two random variables with $\mathbb{V}(X) > 0$ and $\mathbb{V}(Y) > 0$. Then

$$Y = aX + b \Leftrightarrow \rho(X, Y) = 1 \text{ or } \rho(X, Y) = -1. \quad (262)$$

Remarks

- If Y is a linear-affine transformation of X with $\mathbb{V}(Y) > 0$, then $\rho(X, Y) = \pm 1$.
- If $\rho(X, Y) = \pm 1$ and $\mathbb{V}(Y) > 0$, then Y is a linear-affine transformation of X .
- If $-1 < \rho(X, Y) < 1$ and $\mathbb{V}(Y) > 0$, Y is not a linear-affine transformation of X .
- $\rho(X, Y)$ is commonly interpreted as a “measure of linear dependence”.
- A random variable is either a linear-affine transformation of another or not.

Correlation and linear-affine transformations

Proof

We content with showing the \Rightarrow direction. We assume that $Y = aX + b$ with $\mathbb{V}(X) > 0$ and $\mathbb{V}(Y) > 0$ hold and first show that the above implies that

$$\mathbb{S}(Y) = \pm a\mathbb{S}(X) \text{ and } \mathbb{C}(X, Y) = a\mathbb{V}(X). \quad (263)$$

To see this, we first note that as seen above

$$\mathbb{E}(Y) = a\mathbb{E}(X) + b \text{ and } \mathbb{V}(Y) = a^2\mathbb{V}(X). \quad (264)$$

Thus $a \neq 0$ and if $a > 0$, then $\mathbb{S}(Y) = a\mathbb{S}(X)$, and if $a < 0$, then $\mathbb{S}(Y) = -a\mathbb{S}(X) > 0$.

Next, with respect to

$$\mathbb{C}(X, Y) = \mathbb{E}((Y - \mathbb{E}(Y))(X - \mathbb{E}(X))), \quad (265)$$

we note that

$$Y - \mathbb{E}(Y) = aX + b - \mathbb{E}(Y) = aX + b - a\mathbb{E}(X) - b = a(X - \mathbb{E}(X)). \quad (266)$$

We thus obtain

$$\mathbb{C}(X, Y) = \mathbb{E}(a(X - \mathbb{E}(X))^2) = a\mathbb{E}((X - \mathbb{E}(X))^2) = a\mathbb{V}(X). \quad (267)$$

With (263), it then follows

$$\rho(X, Y) = \frac{\mathbb{C}(X, Y)}{\mathbb{S}(X)\mathbb{S}(Y)} = \frac{a\mathbb{V}(X)}{\mathbb{S}(X)(\pm a\mathbb{S}(X))} = \pm \frac{a\mathbb{V}(X)}{a\mathbb{V}(X)} = \pm 1. \quad (268)$$

□

Expectation and covariance

- Expectation
- Variance and standard deviation
- Sample mean, sample variance, sample standard deviation
- Covariance and correlation
- **Sample covariance and sample correlation**
- Exercises

Definition (Sample covariance and sample correlation)

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ denote two-dimensional random vectors. Then

- the sample mean of $(X_1, Y_1), \dots, (X_n, Y_n)$ is defined as

$$\overline{(X, Y)}_n := (\bar{X}_n, \bar{Y}_n) = \left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n Y_i \right), \quad (269)$$

- the sample covariance of $(X_1, Y_1), \dots, (X_n, Y_n)$ is defined as

$$C_n := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}), \quad (270)$$

- the sample correlation coefficient of $(X_1, Y_1), \dots, (X_n, Y_n)$ is defined as

$$R_n := \frac{C_n}{S_{X,n} S_{Y,n}}, \quad (271)$$

where $S_{X,n}$ and $S_{Y,n}$ are the sample standard deviations of X_1, \dots, X_n and Y_1, \dots, Y_n .

Remarks

- $\overline{(X, Y)}_n$, C_n , and R_n denote random variables, $\overline{(x, y)}_n$, c_n , and r_n denote realizations.

Sample covariance and sample correlation

Example (Sample covariance and sample correlation)

- Let $(X_1, X_2), \dots, (X_{10}, Y_{10}) \sim N \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$.
- Assume the following realizations

$$\begin{array}{cccccccccccc} (x_1, y_1) & (x_2, y_2) & (x_3, y_3) & (x_4, y_4) & (x_5, y_5) & (x_6, y_6) & (x_7, y_7) & (x_8, y_8) & (x_9, y_9) & (x_{10}, y_{10}) \\ \hline (0.8, -0.7) & (1.1, 1.6) & (-0.8, 1.1) & (-0.2, 0.1) & (1.1, 0.4) & (0.5, 1.5) & (1.3, -1.2) & (1.8, 0.6) & (0.4, 0.2) & (1.5, -1.0) \end{array}$$

- Sample mean realization

$$\overline{(x, y)}_{10} = \left(\frac{1}{10} \sum_{i=1}^{10} x_i, \frac{1}{10} \sum_{i=1}^{10} y_i \right) = (0.75, 0.26) \quad (272)$$

- Sample standard deviation realizations

$$s_{X,n} = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x}_{10})^2} = 0.79 \text{ and } s_{Y,n} = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (y_i - \bar{y}_{10})^2} = 0.99. \quad (273)$$

- Sample covariance and sample correlation realization

$$c_n = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}_i) = -0.26 \text{ and } r_n = \frac{c_n}{s_{x,n}s_{y,n}} = -0.33. \quad (274)$$

Expectation and covariance

- Expectation
- Variance and standard deviation
- Sample mean, sample variance, sample standard deviation
- Covariance and correlation
- Sample covariance and sample correlation
- **Exercises**

Study Questions

1. Discuss the intuition of the expected value of a random variable.
2. What does it mean for the expectation of a random variable to exist?
3. Compute the expectation of a Bernoulli random variable.
4. State the linearity and multiplication properties of expectations.
5. Write down $\mathbb{E}(X^2)$ in terms of the variance and expectation of the random variable X .
6. For constant a , what is $\mathbb{V}(aX)$?
7. Write down the definition of the covariance and correlation of two random variables X and Y .
8. Express the covariance of two random variables X and Y in terms of expectations.
9. What is the variance of the sum of two random variables X and Y , if X and Y are independent and in general?
10. What is the variance of the difference of two random variables X and Y , if X and Y are independent and in general?

Exercises

Theoretical Exercises

- Prove the theorem on the properties of expectations.
- Prove the theorem on the properties of variances.
- Prove the theorem on the variances of sums/differences of random variables.

Programming Exercises

- Sample $n = 10$ data points of a univariate Gaussian distribution and evaluate the sample mean, sample variance, and sample standard deviation.
- Sample $n = 10$ data points of a bivariate Gaussian distribution and evaluate the sample covariation and sample correlation.
- Validate the theorem on the variances of sums and differences of random variables using a sampling approach in a bivariate Gaussian scenario (\rightarrow Live Programming).

(7) Inequalities and limits

Bibliographic remarks

The material presented in this section is not to be understood as a comprehensive introduction to the respective inequalities and limits in probability theory, but merely serves as a collection of results that will be put to use in later sections. For discussions of the inequalities discussed see e.g. (Wasserman, 2004, Sections 4.1 and 4.2) and Casella and Berger (Sections 3.6 and 4.7 2012)

Inequalities and limits

- Probability inequalities
- Expectation inequalities
- Laws of large numbers
- Central limit theorems
- Exercises

Inequalities and limits

- **Probability inequalities**
- Expectation inequalities
- Laws of large numbers
- Central limit theorems
- Exercises

Theorem (Markov inequality)

Let X denote a random variable with $\mathbb{P}(X \geq 0) = 1$. Then for all $x \in \mathbb{R}$ it holds that

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (275)$$

Theorem (Markov inequality)

Let X denote a random variable with $\mathbb{P}(X \geq 0) = 1$. Then for all $x \in \mathbb{R}$ it holds that

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (275)$$

Remarks

- The Markov inequality relates exceedance probabilities and expectations.
- For example, if $\mathbb{E}(X) = 1$, then $\mathbb{P}(X \geq 100) \leq 0.01$.

Proof (Markov inequality)

We consider the case of a continuous X with PDF p . We first note that

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \xi p(\xi) d\xi = \int_{-\infty}^x \xi p(\xi) d\xi + \int_x^{\infty} \xi p(\xi) d\xi. \quad (276)$$

Proof (Markov inequality)

We consider the case of a continuous X with PDF p . We first note that

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \xi p(\xi) d\xi = \int_{-\infty}^x \xi p(\xi) d\xi + \int_x^{\infty} \xi p(\xi) d\xi. \quad (276)$$

With $\mathbb{P}(X \geq 0) = 1$ it then follows that

$$\mathbb{E}(X) \geq \int_x^{\infty} \xi p(\xi) d\xi \geq \int_x^{\infty} x p(\xi) d\xi = x \int_x^{\infty} p(\xi) d\xi = x \mathbb{P}(X \geq x). \quad (277)$$

Proof (Markov inequality)

We consider the case of a continuous X with PDF p . We first note that

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} \xi p(\xi) d\xi = \int_{-\infty}^x \xi p(\xi) d\xi + \int_x^{\infty} \xi p(\xi) d\xi. \quad (276)$$

With $\mathbb{P}(X \geq 0) = 1$ it then follows that

$$\mathbb{E}(X) \geq \int_x^{\infty} \xi p(\xi) d\xi \geq \int_x^{\infty} x p(\xi) d\xi = x \int_x^{\infty} p(\xi) d\xi = x \mathbb{P}(X \geq x). \quad (277)$$

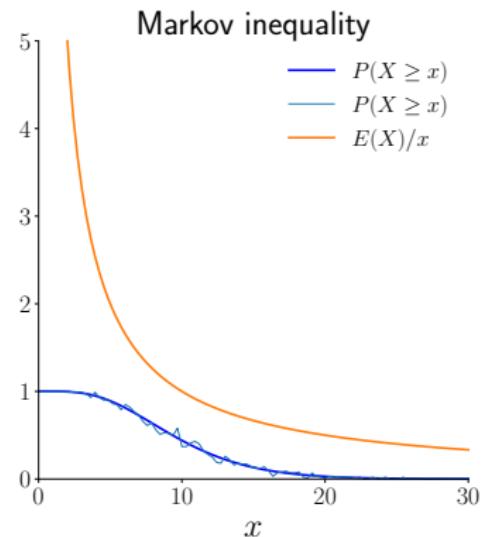
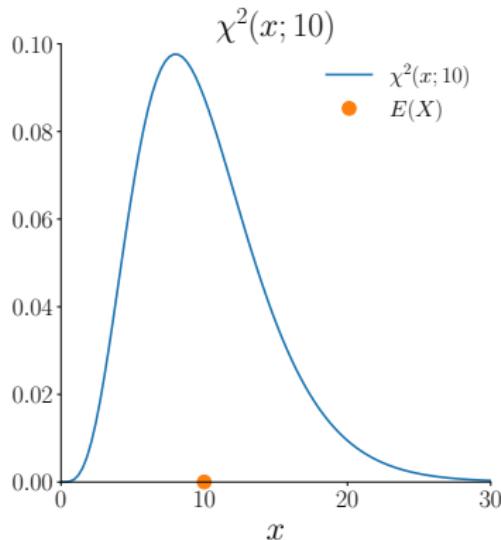
Hence

$$\mathbb{E}(X) \geq x \mathbb{P}(X \geq x) \Leftrightarrow \mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}. \quad (278)$$

□

Probability inequalities

Example: $X \sim \chi^2(n)$ with $\mathbb{E}(X) = n$



Theorem (Chebychev inequality)

Let X denote a random variable with variance $\mathbb{V}(X)$. Then for all $x \in \mathbb{R}$

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}. \quad (279)$$

Theorem (Chebychev inequality)

Let X denote a random variable with variance $\mathbb{V}(X)$. Then for all $x \in \mathbb{R}$

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}. \quad (279)$$

Remarks

- The Chebychev inequality relates deviations from expectations to variances.
- Note that

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq 3\sqrt{\mathbb{V}(X)}\right) \leq \frac{\mathbb{V}(X)}{\left(3\sqrt{\mathbb{V}(X)}\right)^2} = \frac{1}{9}.$$

Proof (Chebychev inequality)

We first note that for $a, b \in \mathbb{R}$, it holds that

$$a^2 \geq b^2 \Leftrightarrow |a| \geq b \quad (280)$$

Probability inequalities

Proof (Chebychev inequality)

We first note that for $a, b \in \mathbb{R}$, it holds that

$$a^2 \geq b^2 \Leftrightarrow |a| \geq b \quad (280)$$

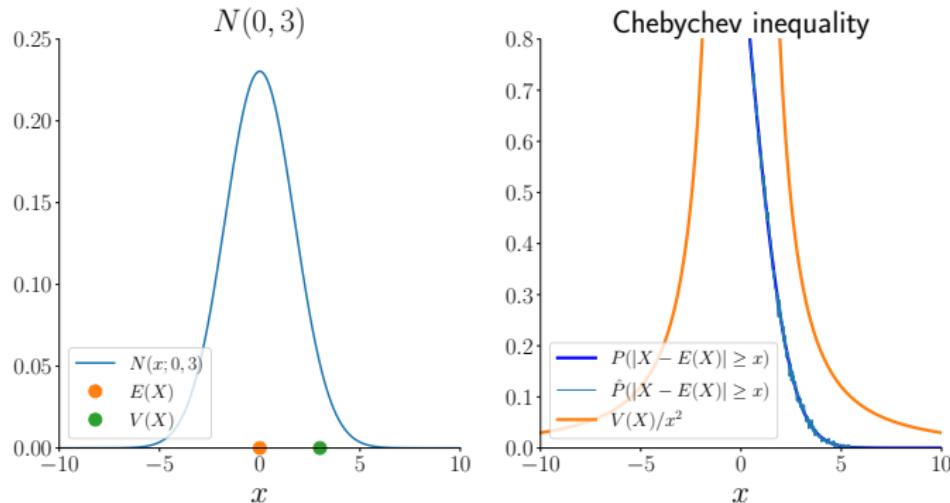
Next, set $Y := X - \mathbb{E}(X)$. Then with the Markov inequality

$$\begin{aligned} \mathbb{P}(Y \geq x^2) &\leq \frac{\mathbb{E}(Y)}{x^2} \\ \Leftrightarrow \mathbb{P}((X - \mathbb{E}(X))^2 \geq x^2) &\leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{x^2} \\ \Leftrightarrow \mathbb{P}(|X - \mathbb{E}(X)| \geq x) &\leq \frac{\mathbb{V}(X)}{x^2}. \end{aligned} \quad (281)$$

□

Probability inequalities

Example: $X \sim N(\mu, \sigma^2)$ with $\mathbb{V}(X) = \sigma^2$



Remark

- Note that here $\mathbb{P}(|X - \mathbb{E}(X)| \geq x) = \mathbb{P}(X - \mathbb{E}(X) \geq x) + \mathbb{P}(X - \mathbb{E}(X) \leq -x)$.

Inequalities and limits

- Probability inequalities
- **Expectation inequalities**
- Laws of large numbers
- Central limit theorems
- Exercises

Theorem (Cauchy-Schwarz inequality)

Let X, Y be two random variables such that $\mathbb{E}(XY)$ exists. Then

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2) \mathbb{E}(Y^2) \quad (282)$$

Theorem (Cauchy-Schwarz inequality)

Let X, Y be two random variables such that $\mathbb{E}(XY)$ exists. Then

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2) \mathbb{E}(Y^2) \quad (282)$$

Remarks

- This is $|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle$ for random variables.
- The correlation inequality is a direct consequence.

Proof (Cauchy-Schwarz inequality)

We consider the case that $0 < \mathbb{E}(X^2) < \infty$ and $0 < \mathbb{E}(Y^2) < \infty$. For all $a, b \in \mathbb{R}$, it holds that

$$0 \leq \mathbb{E}((aX + bY)^2) \text{ and } 0 \leq \mathbb{E}((aX - bY)^2). \quad (283)$$

Hence,

$$0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) + 2ab\mathbb{E}(XY) \text{ and } 0 \leq a^2\mathbb{E}(X^2) + b^2\mathbb{E}(Y^2) - 2ab\mathbb{E}(XY). \quad (284)$$

Setting $a := \sqrt{\mathbb{E}(Y^2)}$ and $b := \sqrt{\mathbb{E}(X^2)}$ then yields

$$\begin{aligned} 0 &\leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) + 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \\ &\Leftrightarrow -2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2) \\ &\Leftrightarrow -\sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)} \\ &\Leftrightarrow -\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}, \end{aligned} \quad (285)$$

and similarly

$$\begin{aligned} 0 &\leq \mathbb{E}(Y^2)\mathbb{E}(X^2) + \mathbb{E}(X^2)\mathbb{E}(Y^2) - 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \\ &\Leftrightarrow 2\sqrt{\mathbb{E}(Y^2)}\sqrt{\mathbb{E}(X^2)}\mathbb{E}(XY) \leq 2\mathbb{E}(X^2)\mathbb{E}(Y^2) \\ &\Leftrightarrow \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)} \\ &\Leftrightarrow \mathbb{E}(XY) \leq \sqrt{\mathbb{E}(Y^2)\mathbb{E}(X^2)}. \end{aligned} \quad (286)$$

Together, the above imply the Cauchy-Schwarz inequality. See DeGroot and Schervish (2012, Theorem 4.6.2) for a full proof.

□

Example (Bivariate Gaussian)

- Let $(X, Y) \sim N(\mu, \Sigma)$ with $\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{YX} & \sigma_{YY} \end{pmatrix}$
- Note that $\mathbb{C}(X, Y) = \sigma_{XY}$, the covariance translation theorem implies that

$$\mathbb{E}(XY)^2 = (\sigma_{XY} + \mu_X \mu_Y)^2 \quad (287)$$

- Similarly, with $\mathbb{V}(X) = \sigma_{XX}$ and $\mathbb{V}(Y) = \sigma_{YY}$ the variance translation theorem implies that

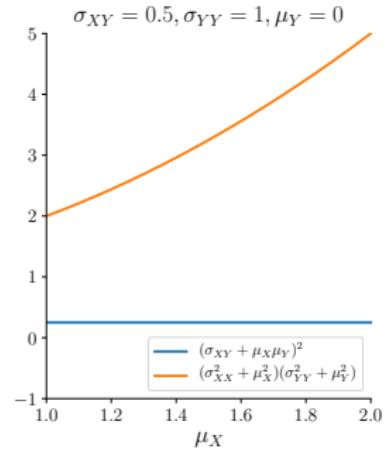
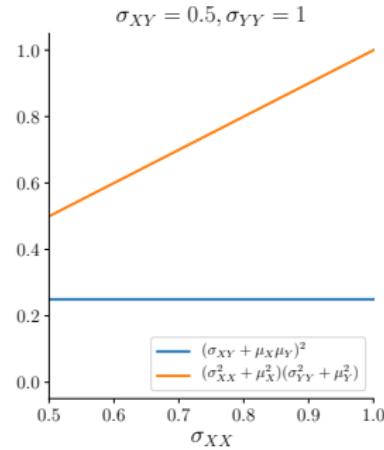
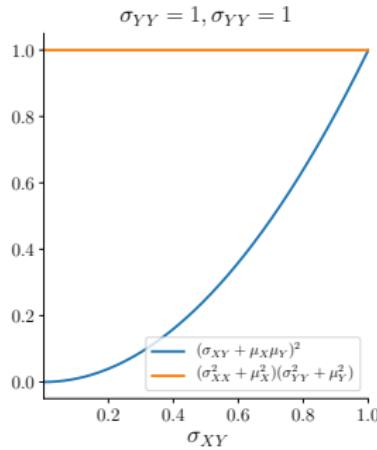
$$\mathbb{E}(X^2) = \sigma_{XX}^2 + \mu_X^2 \text{ and } \mathbb{E}(Y^2) = \sigma_{YY}^2 + \mu_Y^2 \quad (288)$$

- The Cauchy-Schwarz inequality thus states that

$$(\sigma_{XY} + \mu_X \mu_Y)^2 \leq (\sigma_{XX}^2 + \mu_X^2)(\sigma_{YY}^2 + \mu_Y^2) \quad (289)$$

Expectation inequalities

Example (Bivariate Gaussian)



Theorem (Correlation inequality)

Let X and Y denote two random variables with $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Then

$$\rho(X, Y)^2 = \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1. \quad (290)$$

Theorem (Correlation inequality)

Let X and Y denote two random variables with $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Then

$$\rho(X, Y)^2 = \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1. \quad (290)$$

Remark

- $\rho(X, Y)^2 \leq 1 \Leftrightarrow |\rho(X, Y)| \leq 1 \Leftrightarrow \rho(X, Y) \in [-1, 1]$.

Expectation inequalities

Theorem (Correlation inequality)

Let X and Y denote two random variables with $\mathbb{V}(X), \mathbb{V}(Y) > 0$. Then

$$\rho(X, Y)^2 = \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1. \quad (290)$$

Remark

- $\rho(X, Y)^2 \leq 1 \Leftrightarrow |\rho(X, Y)| \leq 1 \Leftrightarrow \rho(X, Y) \in [-1, 1]$.

Proof

According to the Cauchy-Schwarz inequality for random variables U, V , it holds that

$$(\mathbb{E}(UV))^2 \leq \mathbb{E}(U^2)\mathbb{E}(V^2). \quad (291)$$

Set $U := X - \mathbb{E}(X)$ and $V := Y - \mathbb{E}(Y)$. Then according to the Cauchy-Schwarz inequality

$$\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^2) \leq \mathbb{E}\left((X - \mathbb{E}(X))^2\right)\mathbb{E}\left((Y - \mathbb{E}(Y))^2\right) \quad (292)$$

Hence,

$$\mathbb{C}(X, Y)^2 \leq \mathbb{V}(X)\mathbb{V}(Y) \Leftrightarrow \frac{\mathbb{C}(X, Y)^2}{\mathbb{V}(X)\mathbb{V}(Y)} \leq 1. \quad (293)$$

□

Expectation inequalities

Theorem (Jensen's inequality)

Let X be a random variable and g be a convex function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2). \quad (294)$$

Then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)). \quad (295)$$

Conversely, let g be a concave function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda g(x_1) + (1 - \lambda)g(x_2). \quad (296)$$

Then

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)). \quad (297)$$

Theorem (Jensen's inequality)

Let X be a random variable and g be a convex function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2). \quad (294)$$

Then

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)). \quad (295)$$

Conversely, let g be a concave function, i.e.

$$g(\lambda x_1 + (1 - \lambda)x_2) \geq \lambda g(x_1) + (1 - \lambda)g(x_2). \quad (296)$$

Then

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)). \quad (297)$$

Remarks

- For convex g the function's graph lies below the straight line $g(x_1)$ to $g(x_2)$.
- For concave g the function's graph lies above the straight line $g(x_1)$ to $g(x_2)$.
- The logarithm is a concave function, hence $\mathbb{E}(\ln X) \leq \ln \mathbb{E}(X)$.

Proof

We show the inequality for the concave case.

Proof

We show the inequality for the concave case.

- Let f be a tangent line at the point $g(\mathbb{E}(X))$, i.e. is a linear-affine function of the form $f(X) := aX + b$ for some $a, b \in \mathbb{R}$ with $f(\mathbb{E}(X)) = g(\mathbb{E}(X))$.

Proof

We show the inequality for the concave case.

- Let f be a tangent line at the point $g(\mathbb{E}(X))$, i.e. is a linear-affine function of the form $f(X) := aX + b$ for some $a, b \in \mathbb{R}$ with $f(\mathbb{E}(X)) = g(\mathbb{E}(X))$.
- Because g is concave, we have $g(x) \leq ax+b$ for all $x \in \mathbb{R}$ and thus also $g(X) \leq aX+b$.

Proof

We show the inequality for the concave case.

- Let f be a tangent line at the point $g(\mathbb{E}(X))$, i.e. is a linear-affine function of the form $f(X) := aX + b$ for some $a, b \in \mathbb{R}$ with $f(\mathbb{E}(X)) = g(\mathbb{E}(X))$.
- Because g is concave, we have $g(x) \leq ax+b$ for all $x \in \mathbb{R}$ and thus also $g(X) \leq aX+b$.
- Hence,

$$\mathbb{E}(g(X)) \leq \mathbb{E}(aX + b) = a\mathbb{E}(X) + b = f(\mathbb{E}(X)) = g(\mathbb{E}(X)). \quad (298)$$

□

Inequalities and limits

- Probability inequalities
- Expectation inequalities
- **Laws of large numbers**
- Central limit theorems
- Exercises

Overview

Intuitively, both the Weak and the Strong Law of Large Numbers state that for i.i.d. random samples of a distribution, the sample mean approximates the expectation of that distribution for large sample sizes. The “Weak” and the “Strong” form of the Law of Large Numbers differ in the form of random variable convergence considered

Definition (Convergence in probability)

A sequence X_1, X_2, \dots of random variables *converges to a random variable X in probability*, written as

$$X_n \xrightarrow[n \rightarrow \infty]{P} X, \quad (299)$$

if for every $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \epsilon) = 1 \Leftrightarrow \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \quad (300)$$

Definition (Convergence in probability)

A sequence X_1, X_2, \dots of random variables *converges to a random variable X in probability*, written as

$$X_n \xrightarrow[n \rightarrow \infty]{P} X, \quad (299)$$

if for every $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \epsilon) = 1 \Leftrightarrow \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0 \quad (300)$$

Remarks

- $X_n \xrightarrow[n \rightarrow \infty]{P} X$ means that the probability that X_n lies inside the random interval $[X - \epsilon, X + \epsilon]$, no matter how small this interval may be, approaches 1 as $n \rightarrow \infty$.
- Intuitively, for a constant random variable $X := x$, this means that the probability distribution of X_i becomes increasingly concentrated around x as $n \rightarrow \infty$.

Theorem (Weak Law of Large Numbers)

Let X_1, \dots, X_n denote a random sample from a distribution with expectation μ . Let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (301)$$

denote the sample mean. Then \bar{X}_n converges to μ in probability,

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{P} \mu. \quad (302)$$

Theorem (Weak Law of Large Numbers)

Let X_1, \dots, X_n denote a random sample from a distribution with expectation μ . Let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (301)$$

denote the sample mean. Then \bar{X}_n converges to μ in probability,

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{P} \mu. \quad (302)$$

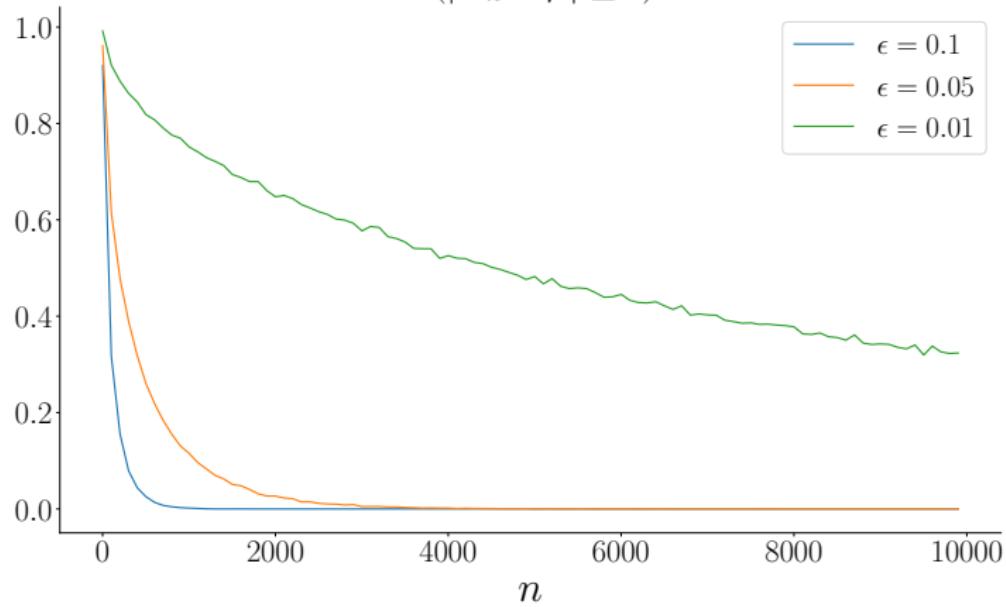
Remark

- $\bar{X}_n \xrightarrow[n \rightarrow \infty]{P} \mu$ means that the probability that the sample mean is identical to the expectation of $X_i, i = 1, \dots, n$ approaches 1 as the sample size $n \rightarrow \infty$.

Laws of large numbers

$$X_1, \dots, X_n \sim N(0, 1)$$

$$\hat{P}(|\bar{X}_n - \mu| \geq \epsilon)$$



Definition (Almost sure convergence)

A sequence X_1, X_2, \dots of random variables *converges almost surely to a random variable X* , written as

$$X_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} X, \quad (303)$$

if for every $\varepsilon > 0$

$$\mathbb{P} \left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon \right) = 1. \quad (304)$$

Definition (Almost sure convergence)

A sequence X_1, X_2, \dots of random variables *converges almost surely to a random variable X* , written as

$$X_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} X, \quad (303)$$

if for every $\varepsilon > 0$

$$\mathbb{P} \left(\lim_{n \rightarrow \infty} |X_n - X| < \varepsilon \right) = 1. \quad (304)$$

Remarks

- Recall that for $(\Omega, \mathcal{A}, \mathbb{P})$ random variables are functions $X : \Omega \rightarrow \mathcal{X}$.
- Let $N \subset \Omega$ be a null set, i.e. $\mathbb{P}(N) = 0$.
- A.s. convergence implies $X_n(\omega) \rightarrow X(\omega)$ for all $\omega \in \Omega \setminus N$.
- A.s. convergence corresponds to pointwise convergence of function sequences
- A.s. convergence implies convergence in probability, but not vice versa.
- A.s. convergence is a strong form of random variable convergence.

Theorem (Strong Law of Large Numbers)

Let X_1, \dots, X_n denote a random sample from a distribution with expectation μ . Let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (305)$$

denote the sample mean. Then \bar{X}_n converges almost surely to μ ,

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu. \quad (306)$$

Theorem (Strong Law of Large Numbers)

Let X_1, \dots, X_n denote a random sample from a distribution with expectation μ . Let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (305)$$

denote the sample mean. Then \bar{X}_n converges almost surely to μ ,

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu. \quad (306)$$

Remark

- $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ means that the probability that $|X_n - \mu|$ is smaller than some arbitrarily small $\epsilon > 0$ is 1 as n goes to infinity.

Inequalities and limits

- Probability inequalities
- Expectation inequalities
- Laws of large numbers
- **Central limit theorems**
- Exercises

Overview

Intuitively, the Central Limit Theorem states that the sum of very many independent and arbitrarily distributed random variables is normally distributed. The “Lindenberg and Lévy” form assumes independent and identically distributed random variables and is easier to prove than the “Liapunov” form, which only assumes independent random variables.

Definition (Convergence in distribution and asymptotic distribution)

A sequence X_1, X_2, \dots of random variables *converges to a random variable X in distribution*, written as

$$X_n \xrightarrow[n \rightarrow \infty]{D} X, \quad (307)$$

if

$$\lim_{n \rightarrow \infty} P_{X_n}(x) = P_X(x). \quad (308)$$

at all points where P_X is continuous. If $X_n \xrightarrow[n \rightarrow \infty]{D} X$, then the distribution of X is referred to as the *asymptotic distribution of X_n* .

Definition (Convergence in distribution and asymptotic distribution)

A sequence X_1, X_2, \dots of random variables *converges to a random variable X in distribution*, written as

$$X_n \xrightarrow[n \rightarrow \infty]{D} X, \quad (307)$$

if

$$\lim_{n \rightarrow \infty} P_{X_n}(x) = P_X(x). \quad (308)$$

at all points where P_X is continuous. If $X_n \xrightarrow[n \rightarrow \infty]{D} X$, then the distribution of X is referred to as the *asymptotic distribution of X_n* .

Remarks

- $X \xrightarrow[n \rightarrow \infty]{D} X$ is a statement about the convergence of CDFs.
- Convergence in probability implies convergence in distribution.
- Almost sure convergence implies convergence in distribution.
- Convergence in distribution is a weak form of convergence.

Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables X_1, \dots, X_n form an independent and identically distributed random sample of size n from a given distribution with expectation μ and variance $0 < \sigma^2 < \infty$.

Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables X_1, \dots, X_n form an independent and identically distributed random sample of size n from a given distribution with expectation μ and variance $0 < \sigma^2 < \infty$. Then for each fixed number x

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{\bar{X}_n - \mu}{\sigma/n^{1/2}} \leq x \right) = \Phi(x) \quad (309)$$

where Φ is the CDF of the standard normal distribution.

Theorem (Central limit theorem (Lindenberg and Lévy))

Let the random variables X_1, \dots, X_n form an independent and identically distributed random sample of size n from a given distribution with expectation μ and variance $0 < \sigma^2 < \infty$. Then for each fixed number x

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{\bar{X}_n - \mu}{\sigma/n^{1/2}} \leq x \right) = \Phi(x) \quad (309)$$

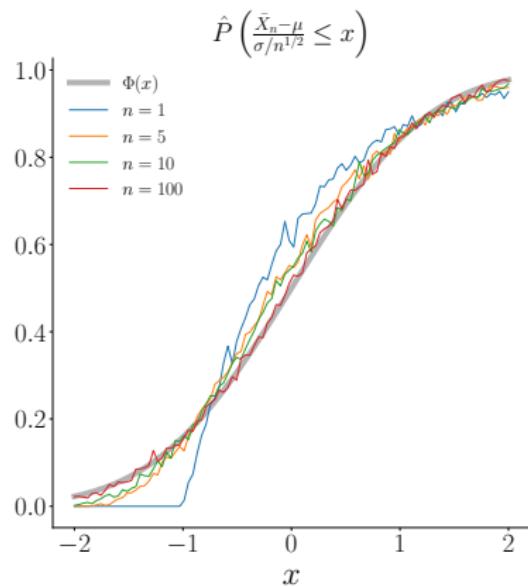
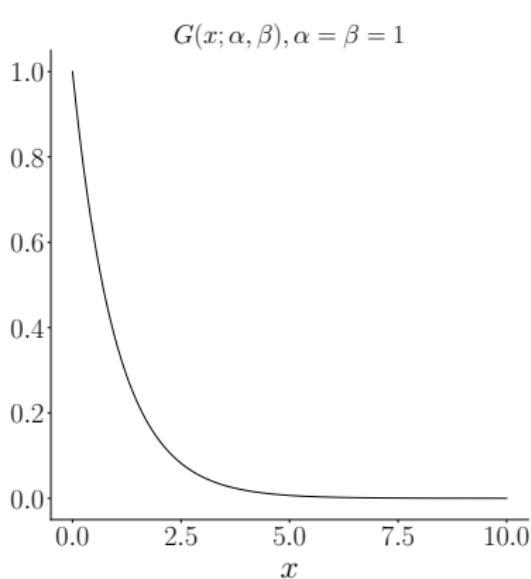
where Φ is the CDF of the standard normal distribution.

Remarks

- For large n , the distribution of $n^{1/2}(\bar{X}_n - \mu)/\sigma$ is approximately $N(0, 1)$.
- For large n , the distribution of \bar{X}_n is approximately $N(\mu, \sigma^2/n)$.
- For large n , the distribution of $\sum_{i=1}^n X_i$ is approximately $N(n\mu, n\sigma^2)$.

Central limit theorems

$X_1, \dots, X_n \sim G(x; \alpha, \beta)$ with $\mu = \frac{\alpha}{\beta}$, $\sigma^2 = \frac{\alpha}{\beta^2}$



Theorem (Central limit theorem (Liapounov))

Let X_1, X_2, \dots be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(|X_i - \mu_i|^3) < \infty \text{ and } \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \quad (310)$$

Theorem (Central limit theorem (Liapounov))

Let X_1, X_2, \dots be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(|X_i - \mu_i|^3) < \infty \text{ and } \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \quad (310)$$

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 = \mathbb{V}(X_i)$ for $i = 1, \dots, n$ and define

$$Y_n := \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}, \quad (311)$$

such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$.

Theorem (Central limit theorem (Liapounov))

Let X_1, X_2, \dots be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(|X_i - \mu_i|^3) < \infty \text{ and } \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \quad (310)$$

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 = \mathbb{V}(X_i)$ for $i = 1, \dots, n$ and define

$$Y_n := \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}, \quad (311)$$

such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$. Then, for each fixed number x ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Y_n \leq x) = \Phi(x), \quad (312)$$

where Φ is the CDF of the standard normal distribution.

Theorem (Central limit theorem (Liapounov))

Let X_1, X_2, \dots be a sequence of independent, but not necessarily identically, distributed random variables, such that

$$\mathbb{E}(|X_i - \mu_i|^3) < \infty \text{ and } \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \mathbb{E}(|X_i - \mu_i|^3)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \quad (310)$$

Let $\mu_i := \mathbb{E}(X_i)$ and $\sigma_i^2 = \mathbb{V}(X_i)$ for $i = 1, \dots, n$ and define

$$Y_n := \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}, \quad (311)$$

such that $\mathbb{E}(Y_n) = 0$ and $\mathbb{V}(Y_n) = 1$. Then, for each fixed number x ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(Y_n \leq x) = \Phi(x), \quad (312)$$

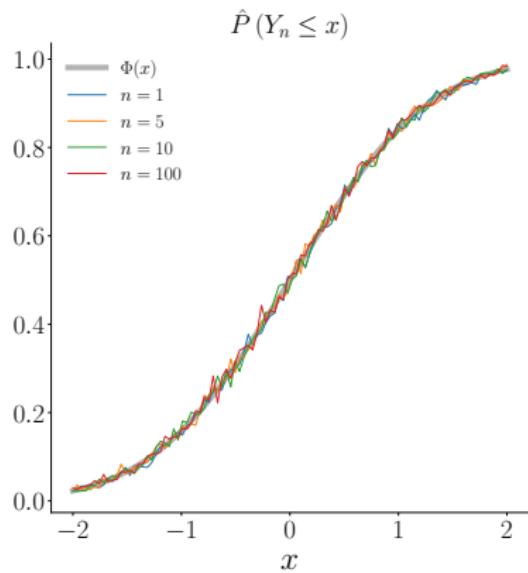
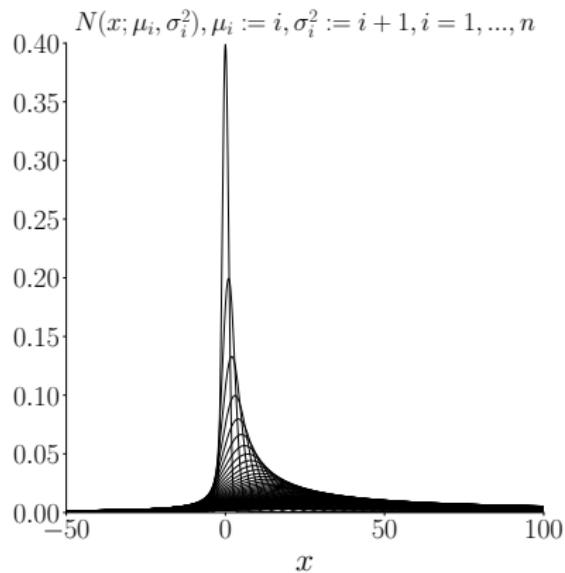
where Φ is the CDF of the standard normal distribution.

Remarks

- For large n , the distribution of $\sum_{i=1}^n X_i$ is approximately $N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.
- The sum of many independent random factors is approximately normally distributed.
- This justifies the ubiquitous assumption of normally distributed observation errors.

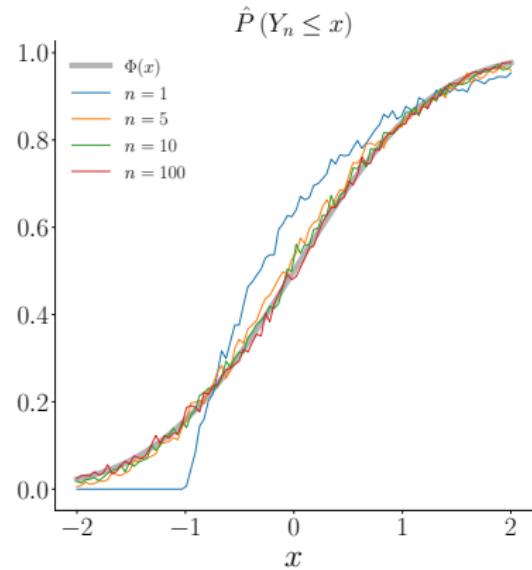
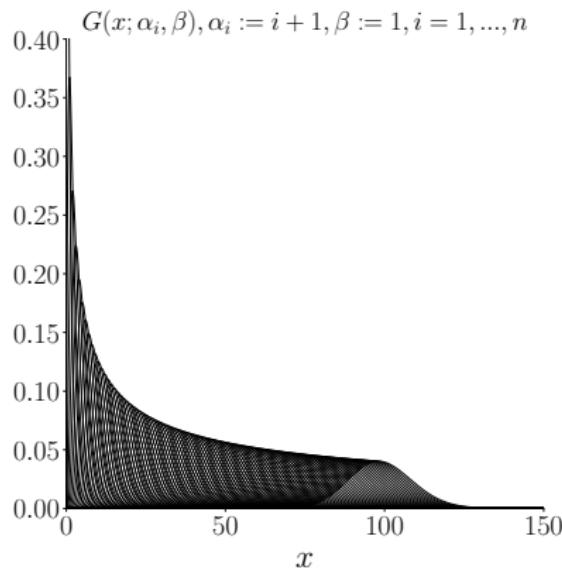
Central limit theorems

$$X_1, \dots, X_n \sim N(x; \mu_i, \sigma_i^2)$$



Central limit theorems

$X_1, \dots, X_n \sim G(x; \alpha_i, \beta)$ with $\mu = \frac{\alpha_i}{\beta}$, $\sigma^2 = \frac{\alpha_i}{\beta^2}$



Inequalities and limits

- Probability inequalities
- Expectation inequalities
- Laws of large numbers
- Central limit theorems
- **Exercises**

Study questions

1. Write down the Markov inequality.
2. Write down the Chebychev inequality.
3. Write down Jensen's inequality for concave functions.
4. Write down the Cauchy-Schwarz inequality.
5. Write down the Correlation inequality.
6. Write down the definition of convergence in probability.
7. Write down the definition of almost-sure convergence.
8. Write down the definition of convergence in distribution.
9. Write down the Weak Law of Large Numbers.
10. Write down the Strong Law of Large Numbers.
11. Write down the Lindenberg-Lévy form of the Central Limit Theorem.
12. Write down the Liapunov form of the Central Limit Theorem.

Exercises

Programming exercises

1. Write simulations that validate the Markov and Chebychev inequalities.
2. Write a simulation that validates the Weak Law of Large Numbers.
3. Write a simulation that validates the Lindenberg-Lévy Central Limit Theorem.
4. Write a simulation that validates the Liapunov Central Limit Theorem.

(8) Maximum likelihood estimation

Bibliographic remarks

The material presented in this section follows Wasserman (2004, Sections 6.1 - 6.3, 9.1, 9.3) and Held and Sabanés Bové (2014, Sections 2.2 - 2.3, C.1.3).

Maximum likelihood estimation

- Statistical models
- Maximum likelihood estimation
- Analytical examples
- Numerical approaches
- Exercises

Maximum likelihood estimation

- **Statistical models**
- Maximum likelihood estimation
- Analytical examples
- Numerical approaches
- Exercises

A statistical model \mathcal{P} is a set of probability distributions.

- A *parametric statistical model* is a statistical model that can be parameterized by a finite number of parameters.
- A *nonparametric statistical model* is a statistical model that cannot be parameterized by a finite number of parameters.

Parametric statistical models

Typical parametric statistical models have the form

$$\mathcal{P} = \{p_\theta(x) | \theta \in \Theta\}, \quad (313)$$

where

- p_θ is a PMF or PDF parameterized by θ ,
- θ is a *parameter (vector)*, and
- Θ is the *parameter space*.

Examples

- $X_1, \dots, X_n \sim p_\mu$ and $p_\mu \in \mathcal{P} := \{\text{Bern}(x; \mu) | \mu \in]0, 1[\}$
- $X_1, \dots, X_n \sim p_{\mu, \sigma^2}$ and $p_{\mu, \sigma^2} \in \mathcal{P} := \{\mathcal{N}(x; \mu, \sigma^2) | (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}\}$

Example (The general linear model)

A very common parametric statistical model is the *general linear model*

$$Y = X\beta + \varepsilon, \varepsilon \sim N(0_n, \sigma^2 I_n) \Leftrightarrow p_{\beta, \sigma^2}(y) = N(y; X\beta, \sigma^2 I_n), \quad (314)$$

where

- Y is an n -dimensional random vector of observable data,
- ε is an n -dimensional random vector of i.i.d. errors,
- $X \in \mathbb{R}^{n \times p}$ is a known (non-random) design matrix, and
- $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$ are unknown parameters.

Thus $Y_1, \dots, Y_n \sim p_{\beta, \sigma^2}$ and $p_{\beta, \sigma^2} \in \mathcal{P} := \{N(y; X\beta, \sigma^2 I_n) | (\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{>0}\}$.

Simple and multiple linear regression, T-tests, ANOVA, and ANCOVA are special cases.

The standard problems of frequentist inference

(1) *Parameter estimation*

The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of $X_1, \dots, X_n \sim p_\theta$.

(2) *Confidence interval evaluation*

The aim of confidence interval evaluation is to provide a quantitative uncertainty statement about a parameter estimate based on the parameter estimator's sampling distribution.

(3) *Hypothesis testing*

The aim of hypothesis testing is to decide, based on the observations X_1, \dots, X_n and in a sensible fashion, whether the true, but unknown, parameter is in one of two mutually exclusive subsets of the parameter space.

Confidence interval evaluation and hypothesis testing make extensive use of *statistics* $h(X_1, \dots, X_n)$ and their distributional properties.

Statistical models

World

True, but unknown, parameter value

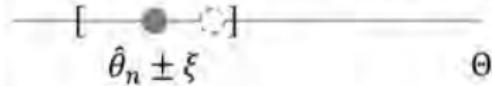


Frequentist inference

Point estimate

$$\theta \leftarrow \hat{\theta}_n$$

Confidence interval estimate $\mathbb{P}(\theta \in [\hat{\theta}_n \pm \xi])$



Hypothesis testing estimate $\theta \in \Theta_0$ vs. $\theta \in \Theta_1$



Definition (Point estimator)

Let $X_1, \dots, X_n \sim p_\theta$ be n independent and identically distributed observations of a parametric statistical model \mathcal{P} with parameter space Θ . A point estimator $\hat{\theta}_n$ for θ is a function of X_1, \dots, X_n that takes on values in Θ .

Remarks

- A point estimator provides a single best guess of some quantity of interest.
- As a function of random variables, a point estimator is a random variable.
- A value taken on by point estimator is referred to as *point estimate*.
- Frequentist inference is concerned with the properties of point estimators.
- *Maximum likelihood estimation* is the most common method for point estimation in parametric statistical methods.

The frequentist sampling intuition

- Let $X_1, \dots, X_n \sim p_\theta$
- Real observed data is considered one possible realization of $X_1, \dots, X_n \sim p_\theta$.
- From a sampling perspective, however, we could sample data and estimators

$x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}$ and $\hat{\theta}_n(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)})$, e.g., $\hat{\theta}_n^{(1)} = \frac{1}{n} \sum_{i=1}^n x_i^{(1)}$

$x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}$ and $\hat{\theta}_n(x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)})$, e.g., $\hat{\theta}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n x_i^{(2)}$

$x_1^{(3)}, x_2^{(3)}, \dots, x_n^{(3)}$ and $\hat{\theta}_n(x_1^{(3)}, x_2^{(3)}, \dots, x_n^{(3)})$, e.g., $\hat{\theta}_n^{(3)} = \frac{1}{n} \sum_{i=1}^n x_i^{(3)}$

$x_1^{(4)}, x_2^{(4)}, \dots, x_n^{(4)}$ and $\hat{\theta}_n(x_1^{(4)}, x_2^{(4)}, \dots, x_n^{(4)})$, e.g., $\hat{\theta}_n^{(4)} = \frac{1}{n} \sum_{i=1}^n x_i^{(4)}$

$x_1^{(5)}, x_2^{(5)}, \dots, x_n^{(5)}$ and $\hat{\theta}_n(x_1^{(5)}, x_2^{(5)}, \dots, x_n^{(5)})$, e.g., $\hat{\theta}_n^{(5)} = \frac{1}{n} \sum_{i=1}^n x_i^{(5)}$

...

- Frequentist inference is interested in the distributional properties of estimators.
- For example, what is the distribution of $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \hat{\theta}_n^{(3)}, \dots$?
- Finite-sample estimator properties concern the distribution of $\hat{\theta}_n$ for fixed n .
- Asymptotic estimator properties concern the distribution of $\hat{\theta}_n$ for $n \rightarrow \infty$.

Maximum likelihood estimation

- Statistical models
- **Maximum likelihood estimation**
- Analytical examples
- Numerical approaches
- Exercises

Definition (Likelihood function and log likelihood function)

Let \mathcal{P} denote a parametric statistical model with PMF or PDF $p_\theta(x)$ and let $X_1, \dots, X_n \sim p_\theta$. The *likelihood function* is defined as

$$L_n : \Theta \rightarrow [0, \infty[, \theta \mapsto L_n(\theta) := \prod_{i=1}^n p_\theta(x_i). \quad (315)$$

The *log likelihood function* is defined as

$$\ell_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto \ell_n(\theta) := \ln L_n(\theta). \quad (316)$$

Remarks

- The value of the likelihood function is the joint density of the data.
- The likelihood function is a function of the parameter.
- In general the likelihood function does not integrate to 1 with respect to θ .

Definition (Maximum likelihood estimator)

The *maximum likelihood estimator* (MLE) is defined as

$$\hat{\theta}_n^{ML} = \arg \max_{\theta \in \Theta} L_n(\theta) = \arg \max_{\theta \in \Theta} \ell_n(\theta). \quad (317)$$

Remarks

- Because \ln is monotonically increasing, the maximum of ℓ_n occurs at the same place as does the maximum of L_n .
- Working with the log likelihood function is often easier.
- Multiplying L_n by a positive constant not depending on θ does not change the MLE, for maximum likelihood estimation, constant contributions to likelihood functions can thus be neglected.
- Maximum likelihood is a standard problem in function maximization.

The general maximum likelihood procedure for parametric statistical models

- (1) Formulation of the log likelihood function.
- (2) Evaluation of the log likelihood function's derivative and setting to zero.
- (3) Solving for critical points, verification of maximum.

Two approaches for maximum likelihood estimation

- (1) Analytical function maximization in classical examples.
- (2) Numerical function maximization in most applied scenarios.

Maximum likelihood estimation

- Statistical models
- Maximum likelihood estimation
- **Analytical examples**
- Numerical approaches
- Exercises

Example (Bernoulli distribution)

Let $X_1, \dots, X_n \sim \text{Bern}(\mu)$ be n i.i.d. Bernoulli distributed random variables.

(1) Formulation of the log likelihood function

We have

$$L_n :]0, 1[\rightarrow]0, 1[, \mu \mapsto L_n(\mu) := \prod_{i=1}^n \mu^{x_i} (1-\mu)^{1-x_i} = \mu^{\sum_{i=1}^n x_i} (1-\mu)^{n - \sum_{i=1}^n x_i}. \quad (318)$$

Taking the logarithm yields

$$\ell_n :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto \ell_n(\mu) = \ln \mu \sum_{i=1}^n x_i + \ln(1-\mu) \left(n - \sum_{i=1}^n x_i \right). \quad (319)$$

Example (Bernoulli distribution)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

$$\begin{aligned}\frac{d}{d\mu} \ell_n(\mu) &= \frac{d}{d\mu} \left(\ln \mu \sum_{i=1}^n x_i + \ln(1-\mu) \left(n - \sum_{i=1}^n x_i \right) \right) \\ &= \frac{d}{d\mu} \ln \mu \sum_{i=1}^n x_i + \frac{d}{d\mu} \ln(1-\mu) \left(n - \sum_{i=1}^n x_i \right) \\ &= \frac{1}{\mu} \sum_{i=1}^n x_i - \frac{1}{1-\mu} \left(n - \sum_{i=1}^n x_i \right)\end{aligned}\tag{320}$$

and hence the *maximum likelihood equation* takes the form

$$\hat{\mu} \sum_{i=1}^n x_i - \frac{1}{1-\hat{\mu}} \left(n - \sum_{i=1}^n x_i \right) = 0.\tag{321}$$

Analytical examples

Example (Bernoulli distribution)

(3) Solving for critical points

We have

$$\begin{aligned} & \frac{1}{\hat{\mu}} \sum_{i=1}^n x_i - \frac{1}{1-\hat{\mu}} \left(n - \sum_{i=1}^n x_i \right) = 0 \\ \Leftrightarrow & \hat{\mu}(1-\hat{\mu}) \left(\frac{1}{\hat{\mu}} \sum_{i=1}^n x_i - \frac{1}{1-\hat{\mu}} \left(n - \sum_{i=1}^n x_i \right) \right) = 0 \\ \Leftrightarrow & \sum_{i=1}^n x_i - \hat{\mu} \sum_{i=1}^n x_i - n\hat{\mu} + \hat{\mu} \sum_{i=1}^n x_i = 0 \\ \Leftrightarrow & n\hat{\mu} = \sum_{i=1}^n x_i \\ \Leftrightarrow & \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i. \end{aligned} \tag{322}$$

Hence $\hat{\mu} = n^{-1} \sum_{i=1}^n x_i$ is a candidate for an MLE of μ .

This can be consolidated, such that $\hat{\mu}_n^{ML} := \frac{1}{n} \sum_{i=1}^n x_i$.

Example (Gaussian distribution)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ be n i.i.d. Gaussian random variables.

(1) Formulation of the log likelihood function

We have

$$\begin{aligned} L_n : \mathbb{R} \times \mathbb{R}_{>0} &\rightarrow \mathbb{R}_{>0}, (\mu, \sigma^2) \mapsto L_n(\mu, \sigma^2) := \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right). \end{aligned} \tag{323}$$

Taking the logarithm yields

$$\ell_n : \mathbb{R} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}, (\mu, \sigma^2) \mapsto \ell_n(\mu, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \tag{324}$$

Analytical examples

Example (Univariate Gaussian distribution)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

$$\frac{d}{d\mu} \ell_n(\mu) = -\frac{d}{d\mu} \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{d}{d\mu} (x_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu). \quad (325)$$

Similarly,

$$\frac{d}{d\sigma^2} \ell_n(\sigma^2) = -\frac{n}{2} \frac{d}{d\sigma^2} \ln \sigma^2 - \frac{d}{d\sigma^2} \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2. \quad (326)$$

Hence the *maximum likelihood equations* take the form

$$\begin{aligned} \sum_{i=1}^n (x_i - \hat{\mu}) &= 0 \\ -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (x_i - \mu)^2 &= 0 \end{aligned} \quad (327)$$

Example (Univariate Gaussian distribution)

(3) Solving for critical points

We have

$$\sum_{i=1}^n (x_i - \hat{\mu}) = 0 \Leftrightarrow \sum_{i=1}^n x_i = n\hat{\mu} \Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (328)$$

Hence $\hat{\mu} = n^{-1} \sum_{i=1}^n x_i$ is a candidate for an MLE of μ .

Substitution yields

$$\begin{aligned} & -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (x_i - \hat{\mu})^2 = 0 \\ & \Leftrightarrow -n\hat{\sigma}^2 + \sum_{i=1}^n (x_i - \hat{\mu})^2 = 0 \\ & \Leftrightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2. \end{aligned} \quad (329)$$

Hence $\hat{\sigma} = n^{-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$ is a candidate for an MLE of σ^2 .

Example (Univariate Gaussian distribution)

Both estimators can be consolidated, such that

$$\begin{aligned}\hat{\mu}_n^{ML} &:= \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\sigma}_n^{2ML} &:= \frac{1}{n} \sum_{i=1}^n \left(x_i - \hat{\mu}_n^{ML} \right)^2.\end{aligned}\tag{330}$$

Maximum likelihood estimation

- Statistical models
- Maximum likelihood estimation
- Analytical examples
- **Numerical approaches**
- Exercises

Numerical maximum likelihood estimation

- ... rests on formulating and implementing the log likelihood function.
- ... is equivalent to negative log likelihood minimization.
- ... is a problem in nonlinear numerical optimization.
- ... can be performed with a plethora of methods, such as
 - deterministic vs. stochastic optimization,
 - constrained vs. unconstrained optimization,
 - local vs. global optimization,
 - gradient-based vs. gradient-free optimization.

Nocedal and Wright (2006) provide an excellent introduction.

We only consider a special *Newton-Raphson method* called *Fisher scoring*.

The univariate Newton-Raphson method

- An iterative approach to find extrema of univariate real-valued functions .

$$f : \mathbb{R} \rightarrow \mathbb{R}, \theta \mapsto f(\theta)$$

- Based on the necessary condition for a maximum at $\tilde{\theta}$, $f'(\tilde{\theta}) = 0$.
- Starts from an initial guess $\theta^{(0)}$ and forms iterands $\theta^{(k)}$, $k = 1, 2, 3, \dots$
- Iterative approximation of $\tilde{\theta}$ by first-order Taylor approximation of f' .
- Maximization of a second-order Taylor approximation of f .
- For maximum likelihood estimation, f is a log likelihood function.

Evaluating a zero of $f'(\theta)$

$$f'(\theta) \approx \tilde{f}'(\theta) = f'(\theta^{(k)}) + f''(\theta^{(k)})(\theta - \theta^{(k)}). \quad (331)$$

We have

$$\begin{aligned} & \tilde{f}'(\tilde{\theta}) = 0 \\ \Leftrightarrow & f'(\theta^{(k)}) + f''(\theta^{(k)})(\tilde{\theta} - \theta^{(k)}) = 0 \\ \Leftrightarrow & f''(\theta^{(k)})(\tilde{\theta} - \theta^{(k)}) = -f'(\theta^{(k)}) \\ \Leftrightarrow & \tilde{\theta} - \theta^{(k)} = -\frac{f'(\theta^{(k)})}{f''(\theta^{(k)})} \\ \Leftrightarrow & \tilde{\theta} = \theta^{(k)} - \frac{f'(\theta^{(k)})}{f''(\theta^{(k)})}. \end{aligned} \quad (332)$$

The last equation implies an update rule for $\theta^{(k)}$.

A univariate Newton-Raphson algorithm

Initialization

0. Define a starting point $\theta^{(0)} \in \mathbb{R}$ and set $k := 0$. If $f'(\theta^{(0)}) = 0$, stop! $\theta^{(0)}$ is a zero of f' . If not, proceed to iterations.

Iterations

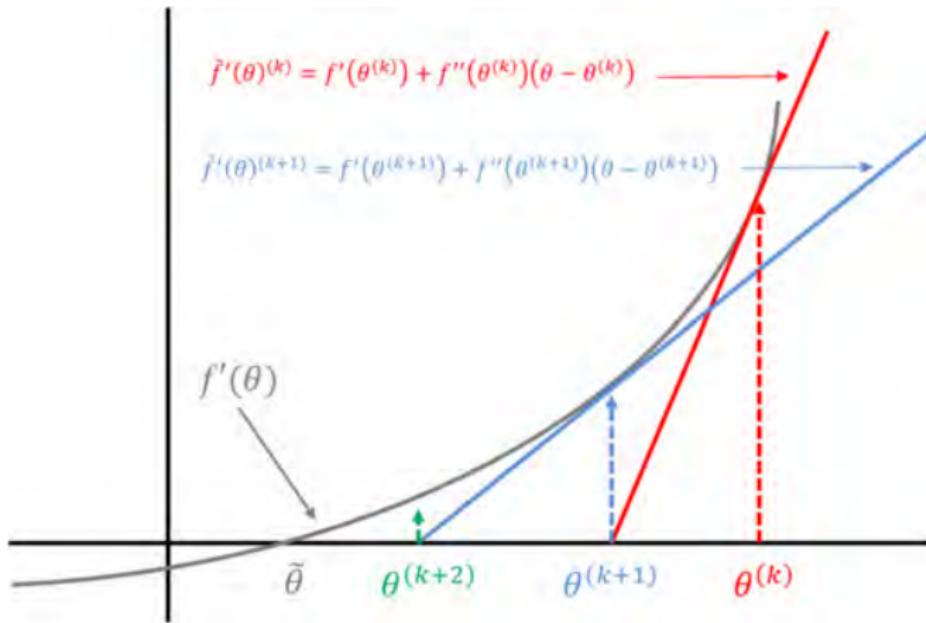
1. Set

$$\theta^{(k+1)} := \theta^{(k)} - \frac{f'(\theta^{(k)})}{f''(\theta^{(k)})}. \quad (333)$$

2. If $f'(\theta^{(k+1)}) = 0$, stop! $\theta^{(k+1)}$ is a zero of f' . If not, go to 3.
3. Set $k := k + 1$ and go to 1.

Numerical approaches

First-order Taylor approximation of $f'(\theta)$ and finding its zero.



Numerical approaches

A second-order Taylor approximation perspective

Consider

$$f(\theta) \approx \bar{f}(\theta) := f(\theta^{(k)}) + f'(\theta^{(k)})(\theta - \theta^{(k)}) + \frac{1}{2}f''(\theta^{(k)})(\theta - \theta^{(k)})^2. \quad (334)$$

Then

$$\begin{aligned}\bar{f}'(\theta) &= \frac{d}{d\theta} f(\theta^{(k)}) + \frac{d}{d\theta} \left(f'(\theta^{(k)})\theta \right) - \frac{d}{d\theta} \left(f'(\theta^{(k)})\theta^{(k)} \right) \\ &\quad + \frac{d}{d\theta} \left(\frac{1}{2}f''(\theta^{(k)})\theta^2 \right) - \frac{d}{d\theta} \left(f''(\theta^{(k)})\theta\theta^{(k)} \right) + \frac{d}{d\theta} \left(\frac{1}{2}(\theta^{(k)})^2 \right) \\ &= f'(\theta^{(k)}) + f''(\theta^{(k)})\theta - f''(\theta^{(k)})\theta^{(k)} \\ &= f'(\theta^{(k)}) + f''(\theta^{(k)})(\theta - \theta^{(k)}).\end{aligned} \quad (335)$$

Hence

$$\bar{f}'(\tilde{\theta}) = 0 \Leftrightarrow f'(\theta^{(k)}) + f''(\theta^{(k)})(\tilde{\theta} - \theta^{(k)}) = 0 \Leftrightarrow \tilde{\theta} = \theta^{(k)} - \frac{f'(\theta^{(k)})}{f''(\theta^{(k)})}. \quad (336)$$

Numerical approaches

The multivariate Newton-Raphson method

- An iterative approach to find extrema of multivariate real-valued functions.

$$f : \mathbb{R}^p \rightarrow \mathbb{R}, \theta \mapsto f(\theta)$$

- Substitution of $f'(\theta^{(k)})$ by the gradient $\nabla f(\theta^{(k)}) := \begin{pmatrix} \frac{\partial}{\partial \theta_1} f(\theta^{(k)}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} f(\theta^{(k)}) \end{pmatrix}$
- Substitution of $f''(\theta^{(k)})^{-1}$ by the inverse of the Hessian

$$H^f(\theta^{(k)}) := \begin{pmatrix} \frac{\partial^2}{\partial \theta_1 \partial \theta_1} f(\theta^{(k)}) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} f(\theta^{(k)}) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} f(\theta^{(k)}) \\ \frac{\partial^2}{\partial \theta_2 \partial \theta_1} f(\theta^{(k)}) & \frac{\partial^2}{\partial \theta_2 \partial \theta_2} f(\theta^{(k)}) & \cdots & \frac{\partial^2}{\partial \theta_2 \partial \theta_p} f(\theta^{(k)}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_p \partial \theta_1} f(\theta^{(k)}) & \frac{\partial^2}{\partial \theta_p \partial \theta_2} f(\theta^{(k)}) & \cdots & \frac{\partial^2}{\partial \theta_p \partial \theta_p} f(\theta^{(k)}) \end{pmatrix}$$

- For maximum likelihood, f is a multivariate log likelihood function.

A multivariate Newton-Raphson algorithm

Initialization

0. Define a starting point $\theta^{(0)} \in \mathbb{R}^p$ and set $k := 0$. If $\nabla f(\theta^{(0)}) = 0$, stop! $\theta^{(0)}$ is a zero of ∇f . If not, proceed to iterations.

Iterations

1. Set

$$\theta^{(k+1)} := \theta^{(k)} - \left(H^f(\theta^{(k)}) \right)^{-1} \nabla f(\theta^{(k)}). \quad (337)$$

2. If $\nabla f(\theta^{(k+1)}) = 0$, stop! $\theta^{(k+1)}$ is a zero of ∇f . If not, go to 3.
3. Set $k := k + 1$ and go to 1.

Fisher scoring

- A numerical approach for maximum likelihood estimation.
- The application of the Newton-Raphson method to log likelihood functions.
- An approach to avoid the computational burden of evaluating $(H^{\ell_n}(\theta^k))^{-1}$.

Definition (Multivariate Fisher scoring algorithm)

For a parametric statistical model with PMF or PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, $\theta \in \Theta \subset \mathbb{R}^p$, and let

$$\ell_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto \ell_n(\theta) := \ln \left(\prod_{i=1}^n p_\theta(x_i) \right). \quad (338)$$

denote the log likelihood function. Then a *multivariate Fisher scoring algorithm* is given by replacing in the multivariate Newton-Raphson method

- the objective function f by the log likelihood function ℓ_n ,
- the gradient of the objective function by the *scoring vector*

$$S_n(\theta) := \nabla \ell_n(\theta) \in \mathbb{R}^p,$$

- the Hessian matrix of the objective by the *expected Fisher information matrix*

$$J_n(\theta) := -\mathbb{E} \left(H^{\ell_n}(\theta) \right) \in \mathbb{R}^{p \times p}.$$

Remark

- Scoring vector and Fisher information are discussed in detail in later sections.

A Fisher scoring algorithm

Initialization

0. Define a starting point $\theta^{(0)} \in \mathbb{R}^p$ and set $k := 0$. If $S_n(\theta^{(0)}) = 0$, stop! $\theta^{(0)}$ is a zero of S_n . If not, proceed to iterations.

Iterations

1. Set

$$\theta^{(k+1)} := \theta^{(k)} + J_n(\theta^{(k)})^{-1} S_n(\theta^{(k)}). \quad (339)$$

2. If $S_n(\theta^{(k+1)}) = 0$, stop! $\theta^{(k+1)}$ is a zero of S_n . If not, go to 3.
3. Set $k := k + 1$ and go to 1.

Remark

- For a detailed discussion of Fisher scoring, see Osborne (1992).

Maximum likelihood estimation

- Statistical models
- Maximum likelihood estimation
- Analytical examples
- Numerical approaches
- **Exercises**

Study Questions

1. Define the notion of a point estimator.
2. Write down the definitions of the likelihood and log likelihood functions.
3. Write down the definition of the maximum likelihood estimator.
4. Write down the general maximum likelihood procedure for parametric statistical models.
5. Derive the maximum likelihood estimator for the parameter of a Bernoulli distribution.
6. Derive the maximum likelihood estimator for the parameters of a Gaussian distribution.
7. Formulate the univariate Newton-Raphson method.
8. Formulate the multivariate Newton-Raphson method.
9. Write down the Fisher scoring algorithm.

Theoretical Exercises

1. Derive the maximum likelihood estimator for the continuous uniform distribution on an interval $[0, \theta]$ (DeGroot and Schervish (2012, Example 7.5.7)).
2. Introduce the notion of moment estimators and discuss one example.
3. Introduce the notions of M- and Z-estimators and discuss one example.

Example (Continuous uniform distribution)

Let $X_1, \dots, X_n \sim U(0, \theta), \theta > 0$ be n i.i.d. continuous uniform random variables.

- We have

$$L_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, \theta \mapsto L_n(\theta) := \prod_{i=1}^n \frac{1}{\theta} = \frac{1}{\theta^n} \text{ for } 0 \leq x_i \leq \theta. \quad (340)$$

- It must hold that $\hat{\theta}_{\text{ML}} \geq x_i$ for all $i = 1, \dots, n$ and that $\hat{\theta}_{\text{ML}}$ maximizes L_n .
- L_n is a decreasing function in θ .
- We are thus interested in the smallest value of θ , such that $\theta \geq x_i, i = 1, \dots, n$.
- This value is the maximum of x_1, \dots, x_n .
- Hence

$$\hat{\theta}_{\text{ML}} = \max\{x_1, \dots, x_n\}. \quad (341)$$

Exercises

Definition (Method of moments estimator)

Let \mathcal{P} denote a parametric statistical model with PMF or PDF p_θ and let $X_1, \dots, X_n \sim p_\theta$. For $j = 1, \dots, k$, let further

$$\mu_j : \Theta \rightarrow \mathbb{R}, \theta \mapsto \mu_j(\theta) := \mathbb{E}_{p_\theta}(X^j) \quad (342)$$

denote the j th moment of $X \sim p_\theta$ and let

$$m_j = \frac{1}{n} \sum_{i=1}^n X_i^j \quad (343)$$

denote the j th sample moment of X_1, \dots, X_n . Then the *method of moments estimator (MME)* is defined as the solution for θ in terms of $m_j, j = 1, \dots, k$ of the simultaneous system of equations

$$m_j = \mu_j(\theta) \text{ for } j = 1, \dots, k. \quad (344)$$

Remarks

- Methods of moment estimators are often suboptimal.
- The Satterthwaite approximation is an example for methods of moments estimation.
- Methods of moment estimation is an old concept, predating the popularization of ML estimation

Exercises

Example (MMEs for the univariate Gaussian)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Then we have for $k := 2$ and $\theta = (\mu, \sigma^2)$ with the variance translation theorem

$$\mu_1(\theta) = \mathbb{E}(X^1) = \mu \text{ and } \mu_2(\theta) = \mathbb{E}(X^2) = \mu^2 + \sigma^2. \quad (345)$$

The set of simultaneous equations of interest for the MMEs μ_{MM} and σ_{MM}^2 of μ and σ^2 is thus given by

$$\frac{1}{n} \sum_{i=1}^n X_i = \mu_{\text{MM}} \text{ and } \frac{1}{n} \sum_{i=1}^n X_i^2 = \mu_{\text{MM}}^2 + \sigma_{\text{MM}}^2. \quad (346)$$

Solving for μ_{MM} and σ_{MM}^2 then yields

$$\mu_{\text{MM}} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n \quad (347)$$

and

$$\sigma_{\text{MM}}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \mu_{\text{MM}}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (348)$$

The MMEs for μ and σ^2 are thus given by

$$\mu_{\text{MM}} = \bar{X}_n \text{ and } \sigma_{\text{MM}}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (349)$$

Definition (M- and Z-estimators)

Let \mathcal{P} denote a parametric statistical model with PMF or PDF p_θ and let $X_1, \dots, X_n \sim p_\theta$. Then

- $\hat{\theta}_M$ is called an *M-estimator*, if it maximizes a function of the form

$$M_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto M_n(\theta) := \frac{1}{n} \sum_{i=1}^n m_\theta(X_i), \quad (350)$$

where $m_\theta : \mathcal{X} \rightarrow \mathbb{R}$ is a known function, and

- $\hat{\theta}_Z$ is called a *Z-estimator* if it minimizes a function of the form

$$\Psi_n : \Theta \subset \mathbb{R}^k \rightarrow \mathbb{R}, \theta \mapsto \Psi_n(\theta) := \frac{1}{n} \left| \sum_{i=1}^n \psi_\theta(X_i) \right| \quad (351)$$

for a known univariate vector-valued function $\psi : \mathcal{X} \rightarrow \mathbb{R}^k$.

Remarks

- M- and Z-estimators can be understood as generalized MLEs.
- m_θ and ψ_θ are not necessarily likelihood functions and likelihood function derivatives.
- M- and Z-estimators are a modern concept and popular in robust statistics.

Example (MLEs vs. M- and Z-estimators)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ and let $X_1, \dots, X_n \sim p_\theta$. Define

$$m_\theta : \mathcal{X} \rightarrow \mathbb{R}, x \mapsto m_\theta(x) := \ln p_\theta(x) \quad (352)$$

Then the M-estimator which maximizes the function

$$M_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto M_n(\theta) := \frac{1}{n} \sum_{i=1}^n m_\theta(X_i) = \frac{1}{n} \sum_{i=1}^n \ln p_\theta(X_i) \quad (353)$$

corresponds to the MLE. Similarly, define

$$\psi_\theta : \mathcal{X} \rightarrow \mathbb{R}^k, x \mapsto \nabla_\theta \ln p_\theta(x) = \left(\frac{\partial}{\partial \theta_j} \ln p_\theta(x) \right)_{1 \leq j \leq k}. \quad (354)$$

Then minimization of

$$\Psi_n : \Theta \rightarrow \mathbb{R}, \theta \mapsto \Psi_n(\theta) := \frac{1}{n} \left| \sum_{i=1}^n \psi_\theta(X_i) \right| = \frac{1}{n} \left| \sum_{i=1}^n \nabla_\theta \ln p_\theta(X_i) \right| \quad (355)$$

corresponds to identifying a zero of the log likelihood function gradient, and thus identifying a candidate value for the maximization of the log likelihood function.

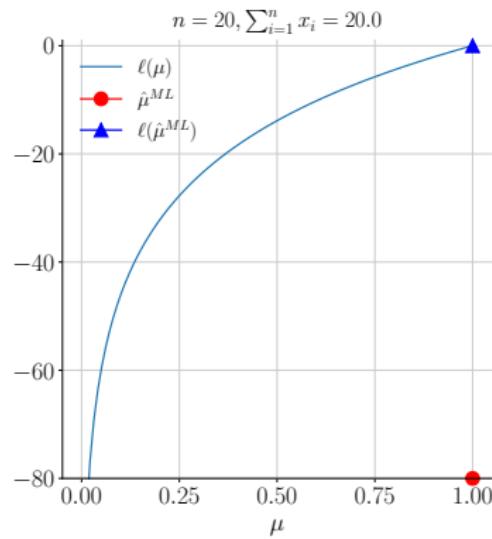
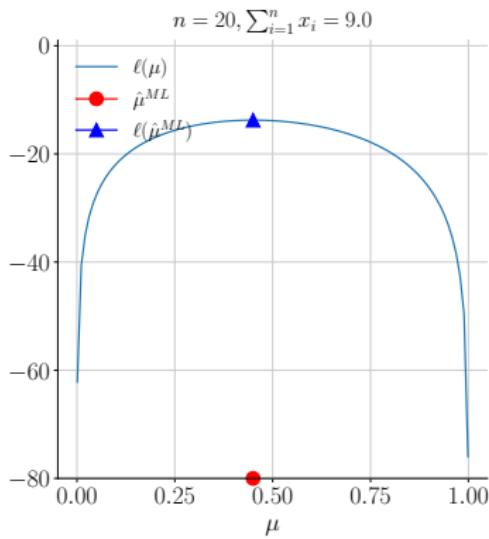
Programming Exercises

1. Let $X_1, \dots, X_n \sim \text{Bern}(\mu)$ be $n = 20$ i.i.d. Bernoulli random variables. Using an optimization routine of your choice, formulate and implement the numerical maximum likelihood estimation of μ for true, but unknown values of $\mu = 0.7$ and $\mu = 1$ based on X_1, \dots, X_n .
2. Let $X_1, \dots, X_n \sim \text{Bern}(\mu)$. For a large number n , sample the X_1, \dots, X_n and evaluate the maximum likelihood estimator $\hat{\mu}^{ML}$. Repeat this m times and create a histogram of the realized $\hat{\mu}_1^{ML}, \dots, \hat{\mu}_m^{ML}$.
3. Write a program that implements a Fisher scoring algorithm for the maximum likelihood estimation of the slope and offset parameters of a simple linear regression model, assuming a known value of the error variance parameter. Compare the results with the analytical estimation of the respective parameters.

Exercises

Programming Exercise 1

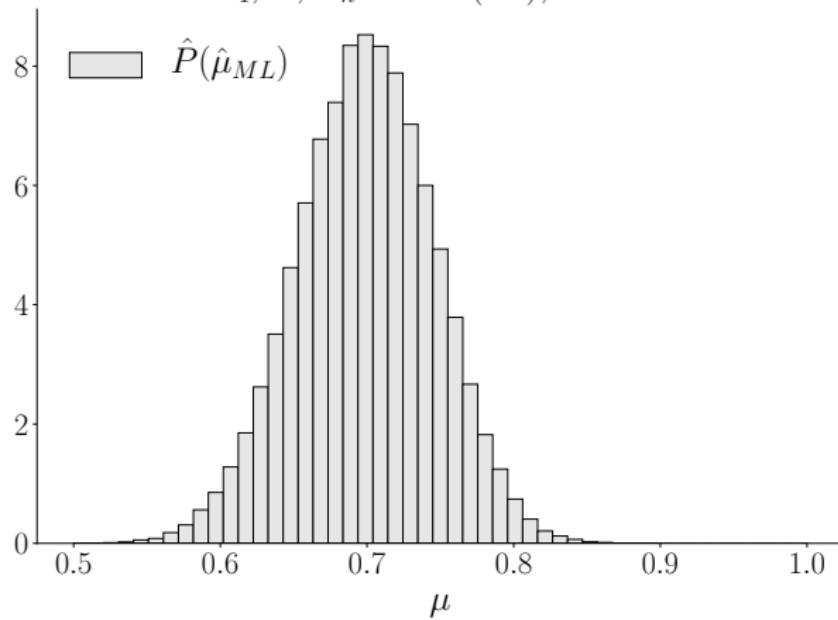
$$\ell :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto \ell(\mu) = \ln \mu \sum_{i=1}^n x_i + \ln(1 - \mu) \left(n - \sum_{i=1}^n x_i \right). \quad (356)$$



Exercises

Programming Exercise 2

$$X_1, \dots, X_n \sim \text{Bern}(0.7), n = 100$$



Programming Exercise 3

We first recall that a simple linear regression model is a GLM of the form

$$y \sim N(X\beta, \sigma^2 I_n), \text{ where } X := \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \beta := \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \sigma^2 > 0 \quad (357)$$

Here, the values x_1, \dots, x_n are referred to as *values of the independent/regressor/predictor variable* x , β_1 is referred to as *offset parameter* and β_2 is referred to as *slope parameter*. To implement a Fisher scoring algorithm for estimating $\beta \in \mathbb{R}^2$, we first consider the simple linear regression model's log likelihood function, score function, and expected Fisher information matrix.

Programming Exercise 3

We first note that the simple linear regression log likelihood function for known error variance σ^2 is given by

$$\ell_n : \mathbb{R}^2 \rightarrow \mathbb{R}, \beta \mapsto \ell(\beta) := -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta), \quad (358)$$

where we can rewrite the value of the log likelihood function as follows

$$\begin{aligned}\ell_n(\beta) &= \ell_n(\beta_1, \beta_2) \\&= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \\&= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (X\beta)_i)^2 \\&= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\beta_1 + x_i\beta_2))^2 \\&= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_1 - x_i\beta_2)^2\end{aligned}\quad (359)$$

Programming Exercise 3

The score function, i.e., gradient of the log likelihood function with respect to β , thus evaluates to

$$\begin{aligned} S_n(\beta) &= \nabla \ell_n(\beta) \\ &= \begin{pmatrix} \frac{\partial}{\partial \beta_1} \ell_n(\beta_1, \beta_2) \\ \frac{\partial}{\partial \beta_2} \ell_n(\beta_1, \beta_2) \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{\partial}{\partial \beta_1} (y_i - \beta_1 - x_i \beta_2)^2 \\ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{\partial}{\partial \beta_2} (y_i - \beta_1 - x_i \beta_2)^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_1 - x_i \beta_2) \\ \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_1 - x_i \beta_2) x_i \end{pmatrix} \\ &= \frac{1}{\sigma^2} \left((y - X\beta)^T X \right)^T \end{aligned} \tag{360}$$

Programming Exercise 3

The Hessian of the log likelihood function with respect to β evaluates to

$$\begin{aligned}
 H^{\ell_n}(\beta) &= \begin{pmatrix} \frac{\partial^2}{\partial \beta_1^2} \ell(\beta_1, \beta_2) & \frac{\partial^2}{\partial \beta_1 \partial \beta_2} \ell(\beta_1, \beta_2) \\ \frac{\partial^2}{\partial \beta_2 \partial \beta_1} \ell(\beta_1, \beta_2) & \frac{\partial^2}{\partial \beta_2^2} \ell(\beta_1, \beta_2) \end{pmatrix} \\
 &= \frac{1}{\sigma^2} \begin{pmatrix} \frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - \beta_1 - x_i \beta_2) & \frac{\partial}{\partial \beta_2} \sum_{i=1}^n (y_i - \beta_1 - x_i \beta_2) \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^n (y_i - \beta_1 - x_i \beta_2) x_i & \frac{\partial}{\partial \beta_2} \sum_{i=1}^n (y_i - \beta_1 - x_i \beta_2) x_i \end{pmatrix} \quad (361) \\
 &= -\frac{1}{\sigma^2} \begin{pmatrix} \frac{\partial}{\partial \beta_1} \sum_{i=1}^n 1 & \frac{\partial}{\partial \beta_2} \sum_{i=1}^n x_i \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^n x_i & \frac{\partial}{\partial \beta_2} \sum_{i=1}^n x_i^2 \end{pmatrix} \\
 &= -\frac{1}{\sigma^2} X^T X
 \end{aligned}$$

such that the expected Fisher information matrix evaluates to

$$J_n(\beta) = -\mathbb{E}(H^{\ell_n}(\beta)) = \frac{1}{\sigma^2} X^T X. \quad (362)$$

Programming Exercise 3

In summary, a Fisher scoring algorithm for the estimation of the parameters of a simple linear regression model thus takes the form

Initialization

0. Define a starting point $\beta^{(0)} \in \mathbb{R}^2$ and set $k := 0$. If $S_n(\beta^{(0)}) = 0$, stop! $\beta^{(0)}$ is a zero of S_n . If not, proceed to iterations.

Iterations

1. Set

$$\beta^{(k+1)} := \beta^{(k)} + (X^T X)^{-1} \left((y - X\beta^{(k)})^T X \right)^T \quad (363)$$

2. If $S_n(\beta^{(k+1)}) = 0$, stop! $\beta^{(k+1)}$ is a zero of S_n . If not, go to 3.
3. Set $k := k + 1$ and go to 1.

Exercises

Programming Exercise 3

```
# assignment exercise 3: Python script for simple linear regression
# linear regression
# least squares estimation
n = 11 # number of data points
r0 = np.ones([n,1]) # offset regression
r1 = np.array([np.arange(0,n)]).T # slope regression
X = np.hstack((r0,r1)) # design matrix - concatenation
p = X.shape[1] # number of columns = number of parameters
beta = np.array([[1],[2]]) # true, but unknown, Beta parameters
mu = np.squeeze(X @ beta) # parameter
sigsqr = 1e-2 # variance parameter
Sigma = sigsqr*np.eye(n) # covariance covariance matrix of the Gaussian process

# data generation
y = np.array([rv.multivariate_normal.rvs(mu, Sigma)]).T # data samples

# analytical parameter estimation
beta_hat = la.pinv(X) @ y # linearized least squares

# numerical parameter estimation
K = 3 # maximum number of iterations
beta_k = np.full([p,K], np.nan) # beta estimates (current array)
beta_k[:,0] = np.random.rand(p) # initial values
for k in range(K-1):
    # the theory of doing a gradient step in both Python and R
    beta_k[:,k+1] = np.squeeze(np.array([beta_k[:,k]]).T + la.inv(X.T @ X) # gradient descent
        @ np.transpose(np.transpose(y-np.array([X @ beta_k[:,k]]).T) @ X)) # gradient descent

print('Analytical estimates')
print(beta_hat)

print('Fisher scoring estimates')
print(np.array([beta_k[:, -1]]).T)
```

(9) Finite-sample estimator properties

Bibliographic remarks

The material presented in this section is based on Wasserman (2004, Chapter 9), Held and Sabanés Bové (2014, Sections 3.1,3.2,4.1,4.2), and DeGroot and Schervish (2012, Section 8.8).

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- Score function and Fisher information
- Cramér-Rao bound
- Mean squared error
- Exercises

Finite-sample estimator properties

- **Error, bias, and unbiasedness**
- Variance and standard error
- Score function and Fisher information
- Cramér-Rao bound
- Mean squared error
- Exercises

Definition (Error, bias, and unbiasedness)

Let \mathcal{P} denote a parametric statistical model with PMF/PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ denote an estimator for θ .

- The *error* of $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n - \theta. \quad (364)$$

Error, bias, and unbiasedness

Definition (Error, bias, and unbiasedness)

Let \mathcal{P} denote a parametric statistical model with PMF/PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ denote an estimator for θ .

- The *error* of $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n - \theta. \quad (364)$$

- The *bias* of $\hat{\theta}_n$ is defined as

$$B(\hat{\theta}_n) := \mathbb{E}_\theta(\hat{\theta}_n) - \theta. \quad (365)$$

- $\hat{\theta}_n$ is called *unbiased*, if

$$B(\hat{\theta}_n) = 0 \Leftrightarrow \mathbb{E}_\theta(\hat{\theta}_n) = \theta \text{ for all } \theta \in \Theta, n \in \mathbb{N}. \quad (366)$$

Otherwise, $\hat{\theta}_n$ is called *biased*.

Remarks

- The error depends on the realization of X_1, \dots, X_n .
- The bias is the expected error over many realizations of X_1, \dots, X_n .
- \mathbb{E}_θ means expectation with respect to p_θ .

Theorem (Unbiasedness of sample mean and sample variance)

Let $X_1, \dots, X_n \sim p_\theta$ be a random sample of a parametric statistical model \mathcal{P} with expectation $\mu := \mathbb{E}(X_i)$ and variance $\sigma^2 := \mathbb{V}(X_i)$ for $i = 1, \dots, n$. Then

- The *sample mean*

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i \tag{367}$$

is an unbiased estimator of the expectation μ , and

- the *sample variance*

$$S^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \tag{368}$$

is an unbiased estimator of the variance σ^2 .

Error, bias, and unbiasedness

Proof

For ease of notation, we set $\mathbb{E} := \mathbb{E}_\theta$ and $\mathbb{V} := \mathbb{V}_\theta$. With the linearity of expectations, we then have

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu,$$

which proves the unbiasedness of the sample mean as an estimator of the expectation.

To show the unbiasedness of the sample variance, we first note that we have

$$\mathbb{V}(\bar{X}) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

We further note that basic algebraic manipulations yield

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu - \bar{X} + \mu)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Error, bias, and unbiasedness

Proof (cont.)

We then have

$$\begin{aligned}\mathbb{E}((n-1)S^2) &= \mathbb{E}\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= \mathbb{E}\left(\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right) \\ &= \sum_{i=1}^n \mathbb{E}((X_i - \mu)^2) - n\mathbb{E}((\bar{X} - \mu)^2) \\ &= n\mathbb{V}(X_i) - n\mathbb{V}(\bar{X}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= n\sigma^2 - \sigma^2 \\ &= (n-1)\sigma^2\end{aligned}$$

Finally, we have

$$\mathbb{E}(S^2) = \mathbb{E}\left(\frac{1}{n-1}(n-1)S^2\right) = \frac{1}{n-1}\mathbb{E}((n-1)S^2) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2,$$

which shows the unbiasedness of the sample variance as an estimator of the variance.

□

Theorem (Biasedness of the sample standard deviation)

Let X_1, \dots, X_n be a random sample of a parametric statistical model \mathcal{P} with variance $\sigma^2 := \mathbb{V}(X_i)$ and standard deviation $\sigma := \sqrt{\mathbb{V}(X_i)}$ for $i = 1, \dots, n$. Then the *sample standard deviation*

$$S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (369)$$

is a biased estimator of the standard deviation σ .

Proof

We first note that $\sqrt{\cdot}$ is a strictly concave function and $\sigma^2 > 0$. Then, with Jensen's inequality $\mathbb{E}(f(X)) < f(\mathbb{E}(X))$ for strictly concave functions, we have

$$\mathbb{E}(S) = \mathbb{E}\left(\sqrt{S^2}\right) < \sqrt{\mathbb{E}(S^2)} = \sqrt{\sigma^2} = \sigma. \quad (370)$$

□

Remark

- Nonlinear transformations of unbiased estimators are often biased.

Finite-sample estimator properties

- Error, bias, and unbiasedness
- **Variance and standard error**
- Score function and Fisher information
- Cramér-Rao bound
- Mean squared error
- Exercises

Definition (Variance and standard error)

Let \mathcal{P} denote a parametric statistical model with PMF/PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ denote an estimator for θ .

- The *variance* of $\hat{\theta}_n$ is defined as

$$\mathbb{V}_\theta(\hat{\theta}_n) := \mathbb{E}_\theta \left((\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2 \right). \quad (371)$$

- The *standard error* of $\hat{\theta}_n$ is defined as

$$\text{SE}(\hat{\theta}_n) := \sqrt{\mathbb{V}_\theta(\hat{\theta}_n)} \quad (372)$$

Remark

- The estimator variance is the variance of the random variable $\hat{\theta}_n$.
- The estimator standard error is the standard deviation of $\hat{\theta}_n$.
- All expectations, variances and the standard error are with respect to p_θ .

Theorem (Standard error of the sample mean)

Let $X_1, \dots, X_n \sim p_\theta$ be a random sample of a parametric statistical model \mathcal{P} with expectation $\mu := \mathbb{E}(X_i)$ and variance $\sigma^2 := \mathbb{V}(X_i)$ for $i = 1, \dots, n$. Then the *standard error of the sample mean*, also referred to as *standard error of the mean* is given by

$$\text{SE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}. \quad (373)$$

Proof

By definition and with $\mathbb{V}_\theta(\bar{X}) = \sigma^2/n$, we have

$$\text{SE}(\bar{X}) = \sqrt{\mathbb{V}_\theta(\bar{X})} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}. \quad (374)$$

□

Remark

- A biased estimator for the standard error of the sample mean is given in terms of the sample standard deviation by $\hat{\text{SE}}(\bar{X}) = S/\sqrt{n}$.

Example (Standard error of the Bernoulli parameter MLE)

Let $X_1, \dots, X_n \sim \text{Bern}(\mu)$ and let $\hat{\mu}_n^{\text{ML}}$ denote the maximum likelihood estimator for μ . Then the standard error of $\hat{\mu}_n^{\text{ML}}$ is

$$\text{SE}(\hat{\mu}_n^{\text{ML}}) = \sqrt{\frac{\mu(1-\mu)}{n}}. \quad (375)$$

Proof

We have

$$\text{SE}(\hat{\mu}_n^{\text{ML}}) = \sqrt{\mathbb{V}(\hat{\mu}_n^{\text{ML}})} = \sqrt{\mathbb{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i)} = \sqrt{\frac{n\mu(1-\mu)}{n^2}} = \sqrt{\frac{\mu(1-\mu)}{n}}, \quad (376)$$

where the third equation follows with the independence of the $X_i, i = 1, \dots, n$ and the fourth equation follows with the variance $\mathbb{V}(X) = \mu(1-\mu)$ of a Bernoulli distributed random variable (cf. Lecture (5)).

□

Remark

- An estimator for the standard error is given by $\hat{\text{SE}}(\hat{\mu}_n^{\text{ML}}) = \sqrt{\frac{\hat{\mu}_n^{\text{ML}}(1-\hat{\mu}_n^{\text{ML}})}{n}}$.

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- **Score function and Fisher information**
- Cramér-Rao bound
- Mean squared error
- Exercises

Definition (Score function and Fisher information)

Let \mathcal{P} denote a statistical model with PMF or PDF p_θ with scalar parameter $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$ denote a random sample from this model, and let ℓ_n denote the respective log likelihood function.

- The first derivative of the log likelihood function ℓ_n is referred to as the *score function*,

$$S_n(\theta) := \frac{d}{d\theta} \ell_n(\theta). \quad (377)$$

For $n = 1$, we write $S(\theta) := S_1(\theta)$ and call $S(\theta)$ the *score function of a random variable*.

- The negative second derivative of the log likelihood function ℓ_n is referred to as *Fisher information* of the random sample X_1, \dots, X_n ,

$$I_n(\theta) := -\frac{d^2}{d\theta^2} \ell_n(\theta). \quad (378)$$

For $n = 1$, we write $I(\theta) := I_1(\theta)$ and call $I(\theta)$ the *Fisher information of a random variable*.

Definition (Expected and observed Fisher information)

Let \mathcal{P} denote a statistical model with PMF or PDF p_θ with scalar parameter $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$ denote a random sample from this model, let ℓ_n denote the respective log likelihood function, and let $\hat{\theta}_n^{\text{ML}}$ denote a maximum likelihood estimator of θ .

- The *observed Fisher information* of a random sample is defined as

$$I_n(\hat{\theta}_n^{\text{ML}}) := -\frac{d^2}{d\theta^2} \ell_n(\hat{\theta}_n^{\text{ML}}), \quad (379)$$

i.e., the observed Fisher information of a random sample is the Fisher information at the location of the maximum likelihood estimate $\hat{\theta}_n^{\text{ML}}$.

- The *expected Fisher information* of a random sample is defined as

$$J_n(\theta) := \mathbb{E}_\theta(I_n(\theta)). \quad (380)$$

For $n = 1$, we write $J(\theta) := J_1(\theta)$ and refer to $J(\theta)$ as the *expected Fisher information of a random variable*.

Theorem (Additivity of Fisher information)

Let $X_1, \dots, X_n \sim p_\theta$ denote a random sample, let ℓ_n denote the respective log likelihood function, and let $I_n(\theta)$ and $J_n(\theta)$ denote the Fisher information and the expected Fisher information of the random sample, respectively. Then

$$I_n(\theta) = nI_1(\theta) \text{ and } J_n(\theta) = nJ_1(\theta). \quad (381)$$

Remarks

- To evaluate $I_n(\theta)$ or $J_n(\theta)$ it suffices to evaluate $I(\theta)$ or $J(\theta)$.
- The observed Fisher information additivity is implied by the additivity of $I_n(\theta)$.

Score function and Fisher information

Proof

We show the result for the expected Fisher information, the result for the Fisher information is implied. By definition and with the linearity of differentiation and expectations, we have

$$\begin{aligned} J_n(\theta) &= \mathbb{E} \left(-\frac{d^2}{d\theta^2} \ell_n(\theta) \right) \\ &= \mathbb{E} \left(-\frac{d^2}{d\theta^2} \ln \left(\prod_{i=1}^n p_\theta(x_i) \right) \right) \\ &= \mathbb{E} \left(-\frac{d^2}{d\theta^2} \sum_{i=1}^n \ln p_\theta(x_i) \right) \\ &= \mathbb{E} \left(-\frac{d^2}{d\theta^2} \sum_{i=1}^n \ln p_\theta(x_1) \right) \\ &= \mathbb{E} \left(-\frac{d^2}{d\theta^2} \ell_1(\theta) n \right) \\ &= n \mathbb{E} \left(-\frac{d^2}{d\theta^2} \ell_1(\theta) \right) \\ &= n J_n(\theta). \end{aligned} \tag{382}$$

□

Score function and Fisher information

Example (Expectation parameter of a Bernoulli sample)

Let $X_1, \dots, X_n \sim \text{Bern}(\mu)$ for $\mu \in]0, 1[$. Then

- The score function of the random sample evaluates to

$$S_n :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto S_n(\mu) := \frac{1}{\mu} \sum_{i=1}^n x_i - \frac{1}{1-\mu} \left(n - \sum_{i=1}^n x_i \right). \quad (383)$$

- The Fisher information of the random sample evaluates to

$$I_n :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto I_n(\mu) := \frac{x}{\mu^2} + \frac{(1-x)^2}{1-\mu}. \quad (384)$$

- The observed Fisher information of the random sample evaluates to

$$I_n :]0, 1[\rightarrow \mathbb{R}, \hat{\mu}_n^{\text{ML}} \mapsto I_n(\hat{\mu}_n^{\text{ML}}) := \frac{x}{\hat{\mu}_n^{\text{ML}}{}^2} + \frac{(1-x)}{1-\hat{\mu}_n^{\text{ML}}}. \quad (385)$$

- The expected Fisher information of the random sample evaluates to

$$J_n :]0, 1[\rightarrow \mathbb{R}, \mu \mapsto J_n(\mu) := \frac{n}{\mu(1-\mu)}. \quad (386)$$

Score function and Fisher information

Example (Expectation parameter of a Bernoulli sample)

Proof

The form of the score function was shown in Lecture (8) in the context of the maximum likelihood estimation of μ . We next consider the Fisher information of a single Bernoulli random variable X and find

$$\begin{aligned} I(\mu) &:= -\frac{d^2}{d\mu^2} \ell_1(\mu) \\ &= -\frac{d^2}{d\mu^2} \ln p_\mu(x) \\ &= -\frac{d^2}{d\mu^2} (x \ln \mu + (1-x) \ln(1-\mu)) \\ &= -\frac{d}{d\mu} \left(\frac{d}{d\mu} (x \ln \mu + (1-x) \ln(1-\mu)) \right) \\ &= -\frac{d}{d\mu} \left(\frac{x}{\mu} + \frac{(1-x)}{1-\mu} \right) \\ &= -\left(-\frac{x}{\mu^2} - \frac{(1-x)^2}{1-\mu} \right) \\ &= \frac{x}{\mu^2} + \frac{(1-x)^2}{1-\mu}. \end{aligned} \tag{387}$$

Example (Expectation of a Bernoulli distribution)

Proof

Furthermore, the expected Fisher information of the random variable X is given by

$$\begin{aligned} J(\mu) &= \mathbb{E}_\mu(I(\mu)) \\ &= \mathbb{E}_\mu \left(\frac{X}{\mu^2} + \frac{(1-X)^2}{1-\mu} \right) \\ &= \frac{\mathbb{E}_\mu(X)}{\mu^2} + \frac{(1-\mathbb{E}_\mu(X))^2}{1-\mu} \\ &= \frac{\mu}{\mu^2} + \frac{(1-\mu)^2}{1-\mu} \\ &= \frac{1}{\mu(1-\mu)}. \end{aligned} \tag{388}$$

With the additivity property of the Fisher information and the definition of the observed Fisher information, it then follows immediately, that

$$I_n(\mu) = \frac{nx}{\mu^2} + \frac{n(1-x)^2}{1-\mu}, \text{ and } J_n(\mu) = \frac{n}{\mu(1-\mu)}. \tag{389}$$

□

Example (Expectation parameter of a Gaussian sample)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and assume that σ^2 is known. Then

- The score function of the random sample evaluates to

$$S_n : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto S_n(\mu) := \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu). \quad (390)$$

- The Fisher information of the random sample evaluates to

$$I_n : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto I_n(\mu) := \frac{n}{\sigma^2}. \quad (391)$$

- The observed Fisher information of the random sample evaluates to

$$I_n(\hat{\mu}_n^{\text{ML}}) = \frac{n}{\sigma^2}. \quad (392)$$

- The expected Fisher information of the random sample evaluates to

$$J_n : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto J_n(\mu) := \frac{n}{\sigma^2}. \quad (393)$$

Score function and Fisher information

Example (Expectation parameter of a Gaussian sample)

Proof

The log likelihood function of a Gaussian random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ for a known variance parameter σ^2 is given by

$$\ell_n : \mathbb{R} \rightarrow \mathbb{R}, \mu \mapsto \ell_n(\mu) := -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (394)$$

The score function thus evaluates to

$$S_n(\mu) = \frac{d}{d\mu} \ell_n(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \quad (395)$$

The Fisher information of the random sample X_1, \dots, X_n evaluates to

$$I_n(\mu) = -\frac{d^2}{d\mu^2} \ell_n(\mu) = -\frac{d}{d\mu} S_n(\mu) = -\frac{1}{\sigma^2} \frac{d}{d\mu} \left(\sum_{i=1}^n x_i - n\mu \right) = \frac{n}{\sigma^2}. \quad (396)$$

The observed Fisher information corresponds to the Fisher information evaluated at the location of the parameter of interest's ML estimator. Finally, the expected Fisher information evaluates to

$$J_n(\mu) = \mathbb{E}_\mu(I_n(\mu)) = \mathbb{E}_\mu \left(\frac{n}{\sigma^2} \right) = \frac{n}{\sigma^2}. \quad (397)$$

Example (Variance parameter of a Gaussian sample)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and assume that μ is known. Then

- the score function of the random sample evaluates to

$$S_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto S_n(\sigma^2) := -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \quad (398)$$

- the Fisher information of the random sample evaluates to

$$I_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto I_n(\sigma^2) := \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^4} \quad (399)$$

- the observed Fisher information of the random sample evaluates to

$$I_n(\hat{\sigma}_{\text{ML}}^2) = \frac{n}{2\hat{\sigma}_{\text{ML}}^4} \quad (400)$$

- the expected Fisher information of the random sample evaluates to

$$J_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto J_n(\sigma^2) := \frac{n}{2\sigma^4}. \quad (401)$$

Example (Variance parameter of a Gaussian sample)

Proof

The log likelihood function of a Gaussian random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ for a known expectation parameter μ is given by

$$\ell_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \sigma^2 \mapsto \ell_n(\sigma^2) := -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (402)$$

The score function thus evaluates to

$$S_n(\sigma^2) = \frac{d}{d\sigma^2} \ell_n(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2. \quad (403)$$

The Fisher information of the random sample X_1, \dots, X_n evaluates to

$$\begin{aligned} I_n(\sigma^2) &= -\frac{d}{d\sigma^2} S_n(\sigma^2) = -\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2 \right) \\ &= \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^4}. \end{aligned} \quad (404)$$

Example (Variance parameter of a Gaussian sample)

Proof (cont.)

The observed Fisher information corresponds to the Fisher information evaluated at the location of the parameter of interest's ML estimator. We thus have

$$\begin{aligned} I_n(\hat{\sigma}_{\text{ML}}^2) &= \frac{\sum_{i=1}^n (x_i - \mu)^2}{(\hat{\sigma}_{\text{ML}}^2)^3} - \frac{n}{2(\hat{\sigma}_{\text{ML}}^2)^2} \\ &= \frac{\sum_{i=1}^n (x_i - \mu)^2}{\frac{1}{n^3} (\sum_{i=1}^n (x_i - \mu)^2)^3} - \frac{n}{2(\hat{\sigma}_{\text{ML}}^2)^2} \\ &= \frac{1}{\frac{1}{n^3} (\sum_{i=1}^n (x_i - \mu)^2)^2} - \frac{n}{2(\hat{\sigma}_{\text{ML}}^2)^2} \\ &= \frac{n}{(\hat{\sigma}_{\text{ML}}^2)^2} - \frac{n}{2(\hat{\sigma}_{\text{ML}}^2)^2} \\ &= \frac{n}{2(\hat{\sigma}_{\text{ML}}^2)^2} \\ &= \frac{n}{2\hat{\sigma}_{\text{ML}}^4}. \end{aligned} \tag{405}$$

Score function and Fisher information

Example (Variance parameter of a Gaussian sample)

Proof (cont.)

The expected Fisher information evaluates to

$$\begin{aligned} J_n(\sigma^2) &= \mathbb{E}_{\sigma^2}(I_n(\sigma^2)) \\ &= \mathbb{E}_{\sigma^2} \left(\frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^4} \right) \\ &= \frac{1}{\sigma^6} \sum_{i=1}^n \mathbb{E}_{\sigma^2} ((x_i - \mu)^2) - \frac{n}{2\sigma^4} \\ &= \frac{1}{\sigma^6} \sum_{i=1}^n \sigma^2 - \frac{n}{2\sigma^4} \\ &= \frac{n\sigma^2}{\sigma^6} - \frac{n}{2\sigma^4} \\ &= \frac{n}{\sigma^4} - \frac{n}{2\sigma^4} \\ &= \frac{n}{2\sigma^4}. \end{aligned} \tag{406}$$

Theorem (Expectation and variance of the score function)

The expectation the score function of a random variable is given by

$$\mathbb{E}_\theta(S(\theta)) = 0 \quad (407)$$

and the variance of the score function of a random variable is given by

$$\mathbb{V}_\theta(S(\theta)) = J(\theta). \quad (408)$$

Remarks

- The expectation of the derivative of the log likelihood function is zero.
- The expected Fisher information equals the variance of the score function.

Score function and Fisher information

Proof

We only consider the case that p_θ is a PDF and first show $\mathbb{E}_\theta(S(\theta)) = 0$:

$$\begin{aligned}\mathbb{E}_\theta(S(\theta)) &= \int S(\theta)p_\theta(x) dx \\&= \int \frac{d}{d\theta} \ell(\theta)p_\theta(x) dx \\&= \int \frac{d}{d\theta} \ln L(\theta)p_\theta(x) dx \\&= \int \frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta)p_\theta(x) dx \\&= \int \frac{1}{p_\theta(x)} \frac{d}{d\theta} L(\theta)p_\theta(x) dx \\&= \int \frac{d}{d\theta} L(\theta) dx \\&= \frac{d}{d\theta} \int p_\theta(x) dx \\&= \frac{d}{d\theta} 1 \\&= 0.\end{aligned}\tag{409}$$

With the definition of the variance, it immediately follows that $\mathbb{V}_\theta(S(\theta)) = \mathbb{E}_\theta(S(\theta)^2)$.

Score function and Fisher information

Proof (cont.)

We next show that $J(\theta) = \mathbb{E}_\theta (S(\theta)^2)$ and thus $\mathbb{V}_\theta(S(\theta)) = J(\theta)$:

$$\begin{aligned} J(\theta) &= \mathbb{E}_\theta \left(-\frac{d^2}{d\theta^2} \ln L(\theta) \right) \\ &= \mathbb{E}_\theta \left(-\frac{d}{d\theta} \frac{\frac{d}{d\theta} L(\theta)}{L(\theta)} \right) \\ &= \mathbb{E}_\theta \left(-\frac{\frac{d^2}{d\theta^2} L(\theta) L(\theta) - \frac{d}{d\theta} L(\theta) \frac{d}{d\theta} L(\theta)}{L(\theta) L(\theta)} \right) \\ &= -\mathbb{E}_\theta \left(\frac{\frac{d^2}{d\theta^2} L(\theta)}{L(\theta)} \right) + \mathbb{E}_\theta \left(\frac{\left(\frac{d}{d\theta} L(\theta) \right)^2}{(L(\theta))^2} \right) \tag{410} \\ &= - \int \frac{\frac{d^2}{d\theta^2} L(\theta)}{L(\theta)} p_\theta(x) dx + \int \frac{\left(\frac{d}{d\theta} L(\theta) \right)^2}{(L(\theta))^2} p_\theta(x) dx \\ &= - \frac{d^2}{d\theta^2} \int p_\theta(x) dx + \int \left(\frac{1}{L(\theta)} \frac{d}{d\theta} L(\theta) \right)^2 p_\theta(x) dx \\ &= - \frac{d^2}{d\theta^2} 1 + \int \left(\frac{d}{d\theta} \ln L(\theta) \right)^2 p_\theta(x) dx \\ &= \mathbb{E}_\theta (S(\theta)^2). \end{aligned}$$

□

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- Score function and Fisher information
- **Cramér-Rao bound**
- Mean squared error
- Exercises

Cramér-Rao bound

- The smaller the variance of an estimator, the better.
- The Cramér-Rao bound is a lower variance bound for unbiased estimators.
- An unbiased estimator with variance equal to the Cramér-Rao bound has minimal variance among all unbiased estimators and is “optimal” in this sense.
- The Cramér-Rao bound rests on the notion of expected Fisher information.
- Most results in this respect hold only under the “Fisher regularity conditions”.

Fisher regularity conditions

1. Θ is an open interval, i.e., θ must not be at a parameter space boundary.
2. The support of p_θ does not depend on θ .
3. PMFs or PDFs indexed by θ are distinct.
4. The likelihood function is twice continuously differentiable.
5. Integration and differentiation can be exchanged.

Theorem (Cramér-Rao bound)

Let \mathcal{P} denote a parametric statistical model with PMF or PDF p_θ and let $\hat{\theta}$ denote an unbiased estimator for $g(\theta)$. Then

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{\left(\frac{d}{d\theta}g(\theta)\right)^2}{J(\theta)}. \quad (411)$$

In particular, for $g(\theta) := \theta$ and thus $\left(\frac{d}{d\theta}g(\theta)\right)^2 = 1$.

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{1}{J(\theta)}. \quad (412)$$

The right-hand sides of the above are referred to as *Cramér-Rao lower bounds*.

Remarks

- In words, the variance of an unbiased estimator $\hat{\theta}$ for θ is larger or equal to the reciprocal expected Fisher information $J(\theta)$.
- If $\mathbb{V}_\theta(\hat{\theta}) = \frac{1}{J(\theta)}$, the variance of the estimator is as low as possible.
- A highly Fisher-informative unbiased estimator has a low variance lower bound.

Cramér-Rao bound

Proof

We first note that for the random variables $S(\theta)$ and $\hat{\theta}$ we have with the correlation inequality and the fact that $\mathbb{V}_\theta(S(\theta)) = J(\theta)$

$$\begin{aligned} \frac{\mathbb{C}_\theta(S(\theta), \hat{\theta})^2}{\mathbb{V}_\theta(S(\theta))\mathbb{V}_\theta(\hat{\theta})} &\leq 1 \\ \Leftrightarrow \mathbb{V}_\theta(\hat{\theta}) &\geq \frac{\mathbb{C}_\theta(S(\theta), \hat{\theta})^2}{J(\theta)}. \end{aligned} \tag{413}$$

With the translation theorem for covariances, $\mathbb{E}_\theta(S(\theta)) = 0$, and the unbiasedness of $\hat{\theta}$, we then have

$$\mathbb{C}_\theta(S(\theta), \hat{\theta}) = \frac{d}{d\theta}g(\theta) \tag{414}$$

as shown below. Thus

$$\mathbb{V}_\theta(\hat{\theta}) \geq \frac{\left(\frac{d}{d\theta}g(\theta)\right)^2}{J(\theta)}. \tag{415}$$

Cramér-Rao bound

Proof (cont.)

It remains to be shown that $\mathbb{C}_\theta(S(\theta), \hat{\theta}) = \frac{d}{d\theta} g(\theta)$. To this end, we have

$$\begin{aligned}\mathbb{C}_\theta(S(\theta), \hat{\theta}) &= \mathbb{E}_\theta(S(\theta)\hat{\theta}) - \mathbb{E}_\theta(S(\theta))\mathbb{E}_\theta(\hat{\theta}) \\ &= \mathbb{E}_\theta(S(\theta)\hat{\theta}) \\ &= \int S(\theta) \hat{\theta} p_\theta(x) dx \\ &= \int \frac{d}{d\theta} \ln L(\theta) \hat{\theta} p_\theta(x) dx \\ &= \int \frac{\frac{d}{d\theta} L(\theta)}{L(\theta)} \hat{\theta} p_\theta(x) dx \\ &= \int \frac{\frac{d}{d\theta} L(\theta)}{p_\theta(x)} \hat{\theta} p_\theta(x) dx \\ &= \int \frac{d}{d\theta} L(\theta) \hat{\theta} dx \\ &= \frac{d}{d\theta} \int L(\theta) \hat{\theta} dx \\ &= \frac{d}{d\theta} \int \hat{\theta} p_\theta(x) dx \\ &= \frac{d}{d\theta} \mathbb{E}_\theta(\hat{\theta}) \\ &= \frac{d}{d\theta} g(\theta)\end{aligned}\tag{416}$$

□

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- Score function and Fisher information
- Cramér-Rao bound
- **Mean squared error**
- Exercises

Definition (Mean squared error)

Let \mathcal{P} denote a parametric statistical model with PMF/PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ denote an estimator for θ . Then the *mean squared error* of $\hat{\theta}_n$ is defined as

$$\text{MSE}(\hat{\theta}_n) := \mathbb{E}_\theta \left((\hat{\theta}_n - \theta)^2 \right). \quad (417)$$

Remarks

- The MSE is the expected squared deviation of $\hat{\theta}_n$ from θ .
- The variance is the expected squared deviation of $\hat{\theta}_n$ from $\mathbb{E}_\theta(\hat{\theta}_n)$.
- The expectation $\mathbb{E}_\theta(\hat{\theta}_n)$ may or may not coincide with θ .

Theorem (Mean squared error decomposition)

Let \mathcal{P} denote a parametric statistical model with PMF/PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ denote an estimator for θ . Then

$$\text{MSE}(\hat{\theta}_n) = \text{B}(\hat{\theta}_n)^2 + \mathbb{V}_\theta(\hat{\theta}_n) \quad (418)$$

Remarks

- Mean squared error = Bias² + Variance.
- The MSE can be used as a bias-variance trade-off criterion.
- Small biases may be favoured over large variances.

Mean squared error

Proof

Let $\bar{\theta}_n := \mathbb{E}_\theta(\hat{\theta}_n)$. Then

$$\begin{aligned}\mathbb{E}_\theta((\hat{\theta}_n - \theta)^2) &= \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2) \\&= \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\hat{\theta}_n - \bar{\theta}_n)(\bar{\theta}_n - \theta) + (\bar{\theta}_n - \theta)^2) \\&= \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)^2) + 2\mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)(\bar{\theta}_n - \theta)) + \mathbb{E}_\theta((\bar{\theta}_n - \theta)^2) \\&= \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)^2) + 2\mathbb{E}_\theta(\hat{\theta}_n\bar{\theta}_n - \hat{\theta}_n\theta - \bar{\theta}_n\bar{\theta}_n + \bar{\theta}_n\theta) + \mathbb{E}_\theta((\bar{\theta}_n - \theta)^2) \\&= \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)^2) + 2(\bar{\theta}_n\bar{\theta}_n - \bar{\theta}_n\theta - \bar{\theta}_n\bar{\theta}_n + \bar{\theta}_n\theta) + \mathbb{E}_\theta((\bar{\theta}_n - \theta)^2) \\&= \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)^2) + 0 + \mathbb{E}_\theta((\bar{\theta}_n - \theta)^2) \\&= \mathbb{E}_\theta((\bar{\theta}_n - \theta)^2) + \mathbb{E}_\theta((\hat{\theta}_n - \bar{\theta}_n)^2) \\&= \mathbb{E}_\theta((\mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2) + \mathbb{E}_\theta((\hat{\theta}_n - \mathbb{E}_\theta(\hat{\theta}_n))^2) \\&= (\mathbb{E}_\theta(\hat{\theta}_n) - \theta)^2 + \mathbb{V}_\theta(\hat{\theta}_n) \\&= \mathbf{B}(\hat{\theta}_n)^2 + \mathbb{V}_\theta(\hat{\theta}_n).\end{aligned}$$

□

Finite-sample estimator properties

- Error, bias, and unbiasedness
- Variance and standard error
- Score function and Fisher information
- Cramér-Rao bound
- Mean squared error
- **Exercises**

Study Questions

1. Define the bias of an estimator. When is an estimator unbiased?
2. Write down the definition of the variance of an estimator.
3. Define the standard error of an estimator.
4. Define the standard error of the mean.
5. Write down the definitions of the score function of a random variable/sample.
6. Write down the definitions of the Fisher information of a random variable/sample.
7. Write down the definitions of the expected Fisher information of a random variable/sample.
8. What are the expected value and the variance of the score function of a random variable?
9. Formulate the Cramér-Rao bound theorem.
10. Write down the bias-variance decomposition for the mean squared error of an estimator.

Theoretical Exercises

1. Evaluate the Fisher information for the expectation of a Bernoulli random sample $X_1, \dots, X_n \sim B(\mu)$ (Held and Sabanés Bové, 2014, Examples 2.9).
2. Evaluate the Fisher information for the expectation (given a known variance parameter) and for the variance parameter (given a known expectation parameter) of a univariate Gaussian random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ (Held and Sabanés Bové, 2014, Examples 2.9).
3. Let X be a Poisson-distributed random variable with parameter λ . Show that the ML estimator of λ attains the Cramér-Rao bound (Held and Sabanés Bové, 2014, Example 4.11).
4. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$. Show that the estimator

$$\hat{\sigma}^2 := \frac{1}{n+1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \tag{419}$$

for the variance parameter σ^2 has a smaller mean squared error than both the maximum likelihood estimator of σ^2 and the sample variance S^2 (DeGroot and Schervish, 2012, Example 8.7.6).

Theoretical Exercise 2

Let $X_1 \sim \text{Poisson}(\lambda)$. Then

- the ML estimator of λ is given by $\hat{\lambda}_1^{\text{ML}} = X_1$,
- the ML estimator is an unbiased estimator, $\mathbb{E}(\hat{\lambda}_1^{\text{ML}}) = \lambda$.
- the variance of the ML estimator is $\mathbb{V}(\hat{\lambda}_1^{\text{ML}}) = \lambda$,
- the expected Fisher information of X_1 is $J(\lambda) = \lambda^{-1}$,
- and $\hat{\lambda}_{\text{ML}}$ thus attains the Cramér-Rao bound.

Theoretical Exercise 2 (Poisson random variable, expectation, variance)

Let X be a discrete random variable with outcome set $\mathcal{X} := \mathbb{N}^0$ and probability mass function

$$p : \mathbb{N}^0 \rightarrow [0, 1], x \mapsto p(x) := \frac{\lambda^x \exp(-\lambda)}{x!}. \quad (420)$$

Then X is said to be distributed according to a *Poisson distribution* with parameter λ for which we write $X \sim \text{Poisson}(\lambda)$. We abbreviate the PMF of a Poisson random variable by

$$\text{Poisson}(x; \lambda) := \frac{\lambda^x \exp(-\lambda)}{x!}. \quad (421)$$

The expectation and variance of Poisson random variable are given by

$$\mathbb{E}(X) = \mathbb{V}(X) = \lambda. \quad (422)$$

Exercises

Theoretical Exercise 2 (Poisson random variable, expectation, variance)

Proof of $\mathbb{E}(X) = \lambda$.

X is discrete with $\mathcal{X} = \mathbb{N}^0$. Thus

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x \in \mathbb{N}^0} x \cdot \text{Poisson}(x; \lambda) \\ &= \sum_{x \in \mathbb{N}^0} x \cdot \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= 0 \cdot \frac{\lambda^0 \exp(-\lambda)}{0!} + \sum_{x \in \mathbb{N}} x \cdot \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= \sum_{x \in \mathbb{N}} x \cdot \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= \sum_{x \in \mathbb{N}} \frac{\lambda^x \exp(-\lambda)}{(x - 1)!} \\ &= \lambda \exp(-\lambda) \sum_{x \in \mathbb{N}} \frac{\lambda^{x-1}}{(x - 1)!} \\ &= \lambda \exp(-\lambda) \sum_{j \in \mathbb{N}^0} \frac{\lambda^j}{j!} \\ &= \lambda \exp(-\lambda) \exp(\lambda) \\ &= \lambda.\end{aligned}\tag{423}$$

Exercises

Theoretical Exercise 2 (Poisson random variable, expectation, variance)

Proof of $\mathbb{V}(X) = \lambda$

We evaluate $\mathbb{E}(X^2)$ to use the variance translation to evaluate $\mathbb{V}(X)$. We have

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{x \in \mathbb{N}^0} x^2 \cdot \text{Poisson}(x; \lambda) \\ &= \sum_{x \in \mathbb{N}^0} x^2 \cdot \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= 0^2 \cdot \frac{\lambda^0 \exp(-\lambda)}{0!} + \sum_{x \in \mathbb{N}} x^2 \cdot \frac{\lambda^x \exp(-\lambda)}{x!} \\ &= \sum_{x \in \mathbb{N}} x^2 \cdot \frac{\lambda^x \exp(-\lambda)}{x!}\end{aligned}\tag{424}$$

Setting $\xi := x + 1$ then yields

Exercises

Theoretical Exercise 2 (Poisson random variable, expectation, variance)

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{\xi \in \mathbb{N}^0} (\xi + 1)^2 \frac{\lambda^{\xi+1} \exp(-\lambda)}{(\xi + 1)!} \\&= \sum_{\xi \in \mathbb{N}^0} (\xi + 1)^2 \frac{\lambda \lambda^\xi \exp(-\lambda)}{(\xi + 1)\xi!} \\&= \lambda \sum_{\xi \in \mathbb{N}^0} (\xi + 1) \frac{\lambda^\xi \exp(-\lambda)}{\xi!} \\&= \lambda \left(\sum_{\xi \in \mathbb{N}^0} \xi \frac{\lambda^\xi \exp(-\lambda)}{\xi!} + \sum_{\xi \in \mathbb{N}^0} \frac{\lambda^\xi \exp(-\lambda)}{\xi!} \right) \\&= \lambda \left(\mathbb{E}(X) + \sum_{\xi \in \mathbb{N}^0} \text{Poisson}(x; \lambda) \right) \\&= \lambda(\lambda + 1).\end{aligned}\tag{425}$$

The variance translation theorem thus yields

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.\tag{426}$$

□

Exercises

Theoretical Exercise 2 (Poisson ML estimator)

Let $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$ be n i.i.d. Poisson distributed random variables.

(1) Formulation of the log likelihood function

We have

$$L_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}_{>0}, \lambda \mapsto L_n(\lambda) := \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \quad (427)$$

Taking the logarithm yields

$$\begin{aligned} \ell_n : \mathbb{R}_{>0} \rightarrow \mathbb{R}, \lambda \mapsto \ell_n(\lambda) &:= \ln \left(\prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \right) \\ &= \sum_{i=1}^n \ln \left(\frac{\lambda^{x_i}}{x_i!} \exp(-\lambda) \right) \\ &= \sum_{i=1}^n (x_i \ln \lambda - \ln x_i! - \lambda) \\ &= \ln \lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \ln x_i! - n\lambda. \end{aligned} \quad (428)$$

Theoretical Exercise 2 (Poisson ML estimator)

(2) Evaluation of the log likelihood function derivative, setting to zero

We have

$$\begin{aligned}\frac{d}{d\lambda} \ell_n(\lambda) &= \frac{d}{d\lambda} \left(\ln \lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \ln x_i! - n\lambda \right) \\ &= \frac{1}{\lambda} \sum_{i=1}^n x_i - n\end{aligned}\tag{429}$$

and hence the *maximum likelihood equation* takes the form

$$\frac{1}{\hat{\lambda}} \sum_{i=1}^n x_i - n = 0.\tag{430}$$

Exercises

Theoretical Exercise 2 (Poisson ML estimator)

(3) Solving for critical points

We have

$$\begin{aligned} \frac{1}{\hat{\lambda}} \sum_{i=1}^n x_i - n &= 0 \\ \Leftrightarrow \frac{1}{\hat{\lambda}} &= \frac{n}{\sum_{i=1}^n x_i} \\ \Leftrightarrow \hat{\lambda} &= \frac{1}{n} \sum_{i=1}^n x_i \end{aligned} \tag{431}$$

Hence $\hat{\lambda} = n^{-1} \sum_{i=1}^n x_i$ is a candidate for an MLE of λ .

This can be consolidated, such that $\hat{\lambda}_n^{\text{ML}} := \frac{1}{n} \sum_{i=1}^n x_i$.

Exercises

Theoretical Exercise 2

Let $X_1 \sim \text{Poisson}(\lambda)$. Then

- the ML estimator of λ is given by $\hat{\lambda}_1^{\text{ML}} = X_1$,
- the ML estimator is an unbiased estimator, $\mathbb{E}(\hat{\lambda}_1^{\text{ML}}) = \lambda$,
- the variance of the ML estimator is $\mathbb{V}(\hat{\lambda}_1^{\text{ML}}) = \lambda$.

Proof

- The first statement follows directly from

$$\hat{\lambda}_n^{\text{ML}} := \frac{1}{n} \sum_{i=1}^n X_i \text{ for } n := 1. \quad (432)$$

- The second statement follows directly from

$$\mathbb{E}(\hat{\lambda}_1^{\text{ML}}) = \mathbb{E}(X_1) = \lambda. \quad (433)$$

- The third statement follows directly from

$$\mathbb{V}(\hat{\lambda}_1^{\text{ML}}) = \mathbb{V}(X_1) = \lambda. \quad (434)$$

□

Theoretical Exercise 2 (Poisson variable expected Fisher information)

Let $X_1 \sim \text{Poisson}(\lambda)$. Then

- The score function of the random variable X_1 evaluates to

$$S : \mathbb{R} \rightarrow \mathbb{R}, \lambda \mapsto S(\lambda) := \frac{x_1}{\lambda} - 1 \quad (435)$$

- The Fisher information of the random variable X_1 evaluates to

$$I : \mathbb{R} \rightarrow \mathbb{R}, \lambda \mapsto I(\lambda) := \frac{X_1}{\lambda^2} \quad (436)$$

- The expected Fisher information of the random variable X_1 evaluates to

$$J : \mathbb{R} \rightarrow \mathbb{R}, \lambda \mapsto J(\lambda) := \frac{1}{\lambda} \quad (437)$$

Because

$$\mathbb{V} \left(\hat{\lambda}_1^{ML} \right) = \lambda = \frac{1}{\frac{1}{\lambda}} = \frac{1}{J(\lambda)}, \quad (438)$$

$\hat{\lambda}_1^{ML}$ thus attains the Cramér-Rao bound.

Exercises

Theoretical Exercise 2 (Poisson variable expected Fisher information)

Proof

- The value of the score function is given by

$$S(\lambda) = \frac{d}{d\lambda} \ell_1(\lambda) = \frac{1}{\lambda} \sum_{i=1}^1 x_i - 1 = \frac{x_1}{\lambda} - 1 \quad (439)$$

- The value of the Fisher information is given by

$$I(\lambda) = -\frac{d^2}{d\lambda^2} \ell_1(\lambda) = -\frac{d}{d\lambda} S(\lambda) = -\frac{d}{d\lambda} \left(\frac{x_1}{\lambda} - 1 \right) = \frac{x_1}{\lambda^2} \quad (440)$$

- The value of the expected Fisher information is given by

$$J(\lambda) = \mathbb{E}_\lambda(I(\lambda)) = \mathbb{E}_\lambda \left(\frac{X_1}{\lambda^2} \right) = \frac{\mathbb{E}_\lambda(X_1)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda} \quad (441)$$

□

Exercises

Theoretical Exercise 3 (MSEs of Gaussian variance estimators)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ and let

$$\hat{\sigma}_c^2 := c \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (442)$$

such that

- for $c := \frac{1}{n}$, $\hat{\sigma}_c^2$ corresponds to the ML estimator of σ^2
- for $c := \frac{1}{n-1}$, $\hat{\sigma}_c^2$ corresponds to the sample variance estimator of σ^2 , and
- for $c := \frac{1}{n+1}$, $\hat{\sigma}_c^2$ corresponds to the current estimator of interest of σ^2 .

Without proof, we first note that (cf. Lecture 11)

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \chi^2(n-1) \quad (443)$$

and that a χ^2 random variable has expectation $n-1$ and variance $2(n-1)$ (cf. Lecture 6), such that

$$\mathbb{E}(\hat{\sigma}_c^2) = (n-1)c\sigma^2 \text{ and } \mathbb{V}(\hat{\sigma}_c^2) = 2(n-1)c^2\sigma^4. \quad (444)$$

We next show the following result:

$$\text{MSE}(\hat{\sigma}_c^2) = ((n^2 - 1)c^2 - 2(n-1)c)\sigma^4. \quad (445)$$

Exercises

Theoretical Exercise 3 (MSEs of Gaussian variance estimators)

With the bias-variance decomposition of the MSE, we have

$$\begin{aligned}\text{MSE}(\hat{\sigma}_c^2) &= \mathbb{B}(\hat{\sigma}_c^2)^2 + \mathbb{V}(\hat{\sigma}_c^2) \\ &= (\mathbb{E}(\hat{\sigma}_c^2) - \sigma^2)^2 + \mathbb{V}(\hat{\sigma}_c^2) \\ &= ((n-1)c\sigma^2 - \sigma^2)^2 + 2(n-1)c^2\sigma^4 \\ &= (\sigma^2((n-1)c-1))^2 + 2(n-1)c^2\sigma^4 \\ &= ((n-1)c-1)^2\sigma^4 + 2(n-1)c^2\sigma^4 \\ &= (((n-1)c-1)^2 + 2(n-1)c^2)\sigma^4 \\ &= ((n-1)^2c^2 - 2(n-1)c + 1 + 2nc^2 - 2c^2)\sigma^4 \\ &= (n^2c^2 - 2nc^2 + c^2 + 2nc^2 - 2c^2 - 2(n-1)c + 1)\sigma^4 \\ &= (n^2c^2 - c^2 - 2(n-1)c + 1)\sigma^4 \\ &= ((n^2 - 1)c^2 - 2(n-1)c + 1)\sigma^4.\end{aligned}\tag{446}$$

Exercises

Theoretical Exercise 3 (MSEs of Gaussian variance estimators)

Finally, we analytically minimize the coefficient of σ^4 with respect to c . We have

$$\frac{d}{dc} \left((n^2 - 1)c^2 - 2(n-1)c + 1 \right) = 2(n^2 - 1)c - 2(n-1), \quad (447)$$

such that for a minimum value candidate \tilde{c} it holds that

$$\begin{aligned} & 2(n^2 - 1)\tilde{c} - 2(n-1) = 0 \\ \Leftrightarrow & (n+1)(n-1)\tilde{c} - (n-1) = 0 \\ \Leftrightarrow & \tilde{c} = \frac{(n-1)}{(n+1)(n-1)} \\ \Leftrightarrow & \tilde{c} = \frac{1}{n+1}. \end{aligned} \quad (448)$$

With

$$\begin{aligned} \frac{d^2}{dc^2} \left((n^2 - 1)c^2 - 2(n-1)c + 1 \right) &= \frac{d}{dc} 2(n^2 - 1)c - 2(n-1) \\ &= 2(n^2 - 1) \\ &> 0 \text{ for } n \geq 2, \end{aligned} \quad (449)$$

it thus follows that $c := \frac{1}{n+1}$ minimizes the MSE of $\hat{\sigma}_c^2$.

□

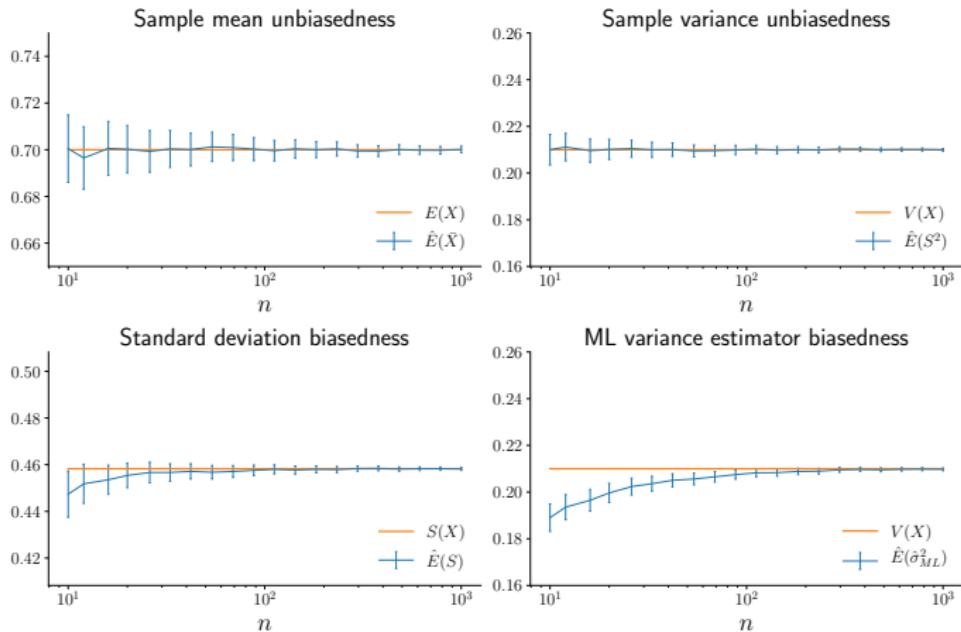
Programming Exercises

1. For $X_1, \dots, X_n \sim \text{Bern}(\mu)$ implement a simulation which validates the unbiasedness of the sample mean, the unbiasedness of the sample variance, the biasedness of the sample standard deviation, and the biasedness of the maximum likelihood variance parameter estimator.
2. For $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ implement a simulation which validates the unbiasedness of the sample mean, the unbiasedness of the sample variance, the biasedness of the sample standard deviation, and the biasedness of the maximum likelihood variance parameter estimator.
3. Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ implement a simulation that validates the bias-variance decompositions of the mean squared errors of the maximum likelihood estimator of σ^2 , the sample variance S^2 , and the estimator $\hat{\sigma}^2$ introduced in the third theoretical exercise above (DeGroot and Schervish, 2012, Example 8.7.6).

Exercises

Programming Exercise 1

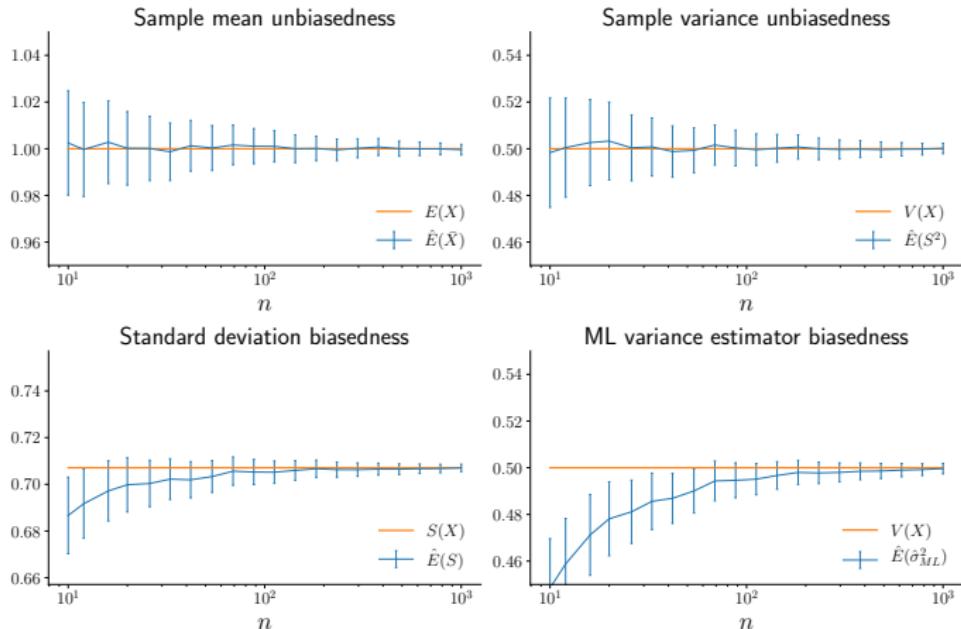
$$X_1, \dots, X_n \sim \text{Bern}(\mu), \mu := 0.7$$



Exercises

Programming Exercise 2

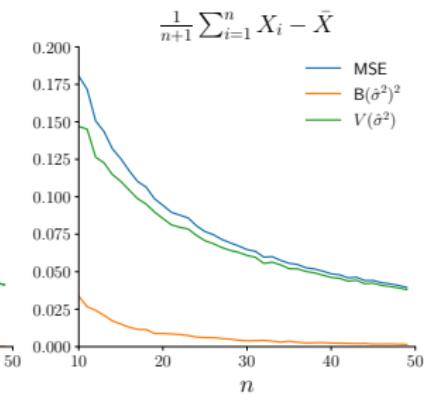
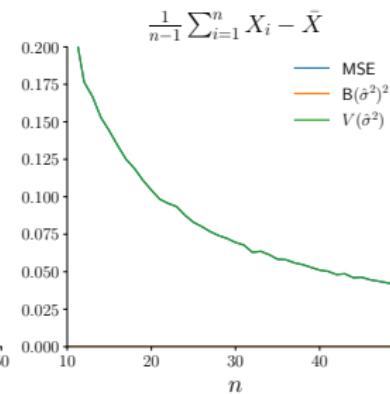
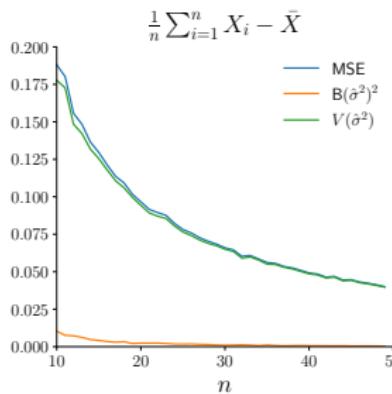
$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \mu := 1, \sigma^2 := 0.5$$



Exercises

Programming Exercise 3

$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \mu := 1, \sigma^2 := 1$$



(10) Asymptotic estimator properties

Bibliographic remarks

The material presented in this section is based on Wasserman (2004), Held and Sabanés Bové (2014), and Casella and Berger (2012, Section 10.1). The formulation and proof of the consistency of the maximum likelihood estimator is based on Held and Sabanés Bové (2014, Section 4.2.2) and the formulation and proof of the asymptotic normality of the maximum likelihood estimator is based on Held and Sabanés Bové (2014, Section 4.2.3).

Overview

- This lecture is a brief introduction to “Asymptotic Statistics”.
- Asymptotic Statistics concerns estimator behavior for large sample sizes.
- Asymptotic Statistics is used to
 - Study the qualitative properties of estimators and
 - Derive estimator property approximations for large sample sizes.
- In “Big Data” sample sizes are very large (or are they?).
- Asymptotic Statistics is thus practically useful and justified.
- Van der Waart (2000) provides a comprehensive introduction.

Overview

Let $\hat{\theta}_n$ be an estimator for the parameter θ of a statistical model based on a random sample $X_1, \dots, X_n \sim p_\theta$ of size n . Then $\hat{\theta}_n$ is called

- *asymptotically unbiased*, if for large sample sizes $n \rightarrow \infty$ the expected value of $\hat{\theta}_n$ corresponds to the true, but unknown, parameter value,
- *consistent*, if for large sample sizes $n \rightarrow \infty$ the probability that $\hat{\theta}_n$ deviates from the true, but unknown, value becomes small,
- *asymptotically normally distributed*, if for large sample sizes $n \rightarrow \infty$, the distribution of $\hat{\theta}_n$ is given by a normal distribution,
- *asymptotically efficient*, if for large sample sizes $n \rightarrow \infty$, the distribution of $\hat{\theta}_n$ is given by a normal distribution with expectation corresponding to the true, but unknown, parameter value and variance corresponding to the reciprocal of the expected Fisher information, i.e., the Cramér-Rao lower bound.

Maximum likelihood estimators are asymptotically unbiased, consistent, asymptotically normally distributed, and asymptotically efficient.

Asymptotic estimator properties

- Asymptotic unbiasedness
- Consistency
- Asymptotic normality
- Asymptotic efficiency
- Maximum likelihood estimator properties
- Exercises

Asymptotic estimator properties

- **Asymptotic unbiasedness**
- Consistency
- Asymptotic normality
- Asymptotic efficiency
- Maximum likelihood estimator properties
- Exercises

Definition (Asymptotic unbiasedness)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ for $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n$ be an estimator for θ . Then $\hat{\theta}_n$ is called *asymptotically unbiased*, if

$$\lim_{n \rightarrow \infty} \mathbb{E}_\theta(\hat{\theta}_n) = \theta \text{ for all } \theta \in \Theta. \quad (450)$$

Remarks

- Asymptotically unbiased estimators are unbiased in the large sample limit.
- An unbiased estimator is necessarily asymptotically unbiased.

Example (An asymptotically unbiased estimator)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ be n i.i.d. Gaussian random variables and let

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (451)$$

denote the maximum likelihood estimator of the variance parameter σ^2 . From the derivation of the unbiasedness of the sample variance (cf. Lecture (9)), we have

$$\mathbb{E}(\hat{\sigma}_n^2) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2\right) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2\right) = \frac{n-1}{n} \sigma^2 \neq \sigma^2. \quad (452)$$

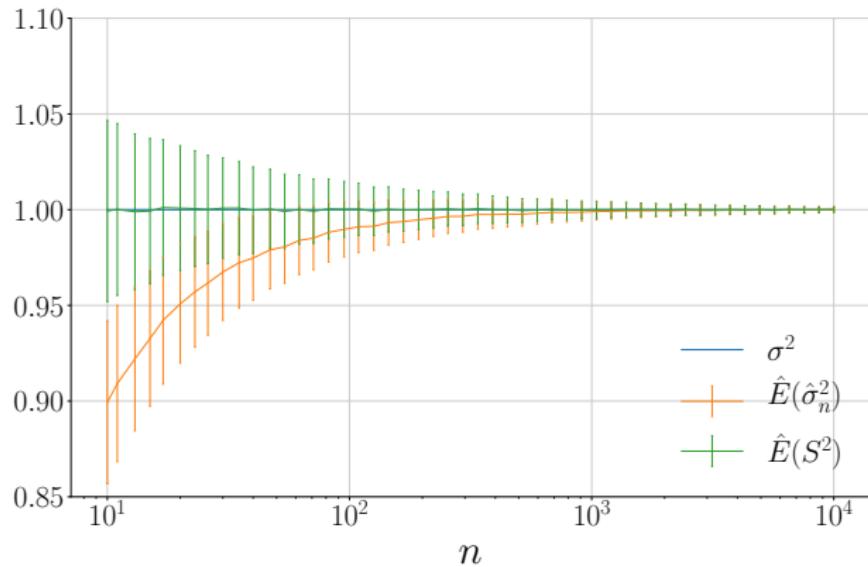
$\hat{\sigma}_n^2$ is thus a biased estimator for σ^2 . However, because $(n-1)/n \rightarrow 1$ for $n \rightarrow \infty$, we have

$$\lim_{n \rightarrow \infty} \mathbb{E}(\hat{\sigma}_n^2) = \lim_{n \rightarrow \infty} \frac{n-1}{n} \sigma^2 = \sigma^2 \lim_{n \rightarrow \infty} \frac{n-1}{n} = \sigma^2. \quad (453)$$

In the limit of large samples, the maximum likelihood estimator for the variance parameter of a Gaussian distribution is thus unbiased.

Asymptotic unbiasedness

Asymptotic unbiasedness of $\hat{\sigma}_{\text{ML}}^2$ and unbiasedness of S^2 for $X_1, \dots, X_n \sim N(0, 1)$.



Asymptotic estimator properties

- Asymptotic unbiasedness
- **Consistency**
- Asymptotic normality
- Asymptotic efficiency
- Maximum likelihood estimator properties
- Exercises

Definition (Consistency)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ for $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n$ be an estimator for θ . Then a sequence of estimators $\hat{\theta}_1, \hat{\theta}_2, \dots$ is called a *consistent sequence of estimators*, if for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| < \epsilon \right) = 1 \Leftrightarrow \lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| \geq \epsilon \right) = 0. \quad (454)$$

If $\hat{\theta}_1, \hat{\theta}_2, \dots$ is a consistent sequence of estimators, then $\hat{\theta}_n$ is referred to as a *consistent estimator*.

Remarks

- As $n \rightarrow \infty$, the probability that $\hat{\theta}_n$ is arbitrarily close to θ becomes high.
- As $n \rightarrow \infty$, the probability that $\hat{\theta}_n$ deviates from θ becomes small.
- These properties hold for all possible true, but unknown, parameter values.
- The convergence type used here is *convergence in probability* (cf. Lecture (7)).
- Consistency can be shown directly or based on mean square error criteria.

Consistency

Theorem (Mean squared error criterion for consistency)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ for $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n$ be an estimator for θ . If

$$\lim_{n \rightarrow \infty} \text{MSE}(\hat{\theta}_n) = \lim_{n \rightarrow \infty} \mathbb{E}_\theta \left((\hat{\theta}_n - \theta)^2 \right) = 0, \quad (455)$$

then $\hat{\theta}_n$ is a consistent estimator.

Proof

From Chebychev's inequality (cf. Lecture (7)), it follows directly that

$$\mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| \geq \epsilon \right) \leq \frac{\mathbb{E}_\theta \left((\hat{\theta}_n - \theta)^2 \right)}{\epsilon^2} \quad (456)$$

Taking limits yields

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \lim_{n \rightarrow \infty} \mathbb{E}_\theta \left((\hat{\theta}_n - \theta)^2 \right). \quad (457)$$

Hence, if $\lim_{n \rightarrow \infty} \mathbb{E}_\theta \left((\hat{\theta}_n - \theta)^2 \right) = 0$, then with $0 \leq \mathbb{P}_\theta(|\hat{\theta}_n - \theta| \geq \epsilon) \leq 1$ it follows that

$$\lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(|\hat{\theta}_n - \theta| \geq \epsilon \right) = 0, \quad (458)$$

i.e., $\hat{\theta}_n$ is a consistent estimator.

□

Theorem (Bias and variance criterion for consistency)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ for $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n$ be an estimator for θ . If

$$\lim_{n \rightarrow \infty} B(\hat{\theta}_n) = 0 \text{ and } \lim_{n \rightarrow \infty} V_\theta(\hat{\theta}_n) = 0, \quad (459)$$

then $\hat{\theta}_n$ is a consistent estimator.

Proof

- o If for $n \rightarrow \infty$ it holds that $B(\hat{\theta}_n) \rightarrow 0$, then also $B(\hat{\theta}_n)^2 \rightarrow 0$.
- o If for $n \rightarrow \infty$ both $B(\hat{\theta}_n)^2 \rightarrow 0$ and $V_\theta(\hat{\theta}_n) \rightarrow 0$, then also $\lim_{n \rightarrow \infty} \text{MSE}(\hat{\theta}_n) = 0$.
- o Thus, with the MSE criterion $\lim_{n \rightarrow \infty} \mathbb{P}_\theta(|\hat{\theta}_n - \theta| \geq \epsilon) = 0$ and $\hat{\theta}_n$ is consistent.

□

Example (Consistency of the sample mean)

With the bias and variance criterion for consistency, the consistency of the sample mean as an estimator of μ for $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ follows directly from

$$B(\bar{X}_n) = 0 \text{ and } \lim_{n \rightarrow \infty} V_\theta(\bar{X}_n) = \lim_{n \rightarrow \infty} \frac{1}{n} \sigma^2 = 0. \quad (460)$$

Asymptotic estimator properties

- Asymptotic unbiasedness
- Consistency
- **Asymptotic normality**
- Asymptotic efficiency
- Maximum likelihood estimator properties
- Exercises

Definition (Asymptotic normality)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$, and let $\hat{\theta}_n$ be an estimator for θ . Let further $\tilde{\theta} \sim N(\mu, \sigma^2)$ be a normally distributed random variable with expectation parameter μ and variance parameter σ^2 . Then, if $\hat{\theta}_n$ converges to $\tilde{\theta}$ in distribution, $\hat{\theta}_n$ is said to be *asymptotically normally distributed* and we write

$$\hat{\theta}_n \xrightarrow{a} N(\mu, \sigma^2). \quad (461)$$

Remark

- Convergence in distribution means that $\lim_{n \rightarrow \infty} P_n(\hat{\theta}_n) = P(\tilde{\theta})$. (cf. Lecture (7)).

Asymptotic estimator properties

- Asymptotic unbiasedness
- Consistency
- Asymptotic normality
- **Asymptotic efficiency**
- Maximum likelihood estimator properties
- Exercises

Definition (Asymptotic efficiency)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ , $X_1, \dots, X_n \sim p_\theta$ be random sample, $\hat{\theta}_n$ be an estimator for θ , and let $J_n(\theta)$ denote the expected Fisher information of the random sample X_1, \dots, X_n . If

$$\hat{\theta}_n \xrightarrow{a} N\left(\theta, J_n(\theta)^{-1}\right), \quad (462)$$

then $\hat{\theta}_n$ is said to be *asymptotically efficient*.

Remarks

- Asymptotic efficiency implies asymptotic normality.
- Asymptotic efficiency implies asymptotic unbiasedness.
- The variance of the asymptotic distribution is called the *asymptotic variance*.
- The variance of an asymptotically efficient estimator attains the Cramér-Rao bound.
- The term *efficiency* is used with variations in the literature.

Asymptotic estimator properties

- Asymptotic unbiasedness
- Consistency
- Asymptotic normality
- Asymptotic efficiency
- **Maximum likelihood estimator properties**
- Exercises

Maximum likelihood estimators are (under the Fisher regularity conditions)

- (1) not necessarily unbiased,
- (2) consistent,
- (3) asymptotically normally distributed,
- (4) asymptotically unbiased, and
- (5) asymptotically efficient.

Example (Non-necessary unbiasedness of MLEs)

Let $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ and for $\mu_{\text{ML}} := n^{-1} \sum_{i=1}^n X_i$ consider the maximum likelihood estimator

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\text{ML}})^2 \quad (463)$$

of the variance parameter σ^2 . Recall that

$$\mathbb{E} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right) = (n-1)\sigma^2. \quad (464)$$

Hence

$$\mathbb{E}(\hat{\sigma}_{\text{ML}}^2) = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\text{ML}})^2 \right) = \frac{n-1}{n} \sigma^2 \neq \sigma^2, \quad (465)$$

and $\hat{\sigma}_{\text{ML}}^2$ is not unbiased.

Theorem (Consistency of maximum likelihood estimators)

Let \mathcal{P} denote a parametric statistical model with PDF p_θ for $\theta \in \Theta$, let $X_1, \dots, X_n \sim p_\theta$, and assume that the Fisher regularity conditions hold. Let $\hat{\theta}_n^{\text{ML}}$ denote the maximum likelihood estimator for the true, but unknown, value θ and let $\hat{\theta}_n^{\text{ML}}$ be defined here as a local maximizer of the likelihood function L_n based on a sample of size n . Then there exists a consistent sequence of maximum likelihood estimators $\hat{\theta}_1^{\text{ML}}, \hat{\theta}_2^{\text{ML}}, \dots$ and $\hat{\theta}_n^{\text{ML}}$ is said to be a consistent estimator of θ .

Remarks

- $\hat{\theta}_n^{\text{ML}}$ converges in probability to the true, but unknown value, θ .
- Formally, $\lim_{n \rightarrow \infty} \mathbb{P}_\theta \left(|\hat{\theta}_n^{\text{ML}} - \theta| < \epsilon \right) = 1$ for every $\epsilon > 0$ and $\theta \in \Theta$.
- There exists proofs of varying mathematical generality and depth.
- Put bluntly, the theorem states that for large sample sizes maximum likelihood estimates equal the true, but unknown, parameter value with certainty.

Maximum likelihood estimator properties

Proof (Consistency of maximum likelihood estimators)

The idea of the proof provided by Held and Sabanés Bové (2014, Section 4.2.2) is to show that there exists a (possibly local) maximum of the likelihood function L_n in an interval $\theta - \epsilon, \theta + \epsilon$ around the true, but unknown, parameter value θ for arbitrary small ϵ with probability 1 as $n \rightarrow \infty$.

Because the likelihood function is assumed to be continuous and $\hat{\theta}_n^{\text{ML}}$ is understood as a local maximizer of the likelihood function, this entails that $|\hat{\theta}_n^{\text{ML}} - \theta| < \epsilon$ with probability 1 as $n \rightarrow \infty$, which in turn corresponds to the convergence in probability of $\hat{\theta}_n^{\text{ML}}$ to the true, but unknown, parameter value θ .

To show that there exists a maximum of the likelihood function L_n in $\theta - \epsilon, \theta + \epsilon$ with probability 1 as $n \rightarrow \infty$ it suffices (with the continuity of the likelihood function) to show that

$$\lim_{n \rightarrow \infty} \mathbb{P}(L_n(\theta) > L_n(\theta - \epsilon)) = 1 \text{ and } \lim_{n \rightarrow \infty} \mathbb{P}(L_n(\theta) > L_n(\theta + \epsilon)) = 1 \quad (466)$$

for arbitrary small $\epsilon > 0$.

Because analogous arguments can be made to show either of these statements, we consider only the first. The aim is thus to show that

$$\lim_{n \rightarrow \infty} \mathbb{P}(L_n(\theta) > L_n(\theta - \epsilon)) = 1, \quad (467)$$

which can be argued with the weak law of large numbers and the positivity of the Kullback-Leibler divergence between two non-identical PDFs with equal support.

Maximum likelihood estimator properties

Proof (Consistency of maximum likelihood estimators)

To set up this line of reasoning, we first note that the event of interest

$$L_n(\theta) > L_n(\theta - \epsilon) \quad (468)$$

can be reformulated as

$$\begin{aligned} L_n(\theta) > L_n(\theta - \epsilon) &\Leftrightarrow \frac{1}{n} \ln L_n(\theta) > \frac{1}{n} \ln L_n(\theta - \epsilon) \\ &\Leftrightarrow \frac{1}{n} \ln L_n(\theta) - \frac{1}{n} \ln L_n(\theta - \epsilon) > 0 \\ &\Leftrightarrow \frac{1}{n} \sum_{i=1}^n \ln p_\theta(x_i) - \frac{1}{n} \sum_{i=1}^n \ln p_{\theta-\epsilon}(x_i) > 0 \\ &\Leftrightarrow \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{p_\theta(x_i)}{p_{\theta-\epsilon}(x_i)} \right) > 0. \end{aligned} \quad (469)$$

Recall that the weak law of large numbers states that the sample mean of n independent and identically distributed random variables $X_i, i = 1, \dots, n$ converges in probability to the expected value of X_i . Applying this theorem to the random variables $\ln(p_\theta(x_i)/p_{\theta-\epsilon}(x_i))$ thus entails that

$$\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{p_\theta(x_i)}{p_{\theta-\epsilon}(x_i)} \right) \xrightarrow[n \rightarrow \infty]{P} \mathbb{E}_\theta \left(\ln \left(\frac{p_\theta(x)}{p_{\theta-\epsilon}(x)} \right) \right). \quad (470)$$

Maximum likelihood estimator properties

Proof (Consistency of maximum likelihood estimators)

Finally, the right-hand side of the above corresponds to the Kullback-Leibler divergence between $p_\theta(x)$ and $p_{\theta-\epsilon}(x)$, which, with the Fisher regularity conditions, are distinct PDFs with equal support.

With the non-negativity of the Kullback-Leibler divergence, it follows that

$$\mathbb{E}_\theta \left(\ln \left(\frac{p_\theta(x)}{p_{\theta-\epsilon}(x)} \right) \right) > 0. \quad (471)$$

⇒ We have thus shown that $\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{p_\theta(x_i)}{p_{\theta-\epsilon}(x_i)} \right)$ converges in probability to a value larger than 0, which corresponds to a limiting probability of 1 for $\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{p_\theta(x_i)}{p_{\theta-\epsilon}(x_i)} \right)$ to be larger than zero.

⇒ In summary, we thus have

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{p_\theta(x_i)}{p_{\theta-\epsilon}(x_i)} \right) > 0 \right) = 1 \Leftrightarrow \lim_{n \rightarrow \infty} \mathbb{P} (L_n(\theta) > L_n(\theta - \epsilon)) = 1. \quad (472)$$

as required for the proof's main argument.

□

Theorem (Asymptotic efficiency of maximum likelihood estimators)

Let \mathcal{P} be a parametric statistical model with PDF p_θ , let $X_1, \dots, X_n \sim p_\theta$ be a random sample, and let $\hat{\theta}_n^{\text{ML}}$ denote the maximum likelihood estimator for the true, but unknown, parameter θ . Let $J_n(\theta)$ denote the expected Fisher information of the random sample and assume that the Fisher regularity conditions hold. Then $\hat{\theta}_n^{\text{ML}}$ is asymptotically efficient, i.e.,

$$\hat{\theta}_n^{\text{ML}} \xrightarrow{a} N(\theta, J_n(\theta)^{-1}). \quad (473)$$

Remarks

- $\hat{\theta}_n^{\text{ML}}$ is asymptotically efficient
 - $\Rightarrow \hat{\theta}_n^{\text{ML}}$ is asymptotically unbiased
 - $\Rightarrow \hat{\theta}_n^{\text{ML}}$ is asymptotically normally distributed
- $\hat{\theta}_n^{\text{ML}} \xrightarrow{a} N(\theta, J_n(\theta)^{-1})$ is used to construct approximate confidence intervals and hypotheses tests in applied statistics as discussed in Lectures (11) and (12).

Maximum likelihood estimator properties

Proof (Asymptotic efficiency of maximum likelihood estimators)

The idea of the proof provided by Held and Sabanés Bové (2014, Section 4.2.3) is to show that

$$\sqrt{J_n(\theta)} (\hat{\theta}_n^{\text{ML}} - \theta) \xrightarrow{a} N(0, 1) \quad (474)$$

and then concluding that $\hat{\theta}_n^{\text{ML}} \xrightarrow{a} N(\theta, J_n(\theta)^{-1})$ as implied by the theorem on the transformation of normal random variables under linear-affine functions (cf. Lecture (5)).

The proof rests on

(1) the *continuous mapping theorem*,

(2) *Slutsky's theorem*

(3) $\frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow[n \rightarrow \infty]{D} N(0, J(\theta))$,

(4) $\frac{1}{n} I_n(\theta) \xrightarrow[n \rightarrow \infty]{P} J(\theta)$.

We first introduce these building blocks individually and then consider their joint application to show the central result.

Maximum likelihood estimator properties

Proof (Asymptotic efficiency of maximum likelihood estimators)

(1) *Continuous mapping theorem*

The continuous mapping theorem (Mann and Wald, 1943) states that convergence in probability and convergence in distribution are preserved under the application of a continuous univariate real-valued function f , i.e.,

- o if $X_n \xrightarrow[n \rightarrow \infty]{P} X$, then also $f(X_n) \xrightarrow[n \rightarrow \infty]{P} f(X)$, and
- o if $X_n \xrightarrow[n \rightarrow \infty]{D} X$, then also $f(X_n) \xrightarrow[n \rightarrow \infty]{D} f(X)$.

(2) *Slutsky's theorem*

Slutsky's theorem (Slutsky, 1925) states that for

$$X_n \xrightarrow[n \rightarrow \infty]{D} X \text{ and } Y_n \xrightarrow[n \rightarrow \infty]{P} a \in \mathbb{R} \quad (475)$$

it holds that

$$X_n + Y_n \xrightarrow[n \rightarrow \infty]{D} X + a \quad (476)$$

and

$$X_n Y_n \xrightarrow[n \rightarrow \infty]{D} aX \quad (477)$$

Maximum likelihood estimator properties

Proof (Asymptotic efficiency of maximum likelihood estimators)

$$(3) \text{ Proof of } \frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow[n \rightarrow \infty]{D} N(0, J(\theta))$$

With the properties of the score function (cf. Lecture (9)), we first note that

$$\frac{1}{\sqrt{n}} S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n S(\theta), \quad \mathbb{E}(S_n(\theta)) = 0, \quad \mathbb{V}(S_n(\theta)) = J(\theta). \quad (478)$$

We next note that for n i.i.d. random variables Y_1, \dots, Y_n with expectation $\mu := \mathbb{E}(Y_i)$ and variance $\sigma^2 := \mathbb{V}(Y_i)$, the central limit theorem in the Lindeberg and Lévy form states that (cf. Lecture (7))

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \xrightarrow[n \rightarrow \infty]{D} N(\mu, \sigma^2). \quad (479)$$

Substitution thus yields

$$\frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow[n \rightarrow \infty]{D} N(0, J(\theta)). \quad (480)$$

Maximum likelihood estimator properties

Proof (Asymptotic efficiency of maximum likelihood estimators)

$$(4) \text{ Proof of } \frac{1}{n} I_n(\theta) \xrightarrow[n \rightarrow \infty]{P} J(\theta)$$

With the properties of the Fisher information (cf. Lecture (7)), we first note that

$$\frac{1}{n} I_n(\theta) = \frac{1}{n} \sum_{i=1}^n I(\theta), \quad \mathbb{E}(I(\theta)) = J(\theta). \quad (481)$$

We next note that for n i.i.d. random variables Y_1, \dots, Y_n with expectation $\mu := \mathbb{E}(Y_i)$ the weak law of large numbers states that (cf. Lecture (7))

$$\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow[n \rightarrow \infty]{P} \mu. \quad (482)$$

Substitution thus yields

$$\frac{1}{n} I_n(\theta) \xrightarrow[n \rightarrow \infty]{P} J(\theta). \quad (483)$$

Maximum likelihood estimator properties

Proof (Asymptotic efficiency of maximum likelihood estimators)

With building blocks (1) - (4) in place, we are now in the position to show

$$\sqrt{J_n(\theta)} (\hat{\theta}_n^{\text{ML}} - \theta) \xrightarrow{a} N(0, 1). \quad (484)$$

To this end, we first consider a first-order Taylor approximation of the score function with expansion point corresponding to the true, but unknown, parameter value,

$$S_n(\tilde{\theta}) \approx S_n(\theta) - I_n(\theta)(\tilde{\theta} - \theta). \quad (485)$$

For $\tilde{\theta} = \hat{\theta}_n^{\text{ML}}$, we then have

$$\begin{aligned} S_n(\hat{\theta}_n^{\text{ML}}) &\approx S_n(\theta) - I_n(\theta)(\hat{\theta}_n^{\text{ML}} - \theta) \\ \Leftrightarrow 0 &\approx S_n(\theta) - I_n(\theta)(\hat{\theta}_n^{\text{ML}} - \theta) \\ \Leftrightarrow \sqrt{n}S_n(\theta) &\approx \sqrt{n}I_n(\theta)(\hat{\theta}_n^{\text{ML}} - \theta) \\ \Leftrightarrow \sqrt{n}(\hat{\theta}_n^{\text{ML}} - \theta) &\approx \sqrt{n} \frac{S_n(\theta)}{I_n(\theta)} \end{aligned} \quad (486)$$

and with $n^{1/2} = n^1 n^{-1/2}$ obtain

$$\sqrt{n}(\hat{\theta}_n^{\text{ML}} - \theta) \approx \left(\frac{1}{n} I_n(\theta) \right)^{-1} \frac{1}{\sqrt{n}} S_n(\theta). \quad (487)$$

Maximum likelihood estimator properties

Proof (Asymptotic efficiency of maximum likelihood estimators)

From Slutsky's theorem we have for

$$\frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow[n \rightarrow \infty]{D} N(0, J(\theta)) \text{ and } \frac{1}{n} I_n(\theta) \xrightarrow[n \rightarrow \infty]{P} J(\theta) \quad (488)$$

that

$$\left(\frac{1}{n} I_n(\theta) \right)^{-1} \frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow[n \rightarrow \infty]{D} J(\theta)^{-1} N(0, J(\theta)). \quad (489)$$

Thus

$$\sqrt{n} (\hat{\theta}_n^{\text{ML}} - \theta) \approx \left(\frac{1}{n} I_n(\theta) \right)^{-1} \frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow[n \rightarrow \infty]{D} J(\theta)^{-1} N(0, J(\theta)). \quad (490)$$

So far, so good.

□

Homework

- Show that if $X_n \xrightarrow[n \rightarrow \infty]{P} X$, then also $X_n^{-1} \xrightarrow[n \rightarrow \infty]{P} X^{-1}$.
- Show that

$$X \approx Y \text{ and } Y \xrightarrow[n \rightarrow \infty]{D} N(\mu, \sigma^2) \Rightarrow X \xrightarrow[n \rightarrow \infty]{D} N(\mu, \sigma^2) \quad (491)$$

or that the first-order Taylor approximation of the Score function at the location of the maximum likelihood estimator is exact.

Asymptotic estimator properties

- Asymptotic unbiasedness
- Consistency
- Asymptotic normality
- Asymptotic efficiency
- Maximum likelihood estimator properties
- Exercises

Summary

Let $\hat{\theta}_n$ be an estimator for the parameter θ of a statistical model based on a random sample $X_1, \dots, X_n \sim p_\theta$ of size n . Then $\hat{\theta}_n$ is called

- *asymptotically unbiased*, if for large sample sizes $n \rightarrow \infty$ the expected value of $\hat{\theta}_n$ corresponds to the true, but unknown, parameter value,
- *consistent*, if for large sample sizes $n \rightarrow \infty$ the probability that $\hat{\theta}_n$ deviates from the true, but unknown, value becomes small,
- *asymptotically normally distributed*, if for large sample sizes $n \rightarrow \infty$, the distribution of $\hat{\theta}_n$ is given by a normal distribution,
- *asymptotically efficient*, if for large sample sizes $n \rightarrow \infty$, the distribution of $\hat{\theta}_n$ is given by a normal distribution with expectation corresponding to the true, but unknown, parameter value and variance corresponding to the reciprocal of the expected Fisher information, i.e., the Cramér-Rao lower bound.

Maximum likelihood estimators are asymptotically unbiased, consistent, asymptotically normally distributed, and asymptotically efficient.

Asymptotic estimator properties

- Asymptotic unbiasedness
- Consistency
- Asymptotic normality
- Asymptotic efficiency
- Maximum likelihood estimator properties
- **Exercises**

Study questions

1. Write down the definition of an asymptotically unbiased estimator.
2. Write down the definition of a consistent estimator.
3. State the mean squared error criterion for estimator consistency.
4. State the bias and variance criterion for estimator consistency.
5. Write down the definition of an asymptotically normally distributed estimator.
6. Write down the definition of an asymptotically efficient estimator.
7. Name five properties of maximum likelihood estimators.
8. Given an example of a biased maximum likelihood estimator.
9. Sketch the proof of the consistency of maximum likelihood estimators
10. Sketch the proof of the asymptotic efficiency of maximum likelihood estimators.

Theoretical exercises

1. Show the consistency of the sample mean without recourse to the mean squared error consistency criterion (Casella and Berger, 2012, Example 10.1.2).
2. Show that the sample variance is a consistent estimator of the variance, if its variance converges to zero for $n \rightarrow \infty$ (Casella and Berger, 2012, Example 5.5.3).
3. Show that sample standard deviation is a consistent estimator of the standard deviation, if the sample variance is consistent (Casella and Berger, 2012, Example 5.5.5).

Exercises

Theoretical Exercise 1

We show that for $X_1, X_2, \dots \sim N(\mu, 1)$, the sample mean is a consistent estimator of μ .

Proof

We first note that with $\bar{X}_n \sim N(\mu, \frac{1}{n})$, we have

$$\begin{aligned}\mathbb{P}(|\bar{X}_n - \mu| < \epsilon) &= \int_{\mu-\epsilon}^{\mu+\epsilon} N\left(\bar{x}_n; \mu, \frac{1}{n}\right) d\bar{x}_n \\ &= \int_{\mu-\epsilon}^{\mu+\epsilon} \left(\frac{n}{2\pi}\right)^{1/2} \exp\left(-\frac{n}{2}(\bar{x}_n - \mu)^2\right) d\bar{x}_n\end{aligned}\tag{492}$$

We next recall that the integration by substitution rule states that

$$\int_a^b f(g(x))g'(x) dx = \int_{g(a)}^{g(b)} f(x) dx.\tag{493}$$

Definition of

$$g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto g(x) := n^{1/2}(x - \mu) \text{ with } g'(x) = n^{1/2}\tag{494}$$

and

$$f : \mathbb{R} \rightarrow \mathbb{R}, y \mapsto f(y) := \left(\frac{1}{2\pi}\right)^{1/2} \exp\left(-\frac{1}{2}y^2\right)\tag{495}$$

then allows for re-expressing the integral term on the right-hand side of eq. (492) as follows:

Exercises

Theoretical Exercise 1

Proof

$$\begin{aligned}\mathbb{P}(|\bar{X}_n - \mu| < \epsilon) &= \int_{\mu-\epsilon}^{\mu+\epsilon} \left(\frac{n}{2\pi} \right)^{1/2} \exp \left(-\frac{n}{2} (\bar{x}_n - \mu)^2 \right) d\bar{x}_n \\ &= \int_{\mu-\epsilon}^{\mu+\epsilon} \left(\frac{1}{2\pi} \right)^{1/2} \exp \left(-\frac{1}{2} \left(n^{1/2} (\bar{x}_n - \mu) \right)^2 \right) n^{1/2} d\bar{x}_n \\ &= \int_{\mu-\epsilon}^{\mu+\epsilon} \left(\frac{1}{2\pi} \right)^{1/2} \exp \left(-\frac{1}{2} g(\bar{x}_n)^2 \right) \cdot g'(\bar{x}_n) d\bar{x}_n \\ &= \int_{g(\mu-\epsilon)}^{g(\mu+\epsilon)} \left(\frac{1}{2\pi} \right)^{1/2} \exp \left(-\frac{1}{2} \bar{x}_n^2 \right) d\bar{x}_n \tag{496} \\ &= \int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} \left(\frac{1}{2\pi} \right)^{1/2} \exp \left(-\frac{1}{2} \bar{x}_n^2 \right) d\bar{x}_n \\ &= \int_{-\epsilon\sqrt{n}}^{\epsilon\sqrt{n}} \left(\frac{1}{2\pi} \right)^{1/2} \exp \left(-\frac{1}{2} z^2 \right) dz \\ &= \mathbb{P}(-\epsilon\sqrt{n} < Z < \epsilon\sqrt{n}),\end{aligned}$$

where $Z \sim N(0, 1)$. Because $\mathbb{P}(-\epsilon\sqrt{n} < Z < \epsilon\sqrt{n}) \rightarrow 1$ for $n \rightarrow \infty$, it then follows that also $\mathbb{P}(|\bar{X}_n - \theta| < \epsilon) \rightarrow 1$ for $n \rightarrow \infty$, and hence \bar{X}_n is a consistent estimator of μ .

□

Exercises

Theoretical Exercise 2

Theorem (Consistency of the sample variance)

Let $X_1, \dots, X_n \sim p_\theta$, $\mu := \mathbb{E}(X_i)$ and $\sigma^2 := \mathbb{V}(X_i) < \infty$. Let further

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (497)$$

denote the sample variance estimator of σ^2 . Then S_n^2 is a consistent estimator of σ^2 , if

$$\mathbb{V}(S_n^2) \rightarrow 0 \text{ for } n \rightarrow \infty. \quad (498)$$

Proof

We first recall Chebycheff's inequality

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X)}{x^2}. \quad (499)$$

For $X = S_n^2$, we thus have

$$\mathbb{P}(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathbb{V}(S_n^2)}{\epsilon^2}. \quad (500)$$

Hence, if $\mathbb{V}(S_n^2) \rightarrow 0$ for $n \rightarrow \infty$, then also $\mathbb{P}(|S_n^2 - \sigma^2| \geq \epsilon)$ for every $\epsilon > 0$ and S_n^2 is a consistent estimator of σ^2 .

□

Theoretical Exercise 3

Theorem (Consistency of the sample standard deviation)

Let $X_1, \dots, X_n \sim p_\theta$, $\mu := \mathbb{E}(X_i)$, $\sigma^2 := \mathbb{V}(X_i) < \infty$, and let S_n^2 denote the sample variance. If the sample variance is a consistent estimator of σ^2 , then the sample standard deviation S_n is a consistent estimator of σ .

Proof

We first note that the continuous mapping theorem states that if X_1, X_2, \dots converges in probability to a random variable X and f is a continuous function, then $f(X_1), f(X_2), \dots$ converges in probability to $f(X)$. Because $\sqrt{\cdot}$ is a continuous function, it thus follows that if $S_2^2, S_3^2, S_4^2, \dots$ converges in probability to σ^2 , then so does S_2, S_3, S_4, \dots to σ .

□

Exercises

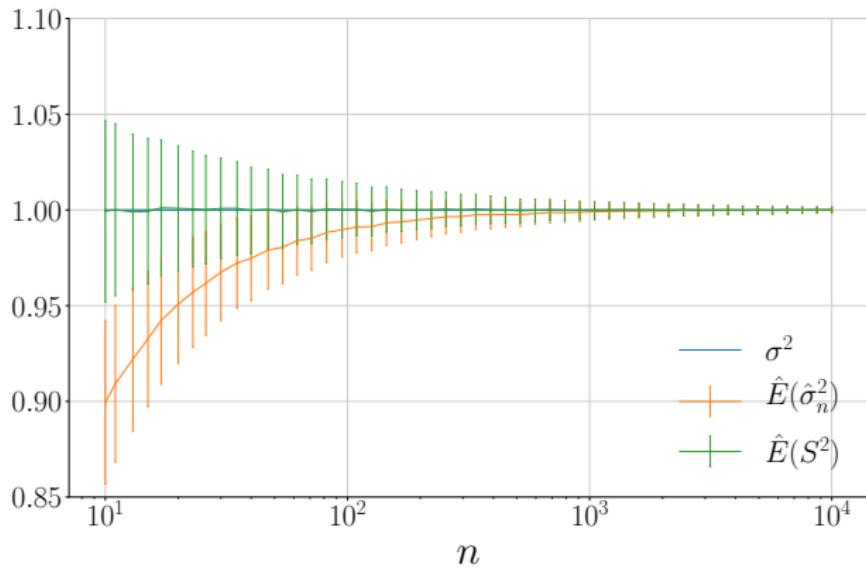
Programming exercises

1. Write a simulation that verifies the asymptotic unbiasedness of the maximum likelihood estimator for the variance parameter of a univariate Gaussian distribution. Include a verification of the unbiasedness of the sample variance.
2. Write a simulation that verifies the asymptotic efficiency of the maximum likelihood estimator for the parameter of a Bernoulli distribution.
3. Write a simulation that verifies the asymptotic efficiency of the maximum likelihood estimator for the variance parameter of a univariate Gaussian distribution.

Exercises

Programming Exercise 1

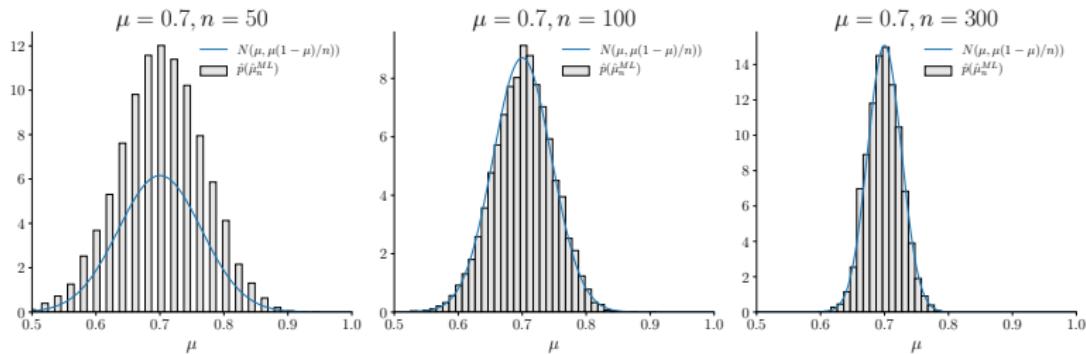
Asymptotic unbiasedness of $\hat{\sigma}_n^{2\text{ML}}$.



Exercises

Programming Exercise 2

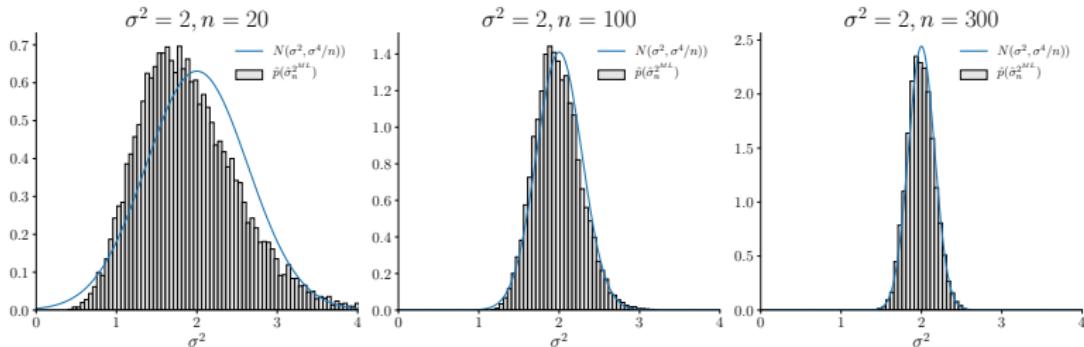
Asymptotic efficiency of $\hat{\mu}_n^{\text{ML}}$ for $\text{Bern}(\mu) \Leftrightarrow \hat{\mu}_n^{\text{ML}} \xrightarrow{a} N\left(\mu, \frac{\mu(1-\mu)}{n}\right)$



Exercises

Programming Exercise 3

Asymptotic efficiency of $\hat{\sigma}_n^{2\text{ML}}$ for $N(\mu, \sigma^2) \Leftrightarrow \hat{\sigma}_n^{2\text{ML}} \xrightarrow{a} N\left(\sigma^2, \frac{2\sigma^4}{n}\right)$



(11) Confidence intervals

Selected statistics

Confidence intervals

- Definition, interpretation, and pivots
- Exact confidence intervals
- Approximate confidence intervals
- Exercises

Selected statistics

Confidence intervals

- Definition, interpretation, and pivots
- Exact confidence intervals
- Approximate confidence intervals
- Exercises

Theorem (Z statistic)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ be a normally distributed random sample and let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \quad (501)$$

denote the sample mean. Then the *Z statistic* is defined as

$$Z := \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu) \quad (502)$$

and is distributed according to standard normal distribution, $Z \sim N(0, 1)$. A PDF for Z is thus given given by

$$N(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right). \quad (503)$$

We denote the associated CDF and inverse CDF by Φ and Φ^{-1} , respectively.

- For a proof, see Casella and Berger (2012, p. 218 - 219)

Theorem (U statistic)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ be a normally distributed random sample and let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \text{ and } S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (504)$$

denote the sample mean and sample variance, respectively. Then the *U statistic* is defined as

$$U := \frac{n-1}{\sigma^2} S_n^2 \quad (505)$$

and is distributed according to a chi-squared distribution with $n-1$ degrees of freedom, $U \sim \chi^2(n-1)$. A PDF for U is thus given by (cf. Lecture (5))

$$\chi^2(u; n-1) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-1}{2}}} u^{\frac{n-1}{2}-1} \exp\left(-\frac{1}{2}u\right). \quad (506)$$

We denote the associated CDF and inverse CDF by Ξ^2 and Ξ^{2-1} , respectively.

- For a proof, see Casella and Berger (2012, p. 219)

Theorem (T statistic)

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ be a normally distributed random sample and let

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \text{ and } S_n := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2} \quad (507)$$

denote the sample mean and sample standard deviation, respectively. Then the *T statistic* is defined as

$$T := \sqrt{n} \frac{\bar{X}_n - \mu}{S_n} \quad (508)$$

and is distributed according to a t-distribution with $n - 1$ degrees of freedom, $T \sim t(n - 1)$. A PDF for T is thus given by (cf. Lecture (5))

$$t(T; n - 1) := \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\Gamma\left(\frac{n-1}{2}\right)} \left(1 + \frac{T^2}{n-1}\right)^{-\frac{n}{2}}. \quad (509)$$

We denote the associate CDF and inverse CDF by ψ and ψ^{-1} , respectively.

- A short proof is provided as Theoretical Exercise 1.

Theorem (Wald statistic)

Let $X_1, \dots, X_n \sim p_\theta$ be a random sample, let $\hat{\theta}_n^{\text{ML}}$ denote the maximum likelihood estimator for θ , and let $I_n(\hat{\theta}_n^{\text{ML}})$ and $J_n(\hat{\theta}_n^{\text{ML}})$ denote the observed Fisher information and the observed expected Fisher information of the random sample, respectively. Then the *Wald statistics* are defined as

$$W_n^I := \sqrt{I_n(\hat{\theta}_n^{\text{ML}})} (\hat{\theta}_n^{\text{ML}} - \theta) \quad \text{and} \quad W_n^J := \sqrt{J_n(\hat{\theta}_n^{\text{ML}})} (\hat{\theta}_n^{\text{ML}} - \theta) \quad (510)$$

and are asymptotically distributed according to standard normal distributions, $W_n^I \xrightarrow{a} N(0, 1)$ and $W_n^J \xrightarrow{a} N(0, 1)$. A PDF for the asymptotic distributions of W_n^I and W_n^J is thus given by

$$N(w_n; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w_n^2\right). \quad (511)$$

We denote the associated CDF and inverse CDF by Φ and Φ^{-1} , respectively.

- For a proof, see Held and Sabanés Bové (2014, Sections 4.2.3) and Lecture (10).

Selected statistics

Confidence intervals

- Definition, interpretation, and pivots
- Exact confidence intervals
- Approximate confidence intervals

Bibliographic remarks

The presented material follows Held and Sabanés Bové (2014, Sections 3.2.2 - 3.2.3) for the definitions of confidence intervals and pivots, Moeschlin (2000a, Sections 4.1 - 4.5) for the exact confidence interval examples, and Wasserman (2004, Sections 6.3.2 and 9.7) for the interpretation of confidence intervals and the notion of approximate confidence intervals, respectively.

Selected statistics

Confidence intervals

- **Definition, interpretation, and pivots**
- Exact confidence intervals
- Approximate confidence intervals

The standard problems of frequentist inference

(1) *Parameter estimation*

The aim of parameter estimation is to find a best guess for the true, but unknown, parameter value of the model, typically based on the observation of $X_1, \dots, X_n \sim p_\theta$.

(2) *Confidence interval evaluation*

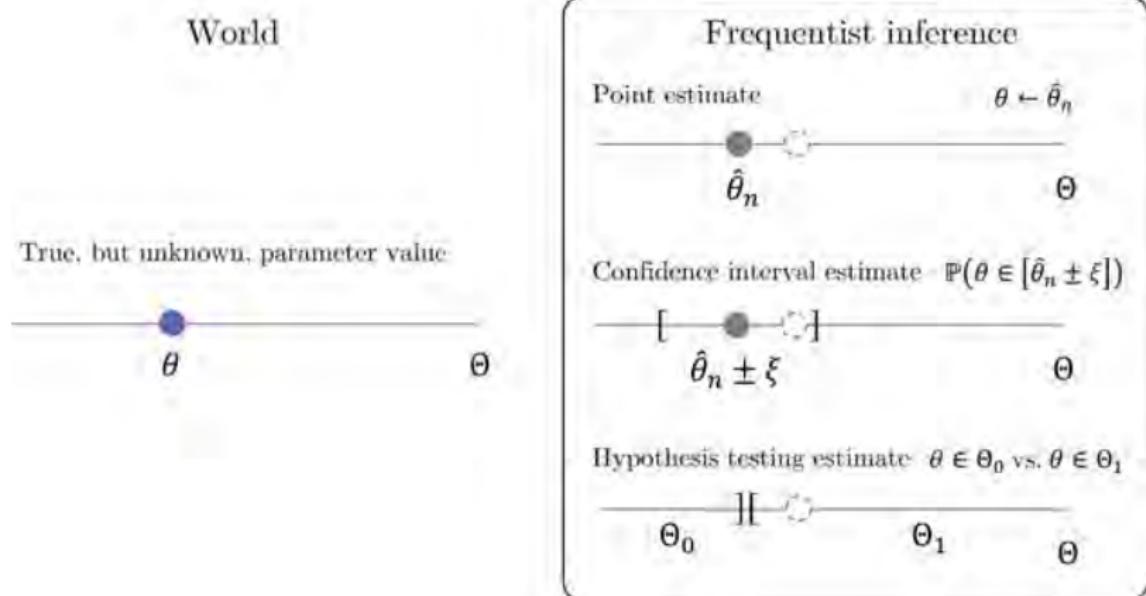
The aim of confidence interval evaluation is to provide a quantitative uncertainty statement about a parameter estimate based on the parameter estimator's sampling distribution.

(3) *Hypothesis testing*

The aim of hypothesis testing is to decide, based on the observations X_1, \dots, X_n and in a sensible fashion, whether, the true, but unknown, parameter is in one of two mutually exclusive subsets of the parameter space.

Confidence interval evaluation and hypothesis testing make extensive use of *statistics* $h(X_1, \dots, X_n)$ and their distributional properties.

Definition, interpretation, and pivots



The frequentist sampling intuition

- Let $X_1, \dots, X_n \sim p_\theta$
- Real observed data is considered one possible realization of $X_1, \dots, X_n \sim p_\theta$.
- From a sampling perspective, however, we could sample data and statistics

$$\begin{aligned} & x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)} \text{ and } h\left(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}\right) \\ & x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)} \text{ and } h\left(x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}\right) \\ & x_1^{(3)}, x_2^{(3)}, \dots, x_n^{(3)} \text{ and } h\left(x_1^{(3)}, x_2^{(3)}, \dots, x_n^{(3)}\right) \\ & x_1^{(4)}, x_2^{(4)}, \dots, x_n^{(4)} \text{ and } h\left(x_1^{(4)}, x_2^{(4)}, \dots, x_n^{(4)}\right) \\ & x_1^{(5)}, x_2^{(5)}, \dots, x_n^{(5)} \text{ and } h\left(x_1^{(5)}, x_2^{(5)}, \dots, x_n^{(5)}\right). \\ & \dots \end{aligned}$$

- Frequentist inference is interested in the distributional properties of statistics
- For example, what is the distribution of

$$h(x_1, \dots, x_n) := \left[\bar{X}_n - 1.96 \frac{S_n}{\sqrt{n}}, \bar{X}_n + 1.96 \frac{S_n}{\sqrt{n}} \right] ? \quad (512)$$

Selected statistics

Confidence intervals

- **Definition, interpretation, and pivots**
- Exact confidence intervals
- Approximate confidence intervals

Definition (δ -confidence interval (δ -CI))

Given a parametric statistical model \mathcal{P} with PMF or PDF p_θ , a sample $X = X_1, \dots, X_n \sim p_\theta$, an $\delta \in]0, 1[$, and two statistics $b_l(X)$ and $b_u(X)$, a δ -confidence interval is an interval $C_n = [b_l, b_u]$, such that

$$\mathbb{P}_\theta(\theta \in C_n) = \mathbb{P}_\theta(b_l(X) \leq \theta \leq b_u(X)) = \delta \text{ for all } \theta \in \Theta. \quad (513)$$

δ is called the *confidence level* or *coverage probability*. The random variables $b_l(X)$ and $b_u(X)$ are the lower and upper bounds of the confidence interval.

Remarks

- C_n is a random interval, because $b_l(X)$ and $b_u(X)$ are random variables.
- θ is fixed (not random) and unknown.
- A δ -confidence interval covers the true θ with probability δ .
- Often $\delta = 0.95$, resulting in *95%-confidence intervals*.
- Confidence intervals with $b_l = -\infty$ or $b_u = \infty$ are called *one-sided*.
- Writing $\mathbb{P}_\theta(C_n \ni \theta)$ may be more intuitive.

Two interpretations of δ confidence intervals

- (1) If an experiment is repeated over and over, the δ -confidence interval will contain the true, but unknown, parameter in $\delta \cdot 100\%$ of all cases. More technically, for repeated random samples form a distribution with unknown parameter θ , a δ -confidence interval will cover θ in $\delta \cdot 100\%$ of all cases.
- (2) Consider a sequence of experiments with unrelated parameters $\theta_1, \theta_2, \dots$ and imagine constructing δ -confidence intervals for the sequence of unrelated parameters $\theta_1, \theta_2, \dots$. Then $\delta \cdot 100\%$ of the confidence intervals will contain the true, but unknown, parameter value.

Definition (Exact and approximate Pivots)

An *exact pivot* is a function of the data X_1, \dots, X_n and the true, but unknown, parameter θ with distribution not depending on θ . An *approximate pivot* is a pivot whose distribution does not asymptotically depend on the true parameter θ .

Remarks

- A pivot is a statistic depending on the true θ , with distribution independent of θ .
- Pivots can be used to construct confidence intervals valid for all values of θ .
- Pivots must not depend on nuisance parameters η .
- For nuisance parameters η , the pivot's distribution must not depend on θ nor η .
- CIs based on exact pivots will be referred to as *exact CIs*.
- CIs based on approximate pivots will be referred to as *approximate CIs*.
- Typical exact and approximate pivots are the T and the Wald statistic, respectively.

Selected statistics

Confidence intervals

- Definition, interpretation, and pivots
- **Exact confidence intervals**
- Approximate confidence intervals
- Exercises

The typical construction of confidence intervals commonly involves

1. The definition of the parametric statistical model.
2. The definition of a statistic and the assessment of its distribution.
3. The verification of the confidence condition.
4. The formulation of the confidence interval.

In the following, we demonstrate the above for exact confidence intervals of

1. The expectation of a normal distribution with known variance.
2. The expectation of a normal distribution with unknown variance.
3. The variance of a normal distribution.

Example (Expectation of a normal distribution, variance known)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ denote a random sample of a parametric statistical model with known variance parameter $\sigma^2 > 0$ and unknown expectation parameter $\mu \in \mathbb{R}$. We develop a δ -confidence interval for μ .

2. Statistic and its distribution

We consider the statistic

$$Z := \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu). \quad (514)$$

We have seen previously, that $Z \sim N(0, 1)$. Note that Z is a function of X_1, \dots, X_n (by means of \bar{X}_n) and μ , while its distribution does not depend on μ . Z is thus an exact pivot.

Example (Expectation of a normal distribution, variance known)

3. Confidence condition

For $\delta \in]0, 1[$, let $z_1 := \Phi^{-1}\left(\frac{1-\delta}{2}\right)$ and $z_2 := \Phi^{-1}\left(\frac{1+\delta}{2}\right)$ denote the respective percentiles of $N(0, 1)$. Note that $(1+\delta)/2 - (1-\delta)/2 = \delta$. For example, for $\delta = 0.95$, $z_1 = \Phi^{-1}(0.025) = -1.96$ and $z_2 = \Phi^{-1}(0.975) = 1.96$. Note that with the symmetry of $N(0, 1)$, $z_1 = -z_2$. Then

$$\mathbb{P}(-z_2 \leq Z \leq z_2) = \delta. \quad (515)$$

Thus, also

$$\begin{aligned} \mathbb{P}\left(-z_2 \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \leq z_2\right) &= \mathbb{P}\left(-\frac{\sigma}{\sqrt{n}}z_2 \leq \bar{X}_n - \mu \leq \frac{\sigma}{\sqrt{n}}z_2\right) \\ &= \mathbb{P}\left(-\bar{X}_n - \frac{\sigma}{\sqrt{n}}z_2 \leq -\mu \leq -\bar{X}_n + \frac{\sigma}{\sqrt{n}}z_2\right) \\ &= \mathbb{P}\left(\bar{X}_n + \frac{\sigma}{\sqrt{n}}z_2 \geq \mu \geq \bar{X}_n - \frac{\sigma}{\sqrt{n}}z_2\right) \\ &= \mathbb{P}\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}}z_2 \leq \mu \leq \bar{X}_n + \frac{\sigma}{\sqrt{n}}z_2\right) \\ &= \delta. \end{aligned} \quad (516)$$

Example (Expectation of a normal distribution, variance known)

4. Formulation of the confidence interval

For a random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with known variance parameter σ^2 and unknown expectation parameter μ , $\delta \in]0, 1[$, and $z_\delta := \Phi^{-1} \left(\frac{1+\delta}{2} \right)$, set

$$C_n := \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_\delta, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_\delta \right]. \quad (517)$$

Then

$$\mathbb{P}(\mu \in C_n) = \delta \quad (518)$$

and C_n is a δ -confidence interval for μ . Note that because \bar{X}_n is a random variable, C_n is random.

□

Example (Expectation of a normal distribution, variance unknown)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ denote a random sample from a parametric statistical model with unknown expectation parameter μ and unknown variance parameter $\sigma^2 > 0$.

0. We develop a δ -confidence interval for μ .

2. Statistic and its distribution

We consider the statistic

$$T := \frac{\sqrt{n}}{S_n} (\bar{X}_n - \mu), \text{ where } S_n := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2}. \quad (519)$$

We have seen previously, that $T \sim t(n-1)$. Note that T is a function of X_1, \dots, X_n (by means of \bar{X}_n and S_n), while its distribution does not depend on neither μ and σ^2 . Also note that σ^2 here attains the role of a nuisance parameter, and neither T nor its distribution depend on σ^2 . The T statistic is thus an exact pivot.

Example (Expectation of a normal distribution, variance unknown)

1. Confidence condition

For $\delta \in]0, 1[$, let $t_1 := \psi^{-1}\left(\frac{1-\delta}{2}; n-1\right)$ and $t_2 := \psi^{-1}\left(\frac{1+\delta}{2}; n-1\right)$ denote the respective percentiles of $t(n-1)$, such that $(1+\delta)/2 - (1-\delta)/2 = \delta$. For example, for $n = 10$ and $\delta = 0.95$, $t_1 = \psi^{-1}(0.025; 9) = -2.26$ and $t_2 = \psi^{-1}(0.975; 9) = 2.26$. Note that with the symmetry of $t(t; n-1)$, $t_1 = -t_2$. Then

$$\mathbb{P}(-t_2 \leq T \leq t_2) = \delta. \quad (520)$$

Thus, also

$$\begin{aligned} \mathbb{P}\left(-t_2 \leq \frac{\sqrt{n}}{S_n}(\bar{X}_n - \mu) \leq t_2\right) &= \mathbb{P}\left(-\frac{S_n}{\sqrt{n}}t_2 \leq \bar{X}_n - \mu \leq \frac{S_n}{\sqrt{n}}t_2\right) \\ &= \mathbb{P}\left(-\bar{X}_n - \frac{S_n}{\sqrt{n}}t_2 \leq -\mu \leq -\bar{X}_n + \frac{S_n}{\sqrt{n}}t_2\right) \\ &= \mathbb{P}\left(\bar{X}_n + \frac{S_n}{\sqrt{n}}t_2 \geq \mu \geq \bar{X}_n - \frac{S_n}{\sqrt{n}}t_2\right) \\ &= \mathbb{P}\left(\bar{X}_n - \frac{S_n}{\sqrt{n}}t_2 \leq \mu \leq \bar{X}_n + \frac{S_n}{\sqrt{n}}t_2\right) \\ &= \delta. \end{aligned} \quad (521)$$

Example (Expectation of a normal distribution, variance unknown)

4. Formulation of the confidence interval

For a random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown expectation parameter μ and unknown variance parameter σ^2 , $\delta \in]0, 1[$, and $t_\delta := \psi^{-1} \left(\frac{1+\delta}{2}; n-1 \right)$, set

$$C_n := \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta \right]. \quad (522)$$

Then

$$\mathbb{P}(\mu \in C_n) = \delta \quad (523)$$

and C_n is a δ -confidence interval for μ . Note that because \bar{X}_n and S are random variables, C_n is random.

□

Example (Variance of a normal distribution)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ denote a random sample of a parametric statistical model with unknown variance parameter $\sigma^2 > 0$ and known or unknown expectation parameter $\mu \in \mathbb{R}$. We develop a δ -confidence interval for σ^2 .

2. Statistic and its distribution

We consider the statistic

$$U := \frac{n-1}{\sigma^2} S_n^2, \text{ where } S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (524)$$

We have seen previously, that $U \sim \chi^2(n-1)$. Note that U is a function of X_1, \dots, X_n (by means of S_n^2) and σ^2 , while its distribution does not depend on σ^2 . U is thus an exact pivot.

Example (Variance of a normal distribution)

1. Confidence condition

For $\delta \in]0, 1[$, let $\xi_1 := \Xi^{2^{-1}}\left(\frac{1-\delta}{2}; n-1\right)$ and $\xi_2 := \Xi^{2^{-1}}\left(\frac{1+\delta}{2}; n-1\right)$ denote the respective percentiles of $\chi^2(n-1)$, such that $(1+\delta)/2 - (1-\delta)/2 = \delta$. For example, for $n = 10$ and $\delta = 0.95$, $\xi_1 := \Xi^{2^{-1}}(0.025; 9) = 2.70$ and $\xi_2 := \Xi^{2^{-1}}(0.975; 9) = 19.0$. Then

$$\mathbb{P}(\xi_1 \leq U \leq \xi_2) = \delta. \quad (525)$$

Thus, also

$$\begin{aligned} \mathbb{P}\left(\xi_1 \leq \frac{n-1}{\sigma^2} S_n^2 \leq \xi_2\right) &= \mathbb{P}\left(\xi_1^{-1} \geq \frac{\sigma^2}{(n-1)S_n^2} \geq \xi_2^{-1}\right) \\ &= \mathbb{P}\left(\frac{(n-1)S_n^2}{\xi_1} \geq \sigma^2 \geq \frac{(n-1)S_n^2}{\xi_2}\right) \\ &= \mathbb{P}\left(\frac{(n-1)S_n^2}{\xi_2} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{\xi_1}\right) \\ &= \delta. \end{aligned} \quad (526)$$

Example (Variance of a normal distribution)

4. Formulation of the confidence interval

For a random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown expectation parameter μ and unknown variance parameter σ^2 , $\delta \in]0, 1[$, $\xi_1 := \Xi^{2^{-1}}\left(\frac{1-\delta}{2}; n-1\right)$ and $\xi_2 := \Xi^{2^{-1}}\left(\frac{1+\delta}{2}; n-1\right)$, set

$$C_n := \left[\frac{(n-1)S_n^2}{\xi_2}, \frac{(n-1)S_n^2}{\xi_1} \right]. \quad (527)$$

Then

$$\mathbb{P}(\sigma^2 \in C_n) = \delta \quad (528)$$

and C_n is a δ -confidence interval for σ^2 . Note that because S_n^2 is a random variables, C_n is random.

□

Selected statistics

Confidence intervals

- Definition, interpretation, and pivots
- Exact confidence intervals
- **Approximate confidence intervals**
- Exercises

The typical construction of confidence intervals commonly involves

1. The definition of the parametric statistical model.
2. The definition of a statistic and the assessment of its distribution.
3. The verification of the confidence condition.
4. The formulation of the confidence interval.

We next demonstrate the above for approximate confidence intervals of

1. Maximum likelihood parameter estimators.
2. The MLE of a Bernoulli distribution parameter.

Example (Maximum likelihood estimators)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim p_\theta$ denote a random sample of a parametric statistical model. We develop a δ -confidence interval for θ .

2. Statistic and its distribution

We consider the Wald statistic

$$W_n^J := \sqrt{J_n(\hat{\theta}_n^{\text{ML}})} (\hat{\theta}_n^{\text{ML}} - \theta) \quad (529)$$

We have seen previously, that $W_n^J \xrightarrow{a} N(0, 1)$. Note that W_n^J is a function of θ , while its asymptotic distribution does not depend on θ . W_n^J is thus an approximate pivot.

Exemplary approximate confidence intervals

Example (Maximum likelihood estimators)

3. Confidence condition

For $\delta \in]0, 1[$, let $z_1 := \Phi^{-1}\left(\frac{1-\delta}{2}\right)$ and $z_2 := \Phi^{-1}\left(\frac{1+\delta}{2}\right)$ denote the respective percentiles of $N(0, 1)$. For example, for $\delta = 0.95$, $z_1 = \Phi^{-1}(0.025) = -1.96$ and $z_2 = \Phi^{-1}(0.975) = 1.96$. Note that with the symmetry of $N(0, 1)$, $z_1 = -z_2$. Then

$$\mathbb{P}(-z_2 \leq W_n^J \leq z_2) \rightarrow \delta \text{ for } n \rightarrow \infty. \quad (530)$$

Thus, also

$$\begin{aligned} & \mathbb{P}\left(-z_2 \leq \sqrt{J_n(\hat{\theta}_n^{\text{ML}})} (\hat{\theta}_n^{\text{ML}} - \theta) \leq z_2\right) \\ &= \mathbb{P}\left(-\sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2 \leq \hat{\theta}_n^{\text{ML}} - \theta \leq \sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2\right) \\ &= \mathbb{P}\left(-\hat{\theta}_n^{\text{ML}} - \sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2 \leq -\theta \leq -\hat{\theta}_n^{\text{ML}} + \frac{1}{SE} z_2\right) \\ &= \mathbb{P}\left(\hat{\theta}_n^{\text{ML}} + \sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2 \geq \theta \geq \hat{\theta}_n^{\text{ML}} - \sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2\right) \\ &= \mathbb{P}\left(\hat{\theta}_n^{\text{ML}} - \sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2 \leq \theta \leq \hat{\theta}_n^{\text{ML}} + \sqrt{J_n(\hat{\theta}_n^{\text{ML}})^{-1}} z_2\right) \\ &\rightarrow \delta \text{ for } n \rightarrow \infty. \end{aligned} \quad (531)$$

Example (Maximum likelihood estimators)

4. Formulation of the confidence interval

For a random sample $X_1, \dots, X_n \sim p_\theta$, $\delta \in]0, 1[$, and $z_\delta := \Phi^{-1} \left(\frac{1+\delta}{2} \right)$, set

$$C_n := \left[\hat{\theta}_n^{\text{ML}} - \sqrt{J_n \left(\hat{\theta}_n^{\text{ML}} \right)^{-1}} z_\delta, \hat{\theta}_n^{\text{ML}} + \sqrt{J_n \left(\hat{\theta}_n^{\text{ML}} \right)^{-1}} z_\delta \right]. \quad (532)$$

Then

$$\mathbb{P}(\theta \in C_n) \rightarrow \delta \text{ for } n \rightarrow \infty \quad (533)$$

and C_n is an approximate δ -confidence interval for θ . Note that because $\hat{\theta}_n^{\text{ML}}$ is a random variable, C_n is random.

□

Example (Expectation of a Bernoulli distribution)

Let $X_1, \dots, X_n \sim \text{Bern}(\mu)$ denote a Bernoulli distributed random sample of size n and let

$$\hat{\mu}_n^{\text{ML}} = \frac{1}{n} \sum_{i=1}^n X_i \quad (534)$$

denote the maximum likelihood estimator of μ . Then an approximate 95%-confidence interval for μ is given by

$$C_n = \left[\hat{\mu}_n^{\text{ML}} - 1.96 \sqrt{\frac{\hat{\mu}_n^{\text{ML}} (1 - \hat{\mu}_n^{\text{ML}})}{n}}, \hat{\mu}_n^{\text{ML}} + 1.96 \sqrt{\frac{\hat{\mu}_n^{\text{ML}} (1 - \hat{\mu}_n^{\text{ML}})}{n}} \right], \quad (535)$$

because $z_{0.95} \approx 1.96$ and

$$J_n \left(\hat{\mu}_n^{\text{ML}} \right) = \frac{n}{\hat{\mu}_n^{\text{ML}} (1 - \hat{\mu}_n^{\text{ML}})}. \quad (536)$$

Example (Expectation of a Bernoulli distribution)

Proof

Let $X \sim \text{Bern}(\mu)$. Then the Fisher information of the random variable X is given by

$$\begin{aligned} I(\mu) &:= -\frac{d^2}{d\mu^2} \ell_1(\mu) \\ &= -\frac{d^2}{d\mu^2} \ln p_\mu(x) \\ &= -\frac{d^2}{d\mu^2} (x \ln \mu + (1-x) \ln(1-\mu)) \\ &= -\frac{d}{d\mu} \left(\frac{d}{d\mu} (x \ln \mu + (1-x) \ln(1-\mu)) \right) \\ &= -\frac{d}{d\mu} \left(\frac{x}{\mu} + \frac{(1-x)}{1-\mu} \right) \\ &= -\left(-\frac{x}{\mu^2} - \frac{(1-x)^2}{1-\mu} \right) \\ &= \frac{x}{\mu^2} + \frac{(1-x)^2}{1-\mu}. \end{aligned} \tag{537}$$

Example (Expectation of a Bernoulli distribution)

Proof

Furthermore, the expected Fisher information of the random variable X is given by

$$\begin{aligned} J(\mu) &= \mathbb{E}_\mu(I(\mu)) \\ &= \mathbb{E}_\mu \left(\frac{X}{\mu^2} + \frac{(1-X)^2}{1-\mu} \right) \\ &= \frac{\mathbb{E}_\mu(X)}{\mu^2} + \frac{(1-\mathbb{E}_\mu(X))^2}{1-\mu} \\ &= \frac{\mu}{\mu^2} + \frac{(1-\mu)^2}{1-\mu} \\ &= \frac{1}{\mu(1-\mu)}. \end{aligned} \tag{538}$$

With the additivity property of the expected Fisher information and the definition of the observed Fisher information, it then follows immediately, that

$$J_n(\mu) = \frac{n}{\mu(1-\mu)} \tag{539}$$

and

$$J_n(\hat{\mu}_n^{\text{ML}}) = \frac{n}{\hat{\mu}_n^{\text{ML}}(1-\hat{\mu}_n^{\text{ML}})}, \tag{540}$$

respectively.

□

Selected statistics

Confidence intervals

- Definition, interpretation, and pivots
- Exact confidence intervals
- Approximate confidence intervals
- **Exercises**

Study questions

1. Write down the definition of the T statistic and state its distribution.
2. Write down the definition of the Wald statistics and state their distribution.
3. Define the δ -confidence interval.
4. Give two interpretations of δ -confidence intervals.
5. Define the notions of exact and approximate pivots and δ -confidence intervals.
6. State the steps involved in the typical construction of confidence intervals.
7. Write down the formula of the 95%-confidence interval for the expectation parameter of a univariate Gaussian distribution with known variance.
8. Write down the formula of the 95%-confidence interval for the expectation parameter of a univariate Gaussian distribution with unknown variance.
9. Write down the formula of the 95%-confidence interval for the variance parameter of a univariate Gaussian distribution.
10. Write down the formula of an approximate 95%-confidence interval for a parameter based on a maximum likelihood estimator.

Theoretical exercises

1. Show that the T statistic has a Student's t distribution with $n - 1$ degrees of freedom (Casella and Berger, 2012, Definition 5.3.4 and p.223 - 224).
2. Develop a 95%-confidence interval for the parameter of an exponential distribution (Held and Sabanés Bové, 2014, Example 3.7).
3. By means of example, show that a confidence interval is not a probability statement about a true, but unknown, parameter (Wasserman, 2004, Example 6.14).

Theoretical Exercise 1

We show that the T statistic can be rewritten as the ratio of a standard normal random variable Z and the square root of a chi-squared random variable U divided by its degrees of freedom. With the results of Lecture (5), T is then distributed according to a t distribution with the degrees of freedom of U . We first note that

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S} = \sqrt{n} \frac{(\bar{X} - \mu)/\sigma}{S/\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{S^2/\sigma^2}} \quad (541)$$

Definition of the random variables

$$Z := \sqrt{n} \frac{\bar{X} - \mu}{\sigma} \text{ and } U := \sqrt{S^2/\sigma^2} \quad (542)$$

then yields

$$T = Z/U. \quad (543)$$

With the Z statistics theorem it then follows directly that $Z \sim N(0, 1)$. Moreover, with the Gaussian random sample mean and variance theorem, we have

$$(n - 1) \frac{S^2}{\sigma^2} \sim \chi^2(n - 1) \quad (544)$$

and hence U is the square root of a chi-squared random variable divided by $n - 1$.

Theoretical Exercise 2 (Exponential random variable)

Let X be a continuous random variable with outcome set $\mathcal{X} := \mathbb{R}_{\geq 0}$ and probability density function

$$p : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}, x \mapsto p(x) := \lambda \exp(-\lambda x). \quad (545)$$

Then X is said to be distributed according to an *exponential distribution* with rate parameter $\lambda > 0$ for which we write $X \sim \text{Exp}(\lambda)$. We abbreviate the PDF of an exponential random variable by

$$\text{Exp}(x; \lambda) := \lambda \exp(-\lambda x). \quad (546)$$

The CDF of an exponential random variable is given by

$$P : \mathbb{R}_{\geq 0} \rightarrow [0, 1], x \mapsto P(x) := 1 - \exp(-\lambda x). \quad (547)$$

and we denote its inverse by P^{-1} . The expectation and variance of an exponential random variable are given by

$$\mathbb{E}(X) = \lambda^{-1} \text{ and } \mathbb{V}(X) = \lambda^{-2}, \quad (548)$$

respectively.

Theoretical Exercise 2 (Confidence interval)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim \text{Exp}(\lambda)$ denote a random sample of a parametric statistical model with unknown rate parameter $\lambda > 0$. We develop a δ -confidence interval for λ .

2. Statistic and its distribution

We consider the statistic

$$L := \lambda n \bar{X}_n, \tag{549}$$

for which we note without proof that $L \sim G(n, 1)$. Note that L is a function of X_1, \dots, X_n (by means of \bar{X}_n) and λ , while its distribution does not depend on λ . L is thus an exact pivot.

Theoretical Exercise 2 (Confidence interval)

3. Confidence condition

For $\delta \in]0, 1[$, let $l_1 := P^{-1}\left(\frac{1-\delta}{2}\right)$ and $l_2 := P^{-1}\left(\frac{1+\delta}{2}\right)$ denote the respective percentiles of $G(n, 1)$. Note that $(1 - \delta)/2 + (1 + \delta)/2 = \delta$. For example, for $n = 47$, $l_1 = P^{-1}(0.025) = 34.5$ and $l_2 = P^{-1}(0.975) = 61.3$. We thus have

$$\mathbb{P}(l_1 \leq L \leq l_2) = \delta. \quad (550)$$

Hence, also

$$\begin{aligned} \mathbb{P}(l_1 \leq L \leq l_2) &= \mathbb{P}(l_1 \leq \lambda n \bar{X}_n \leq l_2) \\ &= \mathbb{P}\left(\frac{l_1}{n \bar{X}_n} \leq \lambda \leq \frac{l_2}{n \bar{X}_n}\right) \\ &= \delta. \end{aligned} \quad (551)$$

Theoretical Exercise 2 (Confidence interval)

4. Formulation of the confidence interval

For a random sample $X_1, \dots, X_n \sim \text{Exp}(\lambda)$, $\delta \in]0, 1[$, and

$$l_1 := P^{-1} \left(\frac{1-\delta}{2} \right) \text{ and } l_2 := P^{-1} \left(\frac{1+\delta}{2} \right) \quad (552)$$

set

$$C_n := \left[\frac{l_1}{n\bar{X}_n}, \frac{l_2}{n\bar{X}_n} \right]. \quad (553)$$

Then

$$\mathbb{P}(\lambda \in C_n) = \delta \quad (554)$$

and C_n is a δ -confidence interval for λ . Note that because \bar{X}_n is a random variable, C_n is random.

Theoretical Exercise 3

Consider the a discrete random variable with outcome space $\mathcal{X}_\theta := \{\theta - 1, \theta + 1\}$ for $\theta \in \mathbb{R}$ and PMF

$$p_\theta : \mathcal{X} \rightarrow [0, 1], x \mapsto p_\theta(x) := \begin{cases} \frac{1}{2} & \text{if } x = \theta - 1 \\ \frac{1}{2} & \text{if } x = \theta + 1 \end{cases}. \quad (555)$$

Let $X_1, X_2 \sim p_\theta$ be a random sample and consider the confidence interval

$$C := \begin{cases} [\frac{1}{2}(X_1 + X_2), \frac{1}{2}(X_1 + X_2)] & \text{if } X_1 \neq X_2 \\ [X_1 - 1, X_1 + 1] & \text{if } X_1 = X_2 \end{cases}. \quad (556)$$

Then C is a 0.75 confidence interval for θ , i.e.,

$$\mathbb{P}(\theta \in C) = 0.75 \text{ and } \mathbb{P}(\theta \notin C) = 0.25. \quad (557)$$

To see this, we first note that C can equivalently be written as

$$C := \begin{cases} \{\frac{1}{2}(X_1 + X_2)\} & \text{if } X_1 \neq X_2 \\ \{X_1 - 1\} & \text{if } X_1 = X_2 \end{cases}. \quad (558)$$

Theoretical Exercise 3

We next note, that we have the following event structure

X_1	X_2	$X_1 - 1$	$\frac{1}{2}(X_1 + X_2)$	C	c	θ
$\theta - 1$	$\theta - 1$	$\theta - 2$	$\theta - 1$	$\{X_1 - 1\}$	$\{\theta - 2\}$	$\theta \notin c$
$\theta - 1$	$\theta + 1$	$\theta - 2$	θ	$\{\frac{1}{2}(X_1 + X_2)\}$	$\{\theta\}$	$\theta \in c$
$\theta + 1$	$\theta - 1$	θ	θ	$\{\frac{1}{2}(X_1 + X_2)\}$	$\{\theta\}$	$\theta \in c$
$\theta + 1$	$\theta + 1$	θ	$\theta + 1$	$\{X_1 - 1\}$	$\{\theta\}$	$\theta \in c$

We thus have

$$\mathbb{P}(\theta \in C) = 1 - \mathbb{P}(\theta \notin C) = 1 - \mathbb{P}(X_1 = \theta - 1, X_2 = \theta - 1) = 1 - 0.25 = 0.75. \quad (559)$$

Finally, consider the case that $X_1 = 15$ and $X_2 = 17$. Then it must hold that $\theta = 16$. However,

$$c = \left\{ \frac{1}{2}(15 + 17) \right\} = \{16\} \quad (560)$$

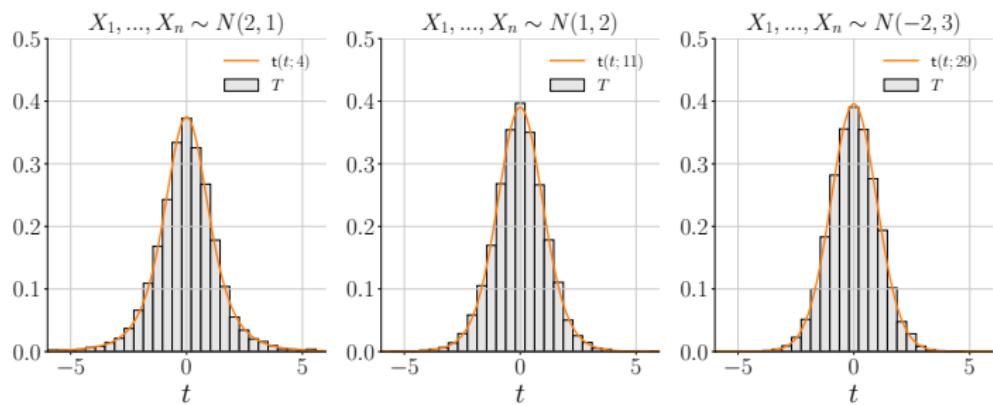
remains a 0.75 confidence interval (only).

Programming exercises

1. Write a simulation that verifies that the T statistic is distributed according to a t-distribution with $n - 1$ degrees of freedom.
2. Write a simulation that verifies that the 95%-confidence interval for the expectation parameter of a Gaussian distribution with unknown variance comprises the true, but unknown, expectation parameter in $\approx 95\%$ of its realizations.
3. Write a simulation that verifies that the approximate 95%-confidence interval for the expectation parameter of a Bernoulli distribution comprises the true, but unknown, expectation parameter in $\approx 95\%$ of its realizations.

Exercises

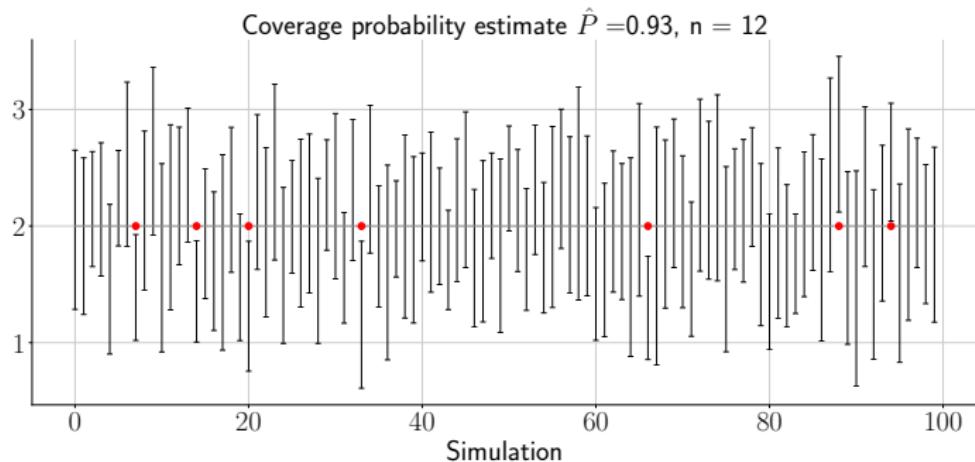
Programming Exercise 1



Programming Exercise 2

For $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, $\delta := 0.95$, and $t_\delta := \psi^{-1} \left(\frac{1+\delta}{2}; n-1 \right)$, we consider

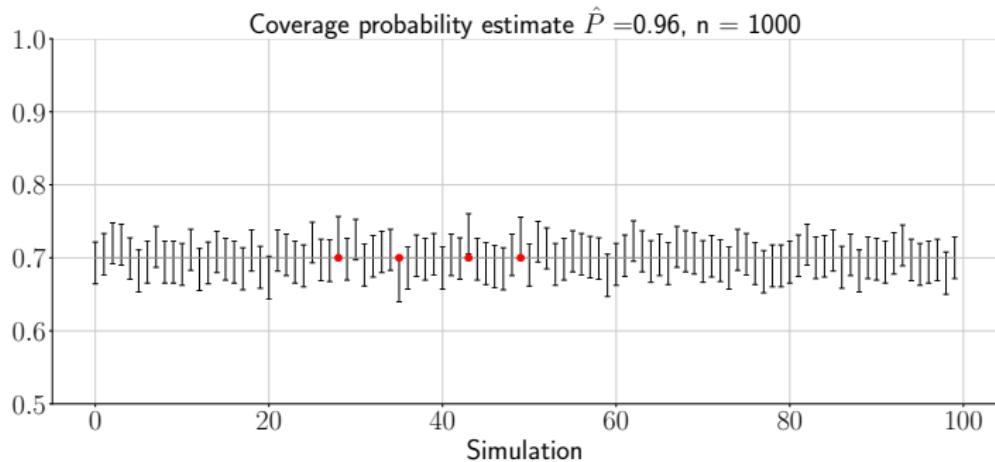
$$C_n := \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta \right]. \quad (561)$$



Programming Exercise 3

For $X_1, \dots, X_n \sim \text{Bern}(\mu)$, $\delta := 0.95$, and $z_\delta := \Phi^{-1} \left(\frac{1+\delta}{2}; n-1 \right)$, we consider

$$C_n := \left[\hat{\mu}_n^{\text{ML}} - \frac{\hat{\mu}_n^{\text{ML}}(1 - \hat{\mu}_n^{\text{ML}})}{n} z_\delta, \hat{\mu}_n^{\text{ML}} + \frac{\hat{\mu}_n^{\text{ML}}(1 - \hat{\mu}_n^{\text{ML}})}{n} z_\delta, \right]. \quad (562)$$



(12) Hypothesis testing

Bibliographic remarks

The presented material follows Ostwald et al. (2019, Supplementary Material, Section 2) for the majority of test-theoretical concepts. The discussion of the Wald test follows Wasserman (2004, Section 10.1), the development of the duality of confidence intervals and hypotheses tests is based on Czado and Schmidt (2011, Section 5.3). For an excellent overview on contemporary views of hypothesis testing, see "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$ " (2019) *The American Statistician*.

Hypothesis testing

- Foundations
- Test construction and examples
 - T test
 - Wald test
- Confidence intervals and hypotheses tests
- Exercises

Hypothesis testing

- **Foundations**
- Test construction and examples
 - T test
 - Wald test
- Confidence intervals and hypotheses tests
- Exercises

World

True, but unknown, parameter value

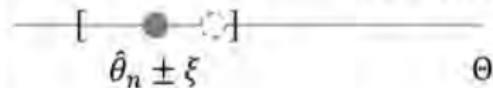


Frequentist inference

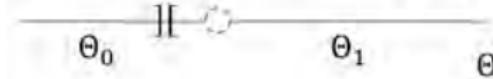
Point estimate



Confidence interval estimate $\mathbb{P}(\theta \in [\hat{\theta}_n \pm \xi])$



Hypothesis testing estimate $\theta \in \Theta_0$ vs. $\theta \in \Theta_1$



Definition (Test hypotheses)

Let \mathcal{P} denote a parametric statistical model governing the distribution of a random sample $X = (X_1, \dots, X_n)$ with a PMF or PDF p_θ , let \mathcal{X} denote the outcome space of the data such that $x \in \mathcal{X}$, and let Θ denote the parameter space of the model. Further, let Θ_0 and Θ_1 denote a partition of the parameter space, such that $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$. Then a *test hypothesis* is a statement about the parameter governing the distribution of X in relation to the parameter space subsets Θ_0 and Θ_1 . Specifically

- $H_0 : \theta \in \Theta_0$ is referred to as *null hypothesis*, and
- $H_1 : \theta \in \Theta_1$ is referred to as *alternative hypothesis*.

Remarks

- We assume that both null and alternative hypothesis exist.
- The null hypothesis is not necessarily the hypothesis $\Theta_0 = \{0\}$.
- The null hypothesis is the hypothesis one is willing to reject.

Definition (Simple and composite hypotheses)

- A *simple hypothesis* refers to a subset of parameter space containing a single element, such as $\Theta_0 := \{\theta_0\}$.
- A *composite hypothesis* refers to a subset of parameter space containing more than one element, such as $\Theta_0 := \mathbb{R}_{\leq 0}$.

Remark

- The often encountered null hypothesis $\Theta_0 = \{0\}$ is an example for a simple hypothesis and is also referred to as *nil hypothesis*.

Definition (Test)

Given a test hypotheses scenario, a *test* is a mapping from the data outcome space to the set $\{0, 1\}$. Formally,

$$\phi : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi(x), \quad (563)$$

where

- 0 represents the act of not rejecting the null hypothesis.
- 1 represents the act of rejecting the null hypothesis.

Remarks

- Rejecting the null hypothesis \Leftrightarrow Accepting the alternative hypothesis.
- Not rejecting the null hypothesis \Leftrightarrow Rejecting the alternative hypothesis.
- Accepting the null hypothesis \Leftrightarrow Rejecting the alternative hypothesis.
- Because X is a random variable, ϕ is also a random variable.

Definition (Standard test)

A *standard test* is given by the composition of a *test statistic*

$$\gamma : \mathcal{X} \rightarrow \mathbb{R} \quad (564)$$

and a *decision rule*

$$\delta : \mathbb{R} \rightarrow \{0, 1\} \quad (565)$$

A standard test can be written as

$$\phi := \delta \circ \gamma : \mathcal{X} \rightarrow \{0, 1\} \quad (566)$$

Remarks

- Because X is random, both γ and δ are random.

Definition (Test rejection region)

The subset of the test statistic's outcome space for which the test takes on the value 1 is referred to as the *rejection region* R of the test. Formally,

$$R := \{\gamma(X) \in \mathbb{R} \mid \phi(X) = 1\} \subset \mathbb{R}. \quad (567)$$

Remarks

- The events $\phi(X) = 1$ and $\gamma(X) \in R$ are equivalent.
- The events $\phi(X) = 1$ and $\gamma(X) \in R$ have the same probability.

Definition (One-sided and two-sided critical value-based tests)

A *critical value-based test* is a standard test with a critical value $c \in \mathbb{R}$ -dependent decision rule.

- A *one-sided* critical value-based test takes the form

$$\phi : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi(x) := 1_{\{\gamma(x) \geq c\}} = \begin{cases} 1 & \gamma(x) \geq c \\ 0 & \gamma(x) < c \end{cases} \quad (568)$$

- A *two-sided* critical value-based test takes the form

$$\phi : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi(x) := 1_{\{|\gamma(x)| \geq c\}} = \begin{cases} 1 & |\gamma(x)| \geq c \\ 0 & |\gamma(x)| < c \end{cases} \quad (569)$$

Remark

- *T tests* are familiar examples of critical value-based tests: using the sample mean and sample standard deviation, a realization of the data X is first transformed into the value of the T statistic, whose size is then compared to a critical value in order to decide for rejecting the null hypothesis or not.

Definition (Test errors)

When conducting a hypothesis test, two kinds of errors can occur:

- Rejecting the null hypothesis ($\phi(X) = 1$), when the null hypothesis is in fact true ($\theta \in \Theta_0$), is referred to as a *Type I error*.
- Not rejecting the null hypothesis ($\phi(X) = 0$), when the null hypothesis is in fact false ($\theta \in \Theta_1$), is referred to as a *Type II error*.

Remark

- Type I errors are usually considered more detrimental than Type II errors.

Definition (Test error probabilities)

- The probability of a Type I error is referred to as the *size* of a test and commonly denoted by $\alpha = [0, 1]$, $\alpha := \mathbb{P}_{\Theta_0}(\phi(X) = 1)$. Its complementary probability $\mathbb{P}_{\Theta_0}(\phi(X) = 0) = 1 - \alpha$ is referred to as the *specificity* of a test.
- The probability of a Type II error $P_{\Theta_1}(\phi(X) = 0)$ lacks a common denomination. Its complementary probability $\beta := \mathbb{P}_{\Theta_1}(\phi(X) = 1)$ is referred to as the *power* of a test.

Remarks

- The \mathbb{P} subscripts Θ_0 and Θ_1 indicate that null/alternative hypothesis hold.
- The size of a test is also referred to as the Type I error rate.
- The probability of a Type II error is sometimes denoted by β , but this is inconsistent with the definition of the power function.

Definition (Significance level, conservative, exact, and liberal tests)

A test is said to be of *significance level* $\alpha' \in [0, 1]$, if its size α is smaller than or equal to α' , i.e., if

$$\alpha \leq \alpha'. \quad (570)$$

- A test is called *conservative*, if $\alpha \leq \alpha'$.
- A test is called *exact*, if $\alpha = \alpha'$.
- A test is called *liberal*, if $\alpha > \alpha'$.

Remarks

- The size and the significance level of a test are two different things.
- A liberal test is not of significance level α' .

Definition (Test quality and power function)

For a test ϕ , the *test quality function* is defined as

$$q : \Theta \rightarrow [0, 1], \theta \mapsto q(\theta) := \mathbb{E}_{\mathbb{P}_\theta}(\phi(X)). \quad (571)$$

For $\theta \in \Theta_1$, the test quality function is also referred to as the test's *power function*, and is denoted by

$$\beta : \Theta_1 \rightarrow [0, 1], \theta \mapsto \beta(\theta) := \mathbb{P}_{\Theta_1}(\phi(X) = 1). \quad (572)$$

Remarks

- The test quality function summarizes a test's size and power as function of θ .
- For $\theta \in \Theta_0$, the test quality function value evaluates to

$$\mathbb{E}_{\mathbb{P}_{\Theta_0}}(\phi(X)) = 0 \cdot \mathbb{P}_{\Theta_0}(\phi(X) = 0) + 1 \cdot \mathbb{P}_{\Theta_0}(\phi(X) = 1) = \mathbb{P}_{\Theta_0}(\phi(X) = 1) = \alpha \quad (573)$$

- For $\theta \in \Theta_1$, the test quality function value evaluates to

$$\mathbb{E}_{\mathbb{P}_{\Theta_1}}(\phi(X)) = 0 \cdot \mathbb{P}_{\Theta_1}(\phi(X) = 0) + 1 \cdot \mathbb{P}_{\Theta_1}(\phi(X) = 1) = \mathbb{P}_{\Theta_1}(\phi(X) = 1) = \beta \quad (574)$$

Hypothesis testing

- Foundations
- **Test construction and examples**
 - T test
 - Wald test
- Confidence intervals and hypotheses tests
- Exercises

Test construction

- Because the Type I error rate of a test is considered more important than the Type II error rate of a test, the test size is usually fixed first, e.g., by selecting a significance level such as $\alpha' = 0.05$ and an associated critical value $c_{\alpha'}$ of the test statistic.
- Given a desired significance level, different tests or statistical models (e.g., sample sizes) are then compared in their ability to minimize the probability of the test's Type II error, i.e., maximize the test's power.

Test construction

The construction of a test thus typically involves

1. The definition of a parametric statistical model.
2. The definition of the test hypotheses, test statistic, and test.
3. The assessment of the test statistic distribution.
4. The establishment of Type I error rate control.
5. The assessment of the test's power function.

Remarks

- We demonstrate steps 1. to 5. for a T test
- We demonstrate steps 1. to 4. for a Wald test

Hypothesis testing

- Foundations
- **Test construction and examples**
 - T test
 - Wald test
- Confidence intervals and hypotheses tests
- Exercises

Example (T test)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ denote a random sample from a parametric statistical model with unknown expectation parameter μ and unknown variance parameter $\sigma^2 > 0$.

2. Test hypotheses, test statistic, and test

For the parameter space of the expectation parameter $\Theta := \mathbb{R}$, we consider the test hypotheses

$$\mu \in \Theta_0 :=]-\infty, \mu_0] \text{ and } \mu \in \Theta_1 :=]\mu_0, \infty[\quad (575)$$

A standard one-sided test can then be constructed by considering the test statistic

$$T(X) := \frac{\sqrt{n}}{S_n} (\bar{X}_n - \mu_0) \quad (576)$$

and the test

$$\phi(X) := 1_{\{T(X) \geq c\}}. \quad (577)$$

Example (T test)

3. Test statistic distribution

We have seen previously that for $\mu \in \Theta_0$, $T \sim t(n - 1)$.

4. Type I error rate control

Given the form of the current test and the invertibility of the CDF $\psi(t)$, $\phi(X)$ can be rendered a test of significance level α' by choosing the critical value

$$c_{\alpha'} := \psi^{-1}(1 - \alpha'; n - 1) \quad (578)$$

For example, for $n = 10$ and $\alpha' = 0.05$, $c_{0.05} = \psi^{-1}(0.95; 9) = 1.83$. Rejecting the null hypothesis for an observed T statistic equal to or larger than $c_{0.05} = 1.83$ thus ensures a test size of $\alpha = 0.05$.

Example (T test)

5. Test quality function

The test quality function of the thus defined test can be written as

$$q : \mathbb{R} \rightarrow [0, 1], \mu \mapsto q(\mu) := 1 - \psi_{\delta, n-1}^{-1}(1 - \alpha'), \quad (579)$$

where

- $\psi_{\delta, n-1}$ is the CDF of the non-central t distribution with $n-1$ degrees of freedom,
- $\delta := \sqrt{nd}$ denotes its non-centrality parameter with $d := \frac{\mu - \mu_0}{\sigma}$, and
- ψ_{n-1}^{-1} is the inverse CDF of the t distribution.

Note that in addition to the true, but unknown, value of μ , q depends on

- the significance level α' ,
- the true, but unknown, variance parameter μ and σ^2 ,
- and the sample size n

Hypothesis testing

- Foundations
- **Test construction and examples**
 - T test
 - **Wald test**
- Confidence intervals and hypotheses tests
- Exercises

Example (Wald test)

1. Parametric statistical model

Let $X_1, \dots, X_n \sim p_\theta$ denote a random sample from a parametric statistical model with unknown parameter $\theta \in \Theta$. Let $\hat{\theta}_n$ denote an asymptotically normally distributed estimator for θ , for example $\hat{\theta}_n^{ML}$.

2. Test hypotheses, test statistic, and test

We consider the test hypotheses

$$\theta \in \Theta_0 := \theta_0 \text{ and } \theta \in \Theta_1 := \Theta \setminus \theta_0 \quad (580)$$

A standard two-sided test can then be constructed by considering the test statistic

$$W(X) := \sqrt{J_n(\hat{\theta}_n^{ML})} (\hat{\theta}_n^{ML} - \theta_0) \quad (581)$$

and the test

$$\phi(X) := 1_{\{|W(X)| \geq c\}}. \quad (582)$$

Example (Wald test)

3. Test statistic distribution

We have seen previously, that $W \xrightarrow{a} N(0, 1)$, i.e, its asymptotic distribution for $\theta \in \Theta_0$ is given in terms of the PDF $N(w; 0, 1)$.

4. Type I error rate control

Given the form of the current test and the symmetry of $N(w; 0, 1)$, $\phi(X)$ can be rendered an asymptotically exact test of significance level α' by choosing the critical value

$$c_{\alpha'} := \Phi^{-1} \left(1 - \frac{\alpha'}{2} \right). \quad (583)$$

Hypothesis testing

- Foundations
- Test construction and examples
 - T test
 - Wald test
- **Confidence intervals and hypotheses tests**
- Exercises

Theorem (Duality of confidence intervals and hypotheses tests I)

Let \mathcal{P} denote parametric statistical model governing the distribution of a random sample $X = (X_1, \dots, X_n)$ with outcome space \mathcal{X} and with PMF or PDF p_θ for $\theta \in \Theta \subseteq \mathbb{R}$. Let $[b_l(X), b_u(X)]$ be a δ -confidence interval for θ . Then the test defined by

$$\phi_\theta : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi_\theta(x) := \begin{cases} 0, & \theta \in [b_l(x), b_u(x)] \\ 1, & \theta \notin [b_l(x), b_u(x)] \end{cases} \quad (584)$$

is a test of significance level $\alpha' = 1 - \delta$ for the hypotheses

$$\Theta_0 := \{\theta\} \text{ and } \Theta_1 := \Theta \setminus \{\theta\}. \quad (585)$$

Proof

The significance level $\alpha' = 1 - \delta$ of the test follows from

$$\alpha' \geq \alpha = \mathbb{P}_\theta(\phi(X) = 1) = \mathbb{P}_\theta(\theta \notin [b_l(x), b_u(x)]) = 1 - \mathbb{P}_\theta(\theta \in [b_l(x), b_u(x)]) = 1 - \delta. \quad (586)$$

□

Remark

- δ -confidence intervals can be used to construct tests of significance level $1 - \delta$.

Confidence intervals and hypotheses tests

Theorem (Duality of confidence intervals and hypotheses tests II)

Under the assumptions of the previous Theorem, let

$$\Phi := \{\phi_\theta(X) | \theta \in \Theta\} \quad (587)$$

be a family of tests, such that $\phi_\theta(X)$ is an exact test of significance level α' for the hypotheses

$$\Theta_0 := \{\theta\} \text{ and } \Theta_1 := \Theta \setminus \{\theta\}. \quad (588)$$

Assume further, that the set

$$C := \{\theta \in \Theta | \phi_\theta(X) = 0\} \quad (589)$$

can be written as $C = [b_l(X), b_u(X)]$ for appropriately determined $b_l(X)$ and $b_u(X)$. Then C is a $\delta := 1 - \alpha'$ confidence interval for θ .

Proof

The confidence level $\delta = 1 - \alpha'$ of the confidence interval follows from

$$\delta = \mathbb{P}_\theta(\theta \in C) = \mathbb{P}_\theta(\phi_\theta(X) = 0) = 1 - \mathbb{P}_\theta(\phi_\theta(X) = 1) = 1 - \alpha = 1 - \alpha'. \quad (590)$$

□

Remark

- Tests of significance level α' can be used to construct a $1 - \alpha'$ confidence intervals.

Example (Constructing a hypothesis test from a confidence interval)

We have previously shown that for a random sample $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ with unknown expectation parameter μ and unknown variance parameter σ^2 , $\delta \in]0, 1[$, and $t_\delta := \psi^{-1} \left(\frac{1+\delta}{2}; n-1 \right)$,

$$C_n := \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta \right]. \quad (591)$$

is a δ -confidence interval. With the duality of confidence intervals and hypotheses tests, we may thus define the test

$$\phi_\theta : \mathcal{X} \rightarrow \{0, 1\}, x \mapsto \phi_\theta(x) := \begin{cases} 0, & \mu \in \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta \right] \\ 1, & \mu \notin \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta \right] \end{cases} \quad (592)$$

for the hypotheses $\Theta_0 = \mu$ and $\Theta_1 = \mathbb{R} \setminus \mu$. Then

$$\begin{aligned} \mathbb{P}_\mu (\phi_\mu(X) = 1) &= 1 - \mathbb{P}_\mu (\phi_\mu(X) = 0) \\ &= 1 - \mathbb{P}_\mu \left(\mu \in \left[\bar{X}_n - \frac{S}{\sqrt{n}} t_\delta, \bar{X}_n + \frac{S}{\sqrt{n}} t_\delta \right] \right) \\ &= 1 - \delta. \end{aligned} \quad (593)$$

We thus confirmed that ϕ_θ is a test with significance level $\alpha' = 1 - \delta$.

Hypothesis testing

- Foundations
- Test construction and examples
 - T test
 - Wald test
- Confidence intervals and hypotheses tests
- **Exercises**

Study questions

1. Define the notions of test hypotheses, as well as simple, composite, and nil hypotheses.
2. Define the notion of a statistical test and of a standard statistical test.
3. Write down the definition of a one-sided critical value-based test.
4. Write down the definition of a two-sided critical value-based test.
5. Define the notions of Type I and Type II test errors.
6. Define the size, specificity, power, and significance level of a test.
7. Define the notions of a conservative, exact, and liberal test.
8. Write down the definition of the test quality and power function.
9. State the typical procedure for constructing a hypothesis test.
10. Formulate the duality of confidence intervals and hypotheses tests.

Exercises

Theoretical exercises

1. Derive the test quality function for the T test example.

Programming exercises

1. By means of simulation, show that a two-sided T test with simple null hypothesis $\Theta_0 := \{\mu_0\}$ of significance level α' is exact.
2. By means of simulation, demonstrate that the δ -confidence interval-based test for the expectation parameter of univariate Gaussian distribution is of significance level $\alpha' = 1 - \delta$.
3. By means of simulation, validate the power function of the T test .

Exercises

Theoretical Exercise 1

Definition (Non-central T random variable)

Let $Z \sim N(0, 1)$ and $V \sim \chi^2(n)$ denote two independent random variables and let $\delta \in \mathbb{R}$. Then the distribution of the random variable

$$T_\delta := \frac{Z + \delta}{\sqrt{V/n}} \quad (594)$$

is called *non-central t-distribution* and T_δ is called a *non-central T random variable* with non-centrality parameter δ and n degrees of freedom.

Theorem

A PDF for a non-central T random variable T_δ is (Lehmann, 1986, p. 254)

$$\begin{aligned} p : \mathbb{R} \rightarrow \mathbb{R}_{>0}, t_\delta \mapsto p(t_\delta) &:= \frac{1}{2^{\frac{1}{2}(n-1)} \Gamma(\frac{1}{2}n) \sqrt{\pi n}} \\ &\times \int_0^\infty y^{\frac{1}{2}(n-1)} \exp\left(-\frac{1}{2}y\right) \exp\left(-\frac{1}{2}\left(t_\delta \sqrt{\frac{y}{n}} - \delta\right)^2\right) dy \end{aligned} \quad (595)$$

We denote the associated CDF and inverse CDF by $\psi_{\delta,n}$ and $\psi_{\delta,n}^{-1}$.

Exercises

Theoretical Exercise 1

We first note that by definition

$$q : \Theta \rightarrow [0, 1], \theta \mapsto q(\theta) := \mathbb{E}_{\mathbb{P}_\theta}(\phi(X)). \quad (596)$$

We here thus have

$$\begin{aligned} q : \mathbb{R} &\rightarrow [0, 1], \mu \mapsto q(\mu) := \mathbb{E}_{\mathbb{P}_\mu}(\phi(X)) \\ &= 1 \cdot \mathbb{P}_\mu(\phi(X) = 1) + 0 \cdot \mathbb{P}_\mu(\phi(X) = 0) \\ &= \mathbb{P}_\mu(\phi(X) = 1) \\ &= \mathbb{P}_\mu(T(X) \geq c_{\alpha'}) \\ &= \mathbb{P}_\mu(T(X) \geq \psi_{n-1}^{-1}(1 - \alpha')). \end{aligned} \quad (597)$$

We are thus concerned with the distribution of

$$T(X) := \sqrt{n} \frac{\bar{X}_n - \mu_0}{S_n} \quad (598)$$

as a function of μ . But this follows a non-central t distribution with non-centrality parameter

$$\delta := \sqrt{nd} \text{ for } d := \frac{\mu - \mu_0}{\sigma}, \quad (599)$$

because

Theoretical Exercise 1

$$\begin{aligned}
 T(X) &= \sqrt{n} \frac{\bar{X}_n - \mu_0}{S_n} \\
 &= \sqrt{n} \frac{\bar{X}_n - \mu + \mu - \mu_0}{S_n} \\
 &= \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)/\sigma}{S_n/\sigma}
 \end{aligned} \tag{600}$$

and we have seen previously, that

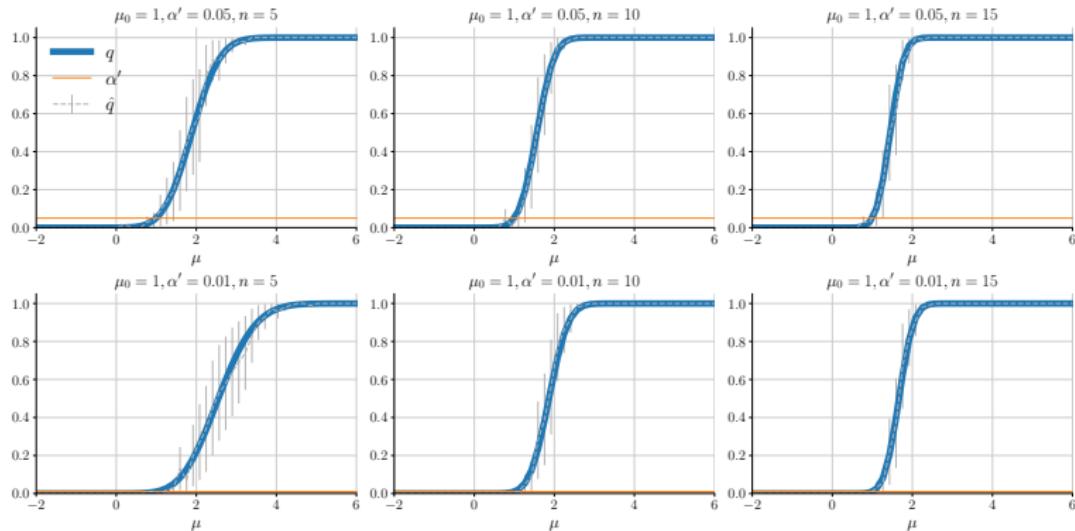
- $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0, 1)$,
- S_n/σ corresponds to the square root of a $\chi^2(n - 1)$ random variable divided by $n - 1$.

Thus, with $\delta := \sqrt{n}(\mu - \mu_0)/\sigma =: \sqrt{n}d$, $T(X)$ has CDF $\psi_{\delta, n-1}$. Hence,

$$\begin{aligned}
 q : \mathbb{R} \rightarrow [0, 1], \mu \mapsto q(\mu) &= \mathbb{P}_\mu(T(X) \geq \psi_{n-1}^{-1}(1 - \alpha')) \\
 &= 1 - \mathbb{P}_\mu(T(X) \leq \psi_{n-1}^{-1}(1 - \alpha')) \\
 &= 1 - \psi_{\delta, n-1}\left(\psi_{n-1}^{-1}(1 - \alpha')\right).
 \end{aligned} \tag{601}$$

Exercises

Programming Exercise 3



(13) Conjugate inference

Bibliographic remarks

Introductions to Bayesian inference can be found in most statistical textbooks (e.g., Wasserman, 2004; DeGroot and Schervish, 2012; Casella and Berger, 2012; Held and Sabanés Bové, 2014).

Conjugate inference

- Foundations
 - The Bayesian paradigm
 - Inference summaries
- Conjugate inference
 - The Beta-Binomial model
 - The univariate Gaussian-Gaussian model
- Exercises

Conjugate inference

- **Foundations**
 - **The Bayesian paradigm**
 - Inference summaries
- Conjugate inference
 - The Beta-Binomial model
 - The univariate Gaussian-Gaussian model
- Exercises

Central postulates of Frequentist inference

- Probabilities are interpreted as limiting relative frequencies and are considered objective properties of the real world.

Central postulates of Frequentist inference

- Probabilities are interpreted as limiting relative frequencies and are considered objective properties of the real world.
- Parameters are fixed, unknown constants, referred to as *true, but unknown*, values. No probability statements are made about parameters.

Central postulates of Frequentist inference

- Probabilities are interpreted as limiting relative frequencies and are considered objective properties of the real world.
- Parameters are fixed, unknown constants, referred to as *true, but unknown*, values. No probability statements are made about parameters.
- Statistical procedures are designed to have good long run frequency properties and are typically assessed by studying their sampling distributions.

Central postulates of Bayesian inference

- Probabilities are interpreted as degrees of belief, not limiting frequencies. Statements like “the probability that it will rain this afternoon is 0.5” are meaningful.

Central postulates of Bayesian inference

- Probabilities are interpreted as degrees of belief, not limiting frequencies. Statements like “the probability that it will rain this afternoon is 0.5” are meaningful.
- Parameters are fixed, unknown constants, about which probabilistic statements quantifying our uncertainty about their true, but unknown, value can be made.

Central postulates of Bayesian inference

- Probabilities are interpreted as degrees of belief, not limiting frequencies. Statements like “the probability that it will rain this afternoon is 0.5” are meaningful.
- Parameters are fixed, unknown constants, about which probabilistic statements quantifying our uncertainty about their true, but unknown, value can be made.
- Probabilistic statements about parameters are made with the help of probability distributions, from which further inferences, such as point or interval estimates, can be derived.

Definition (Probabilistic model)

A *probabilistic model* is a joint distribution over a family of observable random variables $X_{1:n} = (X_1, \dots, X_n)$, commonly modeling data, and a not directly observable random vector θ , commonly modeling parameters, that is specified in terms of a joint PDF or PMF

$$p(x_{1:n}, \theta). \quad (602)$$

Probabilistic models are also referred to as *generative models*.

Definition (Probabilistic model)

A *probabilistic model* is a joint distribution over a family of observable random variables $X_{1:n} = (X_1, \dots, X_n)$, commonly modeling data, and a not directly observable random vector θ , commonly modeling parameters, that is specified in terms of a joint PDF or PMF

$$p(x_{1:n}, \theta). \quad (602)$$

Probabilistic models are also referred to as *generative models*. Typically, probabilistic models are defined in terms of the product of a marginal PDF or PMF of θ and a conditional PDF or PMF of $x_{1:n}$ in the form

$$p(x_{1:n}, \theta) = p(x_{1:n}|\theta)p(\theta). \quad (603)$$

Here, $p(\theta)$ is referred to as *prior distribution* and $p(x_{1:n}|\theta)$ is referred to as the *likelihood*.

Definition (Probabilistic model)

A *probabilistic model* is a joint distribution over a family of observable random variables $X_{1:n} = (X_1, \dots, X_n)$, commonly modeling data, and a not directly observable random vector θ , commonly modeling parameters, that is specified in terms of a joint PDF or PMF

$$p(x_{1:n}, \theta). \quad (602)$$

Probabilistic models are also referred to as *generative models*. Typically, probabilistic models are defined in terms of the product of a marginal PDF or PMF of θ and a conditional PDF or PMF of $x_{1:n}$ in the form

$$p(x_{1:n}, \theta) = p(x_{1:n}|\theta)p(\theta). \quad (603)$$

Here, $p(\theta)$ is referred to as *prior distribution* and $p(x_{1:n}|\theta)$ is referred to as the *likelihood*. Often, the observable random variables $X_{1:n}$ are assumed to be *conditionally independent and identically distributed*, i.e.,

$$p(x_{1:n}|\theta) := \prod_{i=1}^n p(x_i|\theta) \text{ and } p(x_i|\theta) = p(x_1|\theta) \text{ for } i = 2, \dots, n. \quad (604)$$

Definition (Posterior distribution)

Given a probabilistic model $p(x_{1:n}, \theta)$, the conditional distribution of the not directly observable random vector θ given the observable random vector X is referred to as *posterior distribution*. By means of Bayes theorem for PMF or PDFs, the posterior distribution can be evaluated in terms of the PDF or PMF

$$p(\theta|x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{\int p(x_{1:n}|\theta)p(\theta) d\theta}. \quad (605)$$

Definition (Posterior distribution)

Given a probabilistic model $p(x_{1:n}, \theta)$, the conditional distribution of the not directly observable random vector θ given the observable random vector X is referred to as *posterior distribution*. By means of Bayes theorem for PMF or PDFs, the posterior distribution can be evaluated in terms of the PDF or PMF

$$p(\theta|x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{\int p(x_{1:n}|\theta)p(\theta) d\theta}. \quad (605)$$

The denominator of the right-hand side of the above is sometimes referred to as *evidence*, such that a mnemonic for the posterior distribution is given by

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \quad (606)$$

The Bayesian paradigm

Definition (Posterior distribution)

Given a probabilistic model $p(x_{1:n}, \theta)$, the conditional distribution of the not directly observable random vector θ given the observable random vector X is referred to as *posterior distribution*. By means of Bayes theorem for PMF or PDFs, the posterior distribution can be evaluated in terms of the PDF or PMF

$$p(\theta|x_{1:n}) = \frac{p(x_{1:n}|\theta)p(\theta)}{\int p(x_{1:n}|\theta)p(\theta) d\theta}. \quad (605)$$

The denominator of the right-hand side of the above is sometimes referred to as *evidence*, such that a mnemonic for the posterior distribution is given by

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \quad (606)$$

Because the evidence is independent of θ and constant for a given realization of X , it corresponds to a normalization constant of the product of prior and likelihood that renders this product a proper PDF or PMF. If only the functional form of the PDF or PMF of the posterior is of interest, the unnormalized version suffices,

$$p(\theta|x_{1:n}) \propto p(x_{1:n}|\theta)p(\theta). \quad (607)$$

Example (Batch Bayesian estimation)

Assume that n data points $x_i, i = 1, \dots, n$ are available. Then Bayesian estimation typically proceeds as follows:

- (1) Specification of a prior distribution $p(\theta)$.

Example (Batch Bayesian estimation)

Assume that n data points $x_i, i = 1, \dots, n$ are available. Then Bayesian estimation typically proceeds as follows:

- (1) Specification of a prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where often the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_{1:n}|\theta) = \prod_{i=1}^n p(x_i|\theta). \quad (608)$$

Example (Batch Bayesian estimation)

Assume that n data points $x_i, i = 1, \dots, n$ are available. Then Bayesian estimation typically proceeds as follows:

- (1) Specification of a prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where often the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_{1:n}|\theta) = \prod_{i=1}^n p(x_i|\theta). \quad (608)$$

- (3) *Batch Bayesian estimation* then amounts to evaluating

$$p(\theta|x_{1:n}) = \frac{p(\theta) \prod_{i=1}^n p(x_i|\theta)}{\int p(\theta) \prod_{i=1}^n p(x_i|\theta) d\theta}. \quad (609)$$

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_1|\theta) = p(x_2|\theta) = \dots \tag{610}$$

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_1|\theta) = p(x_2|\theta) = \dots \tag{610}$$

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_1|\theta) = p(x_2|\theta) = \dots \quad (610)$$

- (3) *Recursive Bayesian estimation* then amounts to recursively updating the distribution of θ , such that for each $i = 1, 2, \dots$ the posterior distribution at $i - 1$ serves as prior distribution at i , i.e.,

$$\begin{aligned} p(\theta|x_1) &= \frac{p(x_1|\theta)p(\theta)}{\int p(x_1|\theta)p(\theta) d\theta} \\ p(\theta|x_{1:2}) &= \frac{p(x_2|\theta)p(\theta|x_1)}{\int p(x_2|\theta)p(\theta|x_1) d\theta} \\ &\dots \end{aligned} \quad (611)$$

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_1|\theta) = p(x_2|\theta) = \dots \quad (610)$$

- (3) *Recursive Bayesian estimation* then amounts to recursively updating the distribution of θ , such that for each $i = 1, 2, \dots$ the posterior distribution at $i - 1$ serves as prior distribution at i , i.e.,

$$\begin{aligned} p(\theta|x_1) &= \frac{p(x_1|\theta)p(\theta)}{\int p(x_1|\theta)p(\theta) d\theta} \\ p(\theta|x_{1:2}) &= \frac{p(x_2|\theta)p(\theta|x_1)}{\int p(x_2|\theta)p(\theta|x_1) d\theta} \\ &\dots \end{aligned} \quad (611)$$

The Bayesian paradigm

Example (Recursive Bayesian estimation)

Assume that data points are available one at a time in the order x_1, x_2, \dots . Then Bayesian estimation can proceed as follows:

- (1) Specification of the prior distribution $p(\theta)$.
- (2) Specification of the data likelihood given the parameter, where the observed random variables points $X_i, i = 1, \dots, n$ are assumed to be conditionally independent given θ , such that

$$p(x_1|\theta) = p(x_2|\theta) = \dots \quad (610)$$

- (3) *Recursive Bayesian estimation* then amounts to recursively updating the distribution of θ , such that for each $i = 1, 2, \dots$ the posterior distribution at $i - 1$ serves as prior distribution at i , i.e.,

$$\begin{aligned} p(\theta|x_1) &= \frac{p(x_1|\theta)p(\theta)}{\int p(x_1|\theta)p(\theta) d\theta} \\ p(\theta|x_{1:2}) &= \frac{p(x_2|\theta)p(\theta|x_1)}{\int p(x_2|\theta)p(\theta|x_1) d\theta} \\ &\dots \end{aligned} \quad (611)$$

Note that for n data points, batch and recursive Bayesian estimation are equivalent:

$$p(\theta|x_{1:n}) = \frac{p(\theta) \prod_{i=1}^n p(x_i|\theta)}{\int p(\theta) \prod_{i=1}^n p(x_i|\theta) d\theta} = \frac{p(x_n|\theta)p(\theta|x_{1:n-1})}{\int p(x_n|\theta)p(\theta|x_{1:n-1}) d\theta}. \quad (612)$$

The Bayesian paradigm

Definition (Marginal likelihood, model evidence, Bayes factor)

Let $p(x_{1:n}, \theta)$ denote the PMF or PDF a probabilistic model. Then the marginal PDF or PMF of the observable random variables

$$p(x_{1:n}) = \int p(x_{1:n}, \theta) d\theta = \int p(x_{1:n}|\theta)p(\theta) d\theta \quad (613)$$

is referred to as *marginal (data) likelihood* or *model evidence*.

Definition (Marginal likelihood, model evidence, Bayes factor)

Let $p(x_{1:n}, \theta)$ denote the PMF or PDF a probabilistic model. Then the marginal PDF or PMF of the observable random variables

$$p(x_{1:n}) = \int p(x_{1:n}, \theta) d\theta = \int p(x_{1:n}|\theta)p(\theta) d\theta \quad (613)$$

is referred to as *marginal (data) likelihood* or *model evidence*. For a fixed set of data observations $x_{1:n}^*$ and two probabilistic models $p_1(x_{1:n}, \theta_1)$ and $p_2(x_{1:n}, \theta_2)$ with identical data outcome and possibly different parameter spaces, the ratio of the marginal likelihoods

$$\text{BF} = \frac{p_1(x_{1:n}^*)}{p_2(x_{1:n}^*)} \quad (614)$$

is referred to as *Bayes factor* and serves as a basic model comparison criterion in Bayesian statistics.

Remarks

- The evidence is “the probability of a data set under a probabilistic model”.
- $p(x_{1:n})$ is sometimes referred to as prior predictive distribution.

Definition (Posterior predictive distribution)

Let $p(x_{1:n}, \theta)$ denote the PDF or PMF of a probabilistic model. Then the conditional distribution of an observable random variable X_{n+1} given the observed random variables $X_{1:n}$ is referred to as *posterior predictive distribution*.

Definition (Posterior predictive distribution)

Let $p(x_{1:n}, \theta)$ denote the PDF or PMF of a probabilistic model. Then the conditional distribution of an observable random variable X_{n+1} given the observed random variables $X_{1:n}$ is referred to as *posterior predictive distribution*. A PMF or PDF of the posterior predictive distribution is given by

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\theta)p(\theta|x_{1:n}) d\theta, \quad (615)$$

where typically

$$p(x_{n+1}|\theta) = p(x_i|\theta) \text{ for } i = 1, \dots, n \quad (616)$$

Definition (Posterior predictive distribution)

Let $p(x_{1:n}, \theta)$ denote the PDF or PMF of a probabilistic model. Then the conditional distribution of an observable random variable X_{n+1} given the observed random variables $X_{1:n}$ is referred to as *posterior predictive distribution*. A PMF or PDF of the posterior predictive distribution is given by

$$p(x_{n+1}|x_{1:n}) = \int p(x_{n+1}|\theta)p(\theta|x_{1:n}) d\theta, \quad (615)$$

where typically

$$p(x_{n+1}|\theta) = p(x_i|\theta) \text{ for } i = 1, \dots, n \quad (616)$$

Remark

- In contrast to $p_{\hat{\theta}_n}(x_{n+1})$, $p(x_{n+1}|\theta)$ accounts for estimation uncertainty.

Conjugate inference

- **Foundations**

- The Bayesian paradigm
- **Inference summaries**

- Conjugate inference

- The Beta-Binomial model
- The univariate Gaussian-Gaussian model

- Exercises

Bayesian point estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, however, point estimates are often useful.
- Bayesian point estimation is a decision-theoretic problem.
- Decision theory rests on the notions of utility or loss functions.

Definition (Loss function, expected posterior loss, Bayes estimator)

Let $p(x_{1:n}, \theta)$ a probabilistic model and let $\hat{\theta}$ denote a point estimate for the true, but unknown, value of the parameter of the probabilistic model.

Definition (Loss function, expected posterior loss, Bayes estimator)

Let $p(x_{1:n}, \theta)$ a probabilistic model and let $\hat{\theta}$ denote a point estimate for the true, but unknown, value of the parameter of the probabilistic model. Then a real-valued function

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) \quad (617)$$

that measures the undesirability of choosing the point estimate $\hat{\theta}$, if the true, but unknown, parameter value is θ , is called a *loss function*.

Definition (Loss function, expected posterior loss, Bayes estimator)

Let $p(x_{1:n}, \theta)$ a probabilistic model and let $\hat{\theta}$ denote a point estimate for the true, but unknown, value of the parameter of the probabilistic model. Then a real-valued function

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) \quad (617)$$

that measures the undesirability of choosing the point estimate $\hat{\theta}$, if the true, but unknown, parameter value is θ , is called a *loss function*. The expected value of the loss function with respect to the posterior distribution of θ is referred to as the *expected posterior loss* and takes the form

$$l_p : \Theta \rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int l(\hat{\theta}, \theta) p(\theta | x_{1:n}) d\theta. \quad (618)$$

Definition (Loss function, expected posterior loss, Bayes estimator)

Let $p(x_{1:n}, \theta)$ a probabilistic model and let $\hat{\theta}$ denote a point estimate for the true, but unknown, value of the parameter of the probabilistic model. Then a real-valued function

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) \quad (617)$$

that measures the undesirability of choosing the point estimate $\hat{\theta}$, if the true, but unknown, parameter value is θ , is called a *loss function*. The expected value of the loss function with respect to the posterior distribution of θ is referred to as the *expected posterior loss* and takes the form

$$l_p : \Theta \rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int l(\hat{\theta}, \theta) p(\theta | x_{1:n}) d\theta. \quad (618)$$

The point estimator

$$\hat{\theta}_B := \arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) \quad (619)$$

that minimizes the expected posterior loss is known as *Bayes estimator*.

Example (Quadratic loss function, posterior expectation)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model with scalar parameter $\theta \in \Theta \subseteq \mathbb{R}$ and let the *quadratic loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := (\hat{\theta} - \theta)^2. \quad (620)$$

Example (Quadratic loss function, posterior expectation)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model with scalar parameter $\theta \in \Theta \subseteq \mathbb{R}$ and let the *quadratic loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := (\hat{\theta} - \theta)^2. \quad (620)$$

Then the Bayes estimator is the posterior expected value of the parameter, i.e.,

$$\hat{\theta}_B = \arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \int \theta p(\theta | x_{1:n}) d\theta. \quad (621)$$

Example (Quadratic loss function, posterior expectation)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model with scalar parameter $\theta \in \Theta \subseteq \mathbb{R}$ and let the *quadratic loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := (\hat{\theta} - \theta)^2. \quad (620)$$

Then the Bayes estimator is the posterior expected value of the parameter, i.e.,

$$\hat{\theta}_B = \arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \int \theta p(\theta | x_{1:n}) d\theta. \quad (621)$$

This estimator is also referred to as *minimum mean squared error (MMSE)* estimator.

Example (Quadratic loss function, posterior expectation)

Proof

The expected posterior loss function is given by

$$l_p : \Theta \rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int (\hat{\theta} - \theta)^2 p(\theta | x_{1:n}) d\theta \quad (622)$$

Example (Quadratic loss function, posterior expectation)

Proof

The expected posterior loss function is given by

$$l_p : \Theta \rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int (\hat{\theta} - \theta)^2 p(\theta|x_{1:n}) d\theta \quad (622)$$

Its derivative with respect to $\hat{\theta}$ evaluates to

$$\frac{d}{d\hat{\theta}} l_p(\hat{\theta}) = \int \frac{d}{d\hat{\theta}} (\hat{\theta} - \theta)^2 p(\theta|x_{1:n}) d\theta = 2 \int (\hat{\theta} - \theta) p(\theta|x_{1:n}) d\theta. \quad (623)$$

Example (Quadratic loss function, posterior expectation)

Proof

The expected posterior loss function is given by

$$l_p : \Theta \rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int (\hat{\theta} - \theta)^2 p(\theta|x_{1:n}) d\theta \quad (622)$$

Its derivative with respect to $\hat{\theta}$ evaluates to

$$\frac{d}{d\hat{\theta}} l_p(\hat{\theta}) = \int \frac{d}{d\hat{\theta}} (\hat{\theta} - \theta)^2 p(\theta|x_{1:n}) d\theta = 2 \int (\hat{\theta} - \theta) p(\theta|x_{1:n}) d\theta. \quad (623)$$

Setting to zero yields

$$\int (\hat{\theta} - \theta) p(\theta|x_{1:n}) d\theta = 0 \Leftrightarrow \hat{\theta} - \int \theta p(\theta|x_{1:n}) d\theta = 0 \Leftrightarrow \hat{\theta} = \int \theta p(\theta|x_{1:n}) d\theta. \quad (624)$$

□

Example (Zero-one loss function, posterior mode)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model and let the *zero-one loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := 1 - 1_{\hat{\theta}}(\theta) = \begin{cases} 0, & \theta = \hat{\theta} \\ 1, & \theta \neq \hat{\theta} \end{cases} \quad (625)$$

Example (Zero-one loss function, posterior mode)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model and let the *zero-one loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := 1 - 1_{\hat{\theta}}(\theta) = \begin{cases} 0, & \theta = \hat{\theta} \\ 1, & \theta \neq \hat{\theta} \end{cases} \quad (625)$$

Then the Bayes estimator is the posterior mode of the parameter, i.e.,

$$\hat{\theta}_B = \arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \arg \max_{\hat{\theta} \in \Theta} p(\hat{\theta} | x_{1:n}). \quad (626)$$

Example (Zero-one loss function, posterior mode)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model and let the *zero-one loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := 1 - 1_{\hat{\theta}}(\theta) = \begin{cases} 0, & \theta = \hat{\theta} \\ 1, & \theta \neq \hat{\theta} \end{cases} \quad (625)$$

Then the Bayes estimator is the posterior mode of the parameter, i.e.,

$$\hat{\theta}_B = \arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \arg \max_{\hat{\theta} \in \Theta} p(\hat{\theta} | x_{1:n}). \quad (626)$$

This estimator is also referred to as *maximum-a-posteriori (MAP)* estimator.

Example (Zero-one loss, posterior mode)

Proof

The expected posterior loss function is given by

$$\begin{aligned} l_p : \Theta &\rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int (1 - 1_{\hat{\theta}}(\theta)) p(\theta|x_{1:n}) d\theta \\ &= \int 1 p(\theta|x_{1:n}) d\theta - \int 1_{\hat{\theta}}(\theta) p(\theta|x_{1:n}) d\theta \\ &= 1 - p(\hat{\theta}|x_{1:n}) \end{aligned} \tag{627}$$

Example (Zero-one loss, posterior mode)

Proof

The expected posterior loss function is given by

$$\begin{aligned} l_p : \Theta &\rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int (1 - 1_{\hat{\theta}}(\theta)) p(\theta|x_{1:n}) d\theta \\ &= \int 1_{\hat{\theta}}(\theta) p(\theta|x_{1:n}) d\theta - \int 1_{\hat{\theta}}(\theta) p(\theta|x_{1:n}) d\theta \\ &= 1 - p(\hat{\theta}|x_{1:n}) \end{aligned} \tag{627}$$

Minimizing $l_p(\hat{\theta})$ with respect to $\hat{\theta}$ is thus equivalent to maximizing $p(\hat{\theta}|x_{1:n})$ with respect to $\hat{\theta}$.

□

Bayesian interval estimation

- The central entity of Bayesian inference is the posterior distribution.

Bayesian interval estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, interval estimates are often useful.

Bayesian interval estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, interval estimates are often useful.
- Bayesian interval estimates are referred to as *credible intervals*.

Bayesian interval estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, interval estimates are often useful.
- Bayesian interval estimates are referred to as *credible intervals*.
- Credible intervals are “more intuitive” than confidence intervals.

Bayesian interval estimation

- The central entity of Bayesian inference is the posterior distribution.
- In applied settings, interval estimates are often useful.
- Bayesian interval estimates are referred to as *credible intervals*.
- Credible intervals are “more intuitive” than confidence intervals.
- *Credible regions* generalize credible intervals for non-scalar parameters.

Definition (δ -credible region, δ -credible interval)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model with parameter space Θ and let $p(\theta|x_{1:n})$ denote the posterior distribution. Then any region $R_\delta \subset \Theta$ such that

$$\int_{R_\delta} p(\theta|x_{1:n}) d\theta = \delta \text{ for } \delta \in]0, 1[\quad (628)$$

is referred to as a posterior δ -credible region. If $R_\delta \subset \mathbb{R}$ is an interval, then R_δ is also referred to as a δ -credible interval.

Remarks

- δ -credible intervals are typically not uniquely defined

Remarks

- δ -credible intervals are typically not uniquely defined
- ... but neither are confidence intervals

Remarks

- δ -credible intervals are typically not uniquely defined
- ... but neither are confidence intervals
- Approaches for uniquely specifying δ -credible intervals include selecting

Remarks

- δ -credible intervals are typically not uniquely defined
- ... but neither are confidence intervals
- Approaches for uniquely specifying δ -credible intervals include selecting
 - highest posterior density (minimum size) δ -credible intervals,

Remarks

- δ -credible intervals are typically not uniquely defined
- ... but neither are confidence intervals
- Approaches for uniquely specifying δ -credible intervals include selecting
 - highest posterior density (minimum size) δ -credible intervals,
 - symmetric δ -credible intervals around the posterior expectation,

Remarks

- δ -credible intervals are typically not uniquely defined
- ... but neither are confidence intervals
- Approaches for uniquely specifying δ -credible intervals include selecting
 - highest posterior density (minimum size) δ -credible intervals,
 - symmetric δ -credible intervals around the posterior expectation,
 - equal upper and lower tail probability δ -credible intervals.

Conjugate inference

- Foundations
 - The Bayesian paradigm
 - Inference summaries
- **Conjugate inference**
 - The Beta-Binomial model
 - The univariate Gaussian-Gaussian model
- Exercises

Definition (Conjugate family of distributions and hyperparameters)

Let

$$p(\theta, x_{1:n}) = \prod_{i=1}^n p(x_i|\theta)p(\theta) \quad (629)$$

denote a probabilistic model with conditionally independent and identically distributed observed random variables X_1, \dots, X_n and unobserved random variable $\theta \in \Theta$.

Definition (Conjugate family of distributions and hyperparameters)

Let

$$p(\theta, x_{1:n}) = \prod_{i=1}^n p(x_i|\theta)p(\theta) \quad (629)$$

denote a probabilistic model with conditionally independent and identically distributed observed random variables X_1, \dots, X_n and unobserved random variable $\theta \in \Theta$.

Let ϕ denote a family of distributions over Θ . ϕ is called a *conjugate family of distributions with respect to $p(x_i|\theta)$* , if the posterior distribution $p(\theta|x_{1:n})$ is an element of ϕ , irrespective of the prior distribution $p(\theta) \in \phi$ and the observations x_1, \dots, x_n . If both the prior and the posterior distributions are elements of ϕ with respect to $p(x_i|\theta), i = 1, \dots, n$, ϕ is also said to be *closed under sampling*. Because θ is often referred to as parameter, the parameters of the distributions in ϕ are often referred to as *hyperparameters*.

Conjugate inference

- Foundations
 - The Bayesian paradigm
 - Inference summaries
- **Conjugate inference**
 - **The Beta-Binomial model**
 - The univariate Gaussian-Gaussian model
- Exercises

Theorem (The Beta-Binomial model)

Consider the probabilistic model

$$p(x, \theta) = p(x|\theta)p(\theta) := \text{Bin}(x; \theta, n)\text{Beta}(\theta; \alpha, \beta). \quad (630)$$

Theorem (The Beta-Binomial model)

Consider the probabilistic model

$$p(x, \theta) = p(x|\theta)p(\theta) := \text{Bin}(x; \theta, n)\text{Beta}(\theta; \alpha, \beta). \quad (630)$$

Then the posterior distribution is given by

$$p(\theta|x) = \text{Beta}(\theta; \alpha + x, \beta + n - x) \quad (631)$$

and the MMSE and MAP Bayes estimators are

$$\hat{\theta}_{\text{MMSE}} = \frac{\alpha + x}{\alpha + \beta + n} \text{ and } \hat{\theta}_{\text{MAP}} = \frac{\alpha + x - 1}{\alpha + \beta + n - 2}. \quad (632)$$

Example (Binomial random variable)

Let X be a random variable with outcome set $\mathcal{X} := \mathbb{N}_n^0$ and probability mass function

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \binom{n}{x} \mu^x (1 - \mu)^{n-x} \text{ for } \mu \in [0, 1]. \quad (633)$$

Then X is said to be distributed according to a *Binomial distribution* with parameters $\mu \in [0, 1]$ and $n \in \mathbb{N}$, for which we write $X \sim \text{Bin}(\mu, n)$. We denote the probability mass function of a Binomial random variable by

$$\text{Bin}(x; \mu, n) := \binom{n}{x} \mu^x (1 - \mu)^{n-x} \quad (634)$$

Remark

- $\text{Bin}(x; \mu, 1) = \text{Bern}(x; \mu)$.

Example (Beta random variable)

Let X be a random variable with outcome set $\mathcal{X} := [0, 1]$ and probability density function

$$p : \mathcal{X} \rightarrow [0, 1], x \mapsto p(x) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \text{ for } \alpha, \beta \in \mathbb{R}_{>0}, \quad (635)$$

where Γ denotes the Gamma function. Then X is said to be distributed according to a *Beta distribution* with parameters α, β , for which we write $X \sim \text{Beta}(\alpha, \beta)$. We denote the probability density function of a Beta random variable by

$$\text{Beta}(x; \alpha, \beta) := \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}. \quad (636)$$

Remarks

- A Beta random variable can be used to model the distribution of a probability.
- $\text{Beta}(x; 1, 1) = U(x; 0, 1)$.
- For $\alpha < 1, \beta < 1$ the outcome set is $\mathcal{X} :=]0, 1[$.

The Beta-Binomial model

Proof

Posterior distribution

We first note that up to proportionality constants, the posterior distribution is given by

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &= \text{Bin}(x; \theta, n)\text{Beta}(\theta; \alpha, \beta) \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \end{aligned} \tag{637}$$

The Beta-Binomial model

Proof

Posterior distribution

We first note that up to proportionality constants, the posterior distribution is given by

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &= \text{Bin}(x; \theta, n)\text{Beta}(\theta; \alpha, \beta) \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \end{aligned} \tag{637}$$

The right-hand side of the above corresponds to the kernel of the PDF of a Beta distribution with parameters $\alpha + x$ and $\beta + n - x$. The posterior distribution is thus given by

$$p(\theta|x) = \text{Beta}(\theta; \alpha + x, \beta + n - x) \tag{638}$$

The Beta-Binomial model

Proof

Posterior distribution

We first note that up to proportionality constants, the posterior distribution is given by

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &= \text{Bin}(x; \theta, n)\text{Beta}(\theta; \alpha, \beta) \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \end{aligned} \tag{637}$$

The right-hand side of the above corresponds to the kernel of the PDF of a Beta distribution with parameters $\alpha + x$ and $\beta + n - x$. The posterior distribution is thus given by

$$p(\theta|x) = \text{Beta}(\theta; \alpha + x, \beta + n - x) \tag{638}$$

MMSE and MAP Bayes estimators

The expected value and the mode of a random variable $X \sim \text{Beta}(\alpha, \beta)$ are given by

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta} \text{ and } \mathbb{M}(X) = \frac{\alpha - 1}{\alpha + \beta - 2} (\alpha, \beta > 1) \tag{639}$$

Substitution of the posterior distribution parameters thus directly yields the MMSE and MAP Bayes estimators.

The Beta-Binomial model

Remarks

- The MMSE estimator of the Beta-Binomial model can be written as

$$\begin{aligned}\hat{\theta}_{\text{MMSE}} &= \frac{\alpha + x}{\alpha + \beta + n} \\&= \frac{\alpha}{\alpha + \beta + n} + \frac{x}{\alpha + \beta + n} \\&= \frac{\alpha(\alpha + \beta)}{(\alpha + \beta + n)(\alpha + \beta)} + \frac{xn}{(\alpha + \beta + n)n} \\&= \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}\end{aligned}\tag{640}$$

The Beta-Binomial model

Remarks

- The MMSE estimator of the Beta-Binomial model can be written as

$$\begin{aligned}\hat{\theta}_{\text{MMSE}} &= \frac{\alpha + x}{\alpha + \beta + n} \\&= \frac{\alpha}{\alpha + \beta + n} + \frac{x}{\alpha + \beta + n} \\&= \frac{\alpha(\alpha + \beta)}{(\alpha + \beta + n)(\alpha + \beta)} + \frac{xn}{(\alpha + \beta + n)n} \\&= \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}\end{aligned}\tag{640}$$

- It is thus a weighted average of prior expectation and ML estimator.

The Beta-Binomial model

Remarks

- The MMSE estimator of the Beta-Binomial model can be written as

$$\begin{aligned}\hat{\theta}_{\text{MMSE}} &= \frac{\alpha + x}{\alpha + \beta + n} \\&= \frac{\alpha}{\alpha + \beta + n} + \frac{x}{\alpha + \beta + n} \\&= \frac{\alpha(\alpha + \beta)}{(\alpha + \beta + n)(\alpha + \beta)} + \frac{xn}{(\alpha + \beta + n)n} \\&= \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{x}{n}\end{aligned}\tag{640}$$

- It is thus a weighted average of prior expectation and ML estimator.
- The weighting constant are proportional to
 - the number of virtual prior observation $\alpha + \beta$,
 - the number of actual observations n .

Conjugate inference

- Foundations
 - The Bayesian paradigm
 - Inference summaries
- **Conjugate inference**
 - The Beta-Binomial model
 - **The univariate Gaussian-Gaussian model**
- Exercises

Theorem (The univariate Gaussian-Gaussian model)

Consider the probabilistic model

$$p(x_{1:n}, \theta) = \prod_{i=1}^n p(x_i | \theta) p(\theta) := \prod_{i=1}^n N(x_i; \theta, \sigma_x^2) N(\theta; \mu_\theta, \sigma_\theta^2). \quad (641)$$

The univariate Gaussian-Gaussian model

Theorem (The univariate Gaussian-Gaussian model)

Consider the probabilistic model

$$p(x_{1:n}, \theta) = \prod_{i=1}^n p(x_i | \theta) p(\theta) := \prod_{i=1}^n N(x_i; \theta, \sigma_x^2) N(\theta; \mu_\theta, \sigma_\theta^2). \quad (641)$$

Then the posterior distribution is given by

$$p(\theta | x_{1:n}) = N\left(\theta; \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right), \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1}\right) \quad (642)$$

and the MMSE and MAP Bayes estimators are

$$\hat{\theta}_{\text{MMSE}} = \hat{\theta}_{\text{MAP}} = \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right). \quad (643)$$

The univariate Gaussian-Gaussian model

Proof

Posterior distribution

By focusing on the fact that the posterior distribution is a function of θ and hence subsuming multiplicative constants independent of θ in proportionality statements, we have

$$\begin{aligned} p(\theta|x_{1:n}) &\propto \prod_{i=1}^n N(x_i; \theta, \sigma_x^2) N(\theta; \mu_\theta, \sigma_\theta^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{1}{2\sigma_x^2}(x_i - \theta)^2\right) \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \mu_\theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^n (x_i - \theta)^2}{\sigma_x^2} + \frac{(\theta - \mu_\theta)^2}{\sigma_\theta^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2)}{\sigma_x^2} + \frac{\theta^2 - 2\theta\mu_\theta + \mu_\theta^2}{\sigma_\theta^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\sum_{i=1}^n x_i^2 - 2n\bar{x}\theta + n\theta^2}{\sigma_x^2} + \frac{\theta^2 - 2\theta\mu_\theta + \mu_\theta^2}{\sigma_\theta^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{-2n\bar{x}\theta + n\theta^2}{\sigma_x^2} + \frac{\theta^2 - 2\theta\mu_\theta}{\sigma_\theta^2}\right)\right). \end{aligned} \tag{644}$$

The univariate Gaussian-Gaussian model

Proof (cont.)

Hence

$$\begin{aligned} p(\theta|x_{1:n}) &\propto \exp\left(-\frac{1}{2}\left(\frac{\theta^2 n - 2n\bar{x}\theta}{\sigma_x^2} + \frac{\theta^2 - 2\theta\mu_\theta}{\sigma_\theta^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\theta^2 \frac{n}{\sigma_x^2} - 2\theta \frac{n\bar{x}}{\sigma_x^2} + \theta^2 \frac{1}{\sigma_\theta^2} - 2\theta \frac{\mu_\theta}{\sigma_\theta^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\theta^2 \frac{n}{\sigma_x^2} - \frac{1}{2}\theta^2 \frac{1}{\sigma_\theta^2} + \theta \frac{n\bar{x}}{\sigma_x^2} + \theta \frac{\mu_\theta}{\sigma_\theta^2}\right) \\ &= \exp\left(-\frac{1}{2}\theta^2 \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right) + \theta \left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right)\right). \end{aligned} \tag{645}$$

The univariate Gaussian-Gaussian model

Proof (cont.)

Hence

$$\begin{aligned}
 p(\theta|x_{1:n}) &\propto \exp\left(-\frac{1}{2}\left(\frac{\theta^2 n - 2n\bar{x}\theta}{\sigma_x^2} + \frac{\theta^2 - 2\theta\mu_\theta}{\sigma_\theta^2}\right)\right) \\
 &= \exp\left(-\frac{1}{2}\left(\theta^2 \frac{n}{\sigma_x^2} - 2\theta \frac{n\bar{x}}{\sigma_x^2} + \theta^2 \frac{1}{\sigma_\theta^2} - 2\theta \frac{\mu_\theta}{\sigma_\theta^2}\right)\right) \\
 &= \exp\left(-\frac{1}{2}\theta^2 \frac{n}{\sigma_x^2} - \frac{1}{2}\theta^2 \frac{1}{\sigma_\theta^2} + \theta \frac{n\bar{x}}{\sigma_x^2} + \theta \frac{\mu_\theta}{\sigma_\theta^2}\right) \\
 &= \exp\left(-\frac{1}{2}\theta^2 \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right) + \theta \left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right)\right).
 \end{aligned} \tag{645}$$

By defining

$$\phi_1 := \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2}\right)^{-1} \quad \text{and} \quad \phi_2 := \phi_1 \left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2}\right) \tag{646}$$

we then have

$$\begin{aligned}
 p(\theta|x_{1:n}) &\propto \exp\left(-\frac{1}{2\phi_1}\theta^2 + \frac{1}{\phi_1}\theta\phi_2\right) \\
 &\propto \exp\left(-\frac{1}{2\phi_1}\theta^2 + \frac{1}{\phi_1}\theta\phi_2 - \frac{1}{2\phi_1}\phi_2^2\right).
 \end{aligned} \tag{647}$$

The univariate Gaussian-Gaussian model

Proof (cont.)

Hence

$$p(\theta|x_{1:n}) \propto \exp\left(-\frac{1}{2\phi_1}(\theta - \phi_2)^2\right) \quad (648)$$

The univariate Gaussian-Gaussian model

Proof (cont.)

Hence

$$p(\theta|x_{1:n}) \propto \exp\left(-\frac{1}{2\phi_1}(\theta - \phi_2)^2\right) \quad (648)$$

Based on the normalization constant of the Gaussian probability density function, we thus have

$$\begin{aligned} p(\theta|x_{1:n}) &= \frac{1}{\sqrt{2\pi\phi_1}} \exp\left(-\frac{1}{2\phi_1}(\theta - \phi_2)^2\right) \\ &= N(\theta; \phi_2, \phi_1), \end{aligned} \quad (649)$$

such that the posterior distribution is given by a univariate Gaussian probability density function with expectation parameter

$$\phi_2 = \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2} \right)^{-1} \left(\frac{\mu_\theta}{\sigma_\theta^2} + \frac{n\bar{x}}{\sigma_x^2} \right) \quad (650)$$

and variance parameter

$$\phi_1 = \left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2} \right)^{-1}. \quad (651)$$

□

The univariate Gaussian-Gaussian model

Remarks

- The MMSE estimator of the univariate Gaussian-Gaussian model has the form

$$\hat{\theta}_{\text{MMSE}} \propto \frac{1}{\sigma_\theta^2} \mu_\theta + \frac{n}{\sigma_x^2} \bar{x}. \quad (652)$$

The univariate Gaussian-Gaussian model

Remarks

- The MMSE estimator of the univariate Gaussian-Gaussian model has the form

$$\hat{\theta}_{\text{MMSE}} \propto \frac{1}{\sigma_\theta^2} \mu_\theta + \frac{n}{\sigma_x^2} \bar{x}. \quad (652)$$

- It is thus a weighted average of the prior expectation and the ML estimator.

The univariate Gaussian-Gaussian model

Remarks

- The MMSE estimator of the univariate Gaussian-Gaussian model has the form

$$\hat{\theta}_{\text{MMSE}} \propto \frac{1}{\sigma_\theta^2} \mu_\theta + \frac{n}{\sigma_x^2} \bar{x}. \quad (652)$$

- It is thus a weighted average of the prior expectation and the ML estimator.
- The weighting constants are given by
 - the prior precision (reciprocal variance) $1/\sigma_\theta^2$,
 - the data precision $1/\sigma_x^2$ and the number of observations.

The univariate Gaussian-Gaussian model

Remarks

- The MMSE estimator of the univariate Gaussian-Gaussian model has the form

$$\hat{\theta}_{\text{MMSE}} \propto \frac{1}{\sigma_\theta^2} \mu_\theta + \frac{n}{\sigma_x^2} \bar{x}. \quad (652)$$

- It is thus a weighted average of the prior expectation and the ML estimator.
- The weighting constants are given by
 - the prior precision (reciprocal variance) $1/\sigma_\theta^2$,
 - the data precision $1/\sigma_x^2$ and the number of observations.
- The posterior variance parameter

$$\left(\frac{n}{\sigma_x^2} + \frac{1}{\sigma_\theta^2} \right)^{-1} \quad (653)$$

is reciprocally related to the number observations.

Further often encountered conjugate models include

- the Dirichlet-Multinoulli model,
- the Gaussian-Gamma-Gaussian model,
- the multivariate Gaussian-Gaussian model,
- the Gaussian-Wishart-Gaussian model.

Further often encountered conjugate models include

- the Dirichlet-Multinoulli model,
- the Gaussian-Gamma-Gaussian model,
- the multivariate Gaussian-Gaussian model,
- the Gaussian-Wishart-Gaussian model.

A comprehensive theory for conjugate analysis is afforded by

- Conjugate inference in the exponential family.

Conjugate inference

- Foundations
 - The Bayesian paradigm
 - Inference summaries
- Conjugate inference
 - The Beta-Binomial model
 - The univariate Gaussian-Gaussian model
- Exercises

Study questions

1. Write down the definition of a probabilistic model, a generative model, a prior distribution, a likelihood, and a posterior distribution.
2. Write down the distribution of n conditionally independent and identically distributed random variables $X_i, i = 1, \dots, n$ given a parameter random variable θ .
3. Describe the differences and similarities between batch and recursive Bayesian estimation.
4. Write down the definition of the marginal data likelihood (model evidence).
5. Given two probabilistic models and a set of data observations, write down the Bayes factor.
6. Write down the definition of the posterior predictive distribution.
7. Write down the definition of a loss function, the expected posterior loss, and a Bayes estimator.
8. Write down the Bayes estimator under a quadratic loss function.
9. Write down the Bayes estimator under zero-one loss function.
10. Write down the definition of a conjugate family of distributions.

Theoretical exercises

1. Derive the posterior distribution as well as the MMSE and MAP estimators for the Beta-Binomial model.
2. Derive the posterior distribution as well as the MMSE and MAP estimators for the univariate Gaussian-Gaussian model.
3. Show that the Bayes estimator under an absolute error loss function is given by the median (DeGroot and Schervish, 2012, Corollary 7.4.2, Theorem 4.5.3).

Theoretical Exercise 3 (Background)

Definition (Median of a continuous random variable)

Let X denote a continuous random variable with PDF p . Then $m \in \mathbb{R}$ is called *median* of X , if

$$\mathbb{P}(X \leq m) = \int_{-\infty}^m p(x) dx = \frac{1}{2} = \int_m^{\infty} p(x) dx = \mathbb{P}(X \geq m) \quad (654)$$

Theorem (Median inequality)

Let X be a continuous random variable with PDF p and finite expectation and let m denote a median of X . Then for any d , it holds that

$$\mathbb{E}(|X - m|) \leq \mathbb{E}(|X - d|) \quad (655)$$

with equality, if d is also a median of X .

Exercises

Theoretical Exercise 3 (Background)

Proof

We first consider the case $d > m$. Then

$$\mathbb{E}(|X - m|) \leq \mathbb{E}(|X - d|) \Leftrightarrow \mathbb{E}(|X - d|) - \mathbb{E}(|X - m|) \geq 0, \quad (656)$$

because

$$\begin{aligned} & \mathbb{E}(|X - d|) - \mathbb{E}(|X - m|) \\ &= \int_{-\infty}^{\infty} (|x - d| - |x - m|) p(x) dx \\ &= \int_{-\infty}^m (d - m) p(x) dx + \int_m^d (d + m - 2x) p(x) dx + \int_d^{\infty} (m - d) p(x) dx \\ &\geq \int_{-\infty}^m (d - m) p(x) dx + \int_m^d (m - d) p(x) dx + \int_d^{\infty} (m - d) p(x) dx \end{aligned} \quad (657)$$

as can be seen from

- inspection of the drawing below and
- noting that for $d > m$ the smallest value that

$$f : [m, d] \rightarrow \mathbb{R}, x \mapsto f(x) := d + m - 2x \quad (658)$$

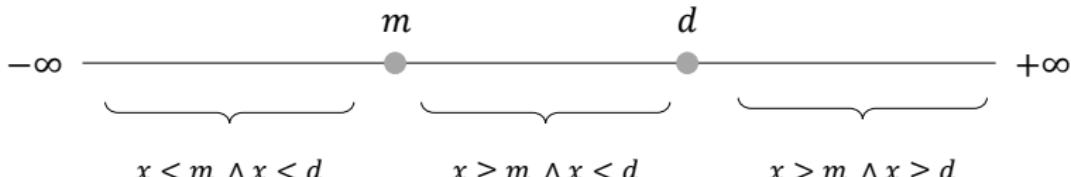
takes on is given by

$$f(d) = d + m - 2d = m - d \quad (659)$$

Exercises

Theoretical Exercise 3 (Background)

Proof



Then

$$|x - d| = -(x - d) = d - x$$

$$|x - m| = -(x - m) = m - x$$

Hence

$$\int_{-\infty}^m (|x - d| - |x - m|)p(x) dx$$

$$= \int_{-\infty}^m (d - x - (m - x))p(x) dx$$

$$= \int_{-\infty}^m (d - m)p(x) dx$$

Then

$$|x - d| = -(x - d) = d - x$$

$$|x - m| = x - m$$

Hence

$$\int_m^d (|x - d| - |x - m|)p(x) dx$$

$$= \int_m^d (d - x - (x - m))p(x) dx$$

$$= \int_m^d (d + m - 2x)p(x) dx$$

Then

$$|x - d| = x - d$$

$$|x - m| = x - m$$

Hence

$$\int_d^{\infty} (|x - d| - |x - m|)p(x) dx$$

$$= \int_d^{\infty} (x - d - (x - m))p(x) dx$$

$$= \int_d^{\infty} (m - d)p(x) dx$$

Theoretical Exercise 3 (Background)

Proof

We thus have

$$\begin{aligned} & \mathbb{E}(|X - d|) - \mathbb{E}(|X - m|) \\ & \geq \int_{-\infty}^m (d - m)p(x) dx + \int_m^d (m - d)p(x) dx + \int_d^{\infty} (m - d)p(x) dx \\ & = (d - m) \int_{-\infty}^m p(x) dx + (m - d) \int_m^d p(x) dx + (m - d) \int_d^{\infty} p(x) dx \\ & = (d - m) \int_{-\infty}^m p(x) dx + (m - d) \left(\int_m^d p(x) dx + \int_d^{\infty} p(x) dx \right) \quad (660) \\ & = (d - m) \int_{-\infty}^m p(x) dx + (m - d) \int_m^{\infty} p(x) dx \\ & = (d - m)\mathbb{P}(X \leq m) + (m - d)\mathbb{P}(X > m) \\ & = d\mathbb{P}(X \leq m) - m\mathbb{P}(X \geq d) + m\mathbb{P}(X > m) - d\mathbb{P}(X > m) \\ & = (d - m)(\mathbb{P}(X \leq m) - \mathbb{P}(X > m)) \end{aligned}$$

Theoretical Exercise 3 (Background)

Proof

Thus

$$\mathbb{E}(|X - d|) - \mathbb{E}(|X - m|) \geq (d - m) (\mathbb{P}(X \leq m) - \mathbb{P}(X > m)). \quad (661)$$

Because m is a median, it follows that

$$\mathbb{P}(X \leq m) = \frac{1}{2} \geq \mathbb{P}(X > m) \quad (662)$$

and thus

$$(d - m) > 0 \text{ and } (\mathbb{P}(X \leq m) - \mathbb{P}(X > m)) \geq 0 \quad (663)$$

Hence, in turn,

$$\mathbb{E}(|X - d|) - \mathbb{E}(|X - m|) \geq 0 \Leftrightarrow \mathbb{E}(|X - m|) \leq \mathbb{E}(|X - d|) \quad (664)$$

A similar line of argument can be developed for $d < m$. Finally, assuming a unique median, for $d = m$, the inequality in the above corresponds to an identity.

Theoretical Exercise 3

Example (Absolute error loss function, median)

Let $p(x_{1:n}, \theta)$ denote a probabilistic model and let the *absolute error loss function* be defined as

$$l : \Theta \times \Theta \rightarrow \mathbb{R}, (\hat{\theta}, \theta) \mapsto l(\hat{\theta}, \theta) := |\hat{\theta} - \theta| \quad (665)$$

Then the Bayes estimator is the median of the posterior distribution, i.e.,

$$\hat{\theta}_B = \arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = \left\{ \hat{\theta} \mid \int_{-\infty}^{\hat{\theta}} p(\theta|x_{1:n}) = \frac{1}{2} = \int_{\hat{\theta}}^{\infty} p(\theta|x_{1:n}) \right\} \quad (666)$$

Theoretical Exercise 3

Example (Absolute error loss function, median)

Proof

The expected posterior loss function is given by

$$\begin{aligned} l_p : \Theta &\rightarrow \mathbb{R}, \hat{\theta} \mapsto l_p(\hat{\theta}) := \int |\hat{\theta} - \theta| p(\theta | x_{1:n}) d\theta \\ &= \int |\theta - \hat{\theta}| p(\theta | x_{1:n}) d\theta \\ &= \mathbb{E}_{p(\theta | x_{1:n})}(|\theta - \hat{\theta}|) \end{aligned} \tag{667}$$

Let m denote a median of θ . The median inequality theorem states that

$$\mathbb{E}_{p(\theta | x_{1:n})}(|\theta - m|) \leq \mathbb{E}_{p(\theta | x_{1:n})}(|\theta - \hat{\theta}|) \tag{668}$$

with equality only for $\hat{\theta} = m$. Thus m minimizes the posterior loss function, i.e.,

$$\arg \min_{\hat{\theta} \in \Theta} l_p(\hat{\theta}) = m. \tag{669}$$

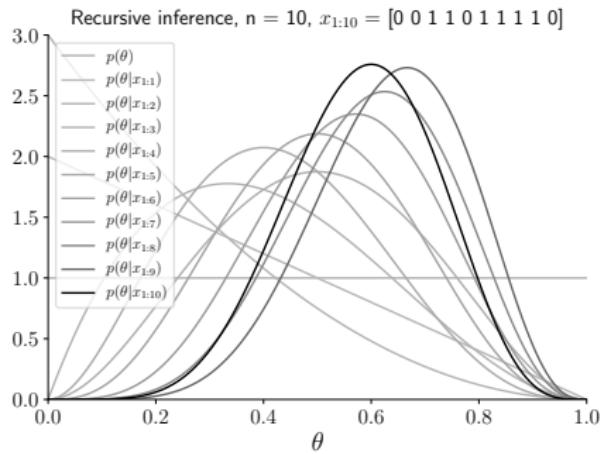
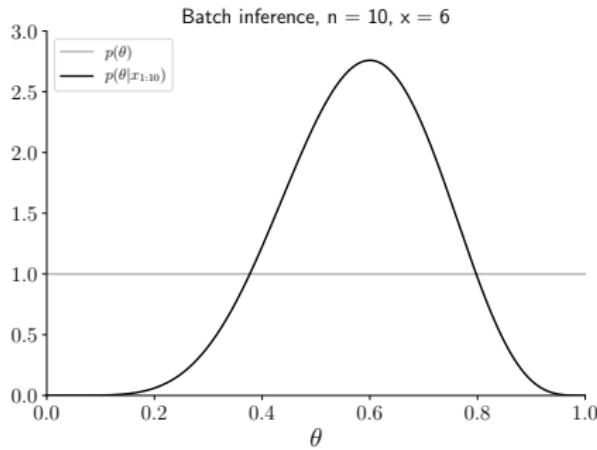
□

Programming exercises

1. For $n = 10$, implement batch and recursive Bayesian estimation for the Beta-Binomial model. Compare the results based on identical samples.
2. Using simulation, study the bias and consistency properties of the posterior expected value of the Beta-Binomial model.
3. Using simulation, study the bias and consistency properties of the posterior expected value of the Gaussian-Gaussian model.

Exercises

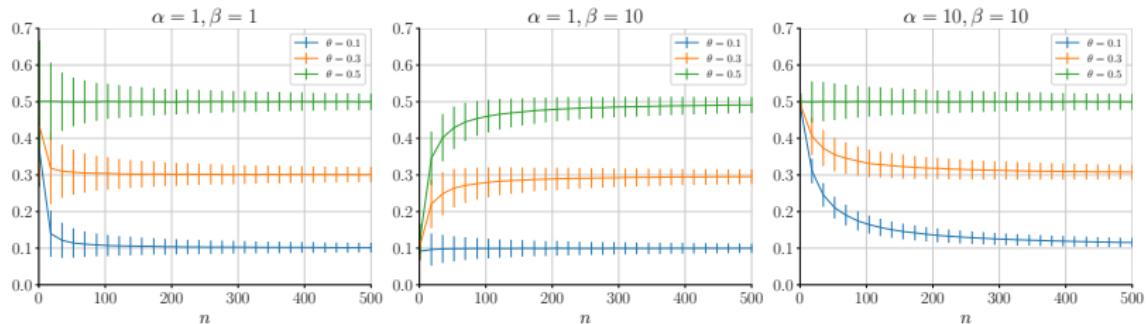
Programming Exercise 1



Exercises

Programming Exercise 2

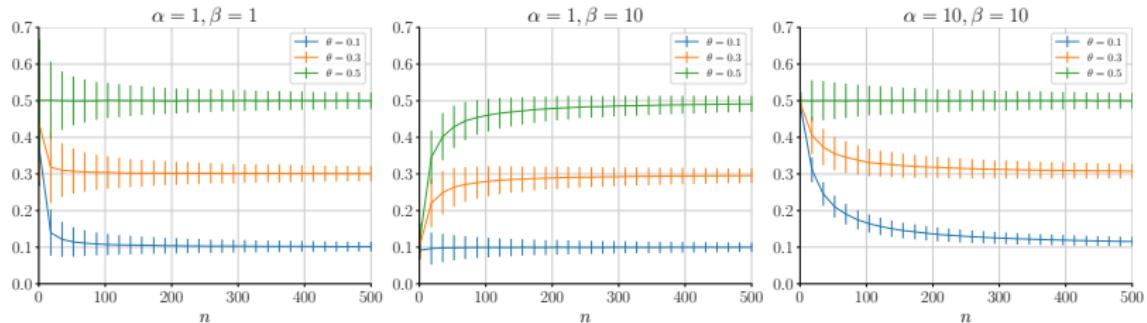
Bias and consistency of the Beta-Binomial posterior expectation



Exercises

Programming Exercise 3

Bias and consistency of the Gaussian-Gaussian posterior expectation



(14) Numerical methods

Bibliographic remarks

The material presented in this section follows Wasserman (2004) and Held and Sabanés Bové (2014). For an introduction to Monte Carlo methods in machine learning, see Andrieu (2003).

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling
- Exercises

Numerical methods

- **Motivation**
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling
- Exercises

Motivation

- The posterior distribution is central to Bayesian inference.
- The prior and likelihood are modeling choices, but the normalization factor

$$p(x_{1:n}) = \int p(x_{1:n}|\theta)p(\theta) d\theta \quad (670)$$

has to be evaluated.

- In conjugate models, the evaluation of the posterior is analytically tractable.
- In non-conjugate models, this may not be see case.

For example, $p(\theta) = N(\theta; \mu, \sigma^2)$ and $p(x_{1:n}|\theta) = \prod_{i=1}^n C(x_i; \theta, 1)$ yields

$$p(\theta|x_{1:n}) \propto \exp\left(\frac{-1}{2\sigma^2}(\theta - \mu)^2\right) \prod_{i=1}^n (1 + (x_i - \theta)^2)^{-1}. \quad (671)$$

The right-hand side cannot be integrated analytically.

- Even if the posterior distribution is analytically tractable, evaluating Bayesian estimators

$$\hat{\theta}_B = \mathbb{E}_{p(\theta|x_{1:n})}(f(\theta)) = \int f(\theta)p(\theta|x_{1:n}) d\theta \quad (672)$$

may not be analytically possible.

⇒ Bayesian inference often requires methods for numerical integration.

Here, we survey the following numerical integration methods

- Quadrature approaches as classical means for numerical integration,
- the Laplace approximation as analytical integral approximation method, and
- Monte Carlo integration, i.e. numerical integration by sampling and estimation.

In addition, we consider

- Importance sampling
- Acceptance-rejection sampling

as specialized sampling schemes for Monte Carlo integration. For brevity, we omit

- discussing the pros and cons of different methods,
- considering Markov chain based sampling methods (MCMC).

Numerical methods

- Motivation
- **Quadrature**
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling
- Exercises

Quadrature

- ... is a synonym for deterministic numerical integration.
- ... is a topic in numerical mathematics.
- ... typically works well for low-dimensional integration problems.
- ... comprises many different methods, such as
 - Riemann sums
 - Trapezoidal rule
 - Simpson's rule
 - Newton-Cotes formulas

We briefly review Riemann sums as an example.

Quadrature

Definition (Riemann sum and Riemann integral)

Let $f : [a, b] \rightarrow \mathbb{R}$ be a univariate real-valued function on $[a, b] \subset \mathbb{R}$.

$$P_n := \{[x_0, x_1], [x_1, x_2], \dots, [x_{n-1}, x_n]\} \quad (673)$$

with

$$a =: x_0 < x_1 < x_2 < \dots < x_n =: b \quad (674)$$

be a partition of the interval $[a, b]$. Then the *Riemann sum* S_n of f over the interval $[a, b]$ based on the partition P_n is defined as

$$S_n := \sum_{i=1}^n f(x_i^*) \Delta x_i \text{ with } \Delta x_i := x_i - x_{i-1} \text{ and } x_i^* \in [x_{i-1}, x_i]. \quad (675)$$

The *definite Riemann integral* of f on $[a, b]$ is defined as

$$\int_a^b f(x) dx = \lim_{\Delta x \rightarrow 0} \sum_{i=1}^n f(x_i^*) \Delta x_i, \quad (676)$$

if the limit on the right-hand side exists.

Remark

- A basic quadrature idea is to approximate integrals by Riemann sums.

Definition (Left rule, right rule, midpoint rule)

Let $f : [a, b] \rightarrow \mathbb{R}$ and assume the aim is to approximate the integral

$$I = \int_a^b f(x) dx \tag{677}$$

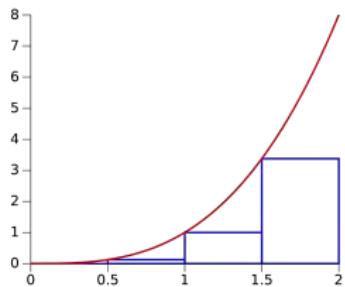
using a Riemann sum S_n . Then setting

- $x_i^* := x_{i-1}$ is called the *left rule*,
- $x_i^* := x_i$ is called the *right rule*, and
- $x_i^* := \frac{1}{2}(x_i + x_{i-1})$ is called the *midpoint rule*.

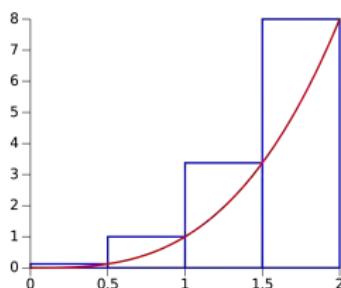
Remarks

- As $\Delta x \rightarrow 0$, the choice of x_i^* does not matter for the Riemann integral.
- Often equipartitions of the form $\Delta x = (b - a)/n$ are used.
- f is approximated by piecewise constant functions.

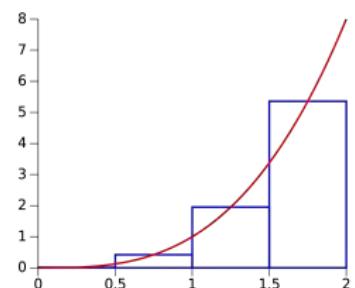
Left rule



Right rule



Midpoint rule



https://en.wikipedia.org/wiki/Riemann_sum

Numerical methods

- Motivation
- Quadrature
- **Laplace approximation**
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling
- Exercises

The Laplace approximation

- ... is a method to approximate posterior expectations of the type

$$\mathbb{E}_{p(\theta|x_{1:n})}(f(\theta)) = \int f(\theta)p(\theta|x_{1:n}) d\theta, \quad (678)$$

such as the posterior expected value

$$\mathbb{E}_{p(\theta|x_{1:n})} = \int \theta p(\theta|x_{1:n}) d\theta. \quad (679)$$

- ... does not require the functional form of the posterior distribution.
- ... is an application of *Laplace's integral approximation method*.

Laplace approximation

Definition (Laplace's integral approximation method)

For a univariate, real-valued, convex, and twice differentiable function f with a minimum at \tilde{x} , a reasonable approximation of the integral

$$I_n = \int_{-\infty}^{\infty} \exp(-nf(x)) dx \quad (680)$$

is given by

$$I_n \approx \exp(-nf(\tilde{x})) \sqrt{\frac{2\pi}{n\kappa}}, \quad (681)$$

where

$$\kappa := f''(\tilde{x}) > 0 \quad (682)$$

denotes the second derivative of f at its minimum location \tilde{x} .

Remarks

- The minimum location \tilde{x} and $\kappa = f''(\tilde{x})$ have to be evaluated.
- The approximation rests on a second-order Taylor approximation of f in \tilde{x} and

$$\int_{-\infty}^{\infty} \exp(-a(x-b)^2) dx = \sqrt{\pi/a}. \quad (683)$$

- “Reasonable” means that the approximation error decreases for $n \rightarrow \infty$.

Laplace approximation

Motivation of Laplace's integral approximation method

Consider the second-order Taylor approximation of f in \tilde{x} ,

$$f(x) \approx f(\tilde{x}) + f'(\tilde{x})(x - \tilde{x}) + \frac{1}{2} f''(\tilde{x})(x - \tilde{x})^2. \quad (684)$$

Because \tilde{x} is a minimum of f , it holds that $f'(\tilde{x}) = 0$ and $\kappa := f''(\tilde{x}) > 0$. We then have

$$\begin{aligned} I_n &= \int_{-\infty}^{\infty} \exp(-nf(x)) dx \\ &\approx \int_{-\infty}^{\infty} \exp\left(-n\left(f(\tilde{x}) + \frac{1}{2}\kappa(x - \tilde{x})^2\right)\right) dx \\ &= \int_{-\infty}^{\infty} \exp\left(-nf(\tilde{x}) - \frac{n}{2}\kappa(x - \tilde{x})^2\right) dx \\ &= \int_{-\infty}^{\infty} \exp(-nf(\tilde{x})) \exp\left(-\frac{n}{2}\kappa(x - \tilde{x})^2\right) dx \\ &= \exp(-nf(\tilde{x})) \int_{-\infty}^{\infty} \exp\left(-\frac{n\kappa}{2}(x - \tilde{x})^2\right) dx. \end{aligned} \quad (685)$$

With

$$\int_{-\infty}^{\infty} \exp\left(-a(x - b)^2\right) dx = \sqrt{\pi/a}, \quad (686)$$

it then follows that

$$I_n = \exp(-nf(\tilde{x})) \sqrt{\frac{2\pi}{n\kappa}}. \quad (687)$$

□

Laplace approximation

Definition (Laplace approximation)

Let

$$p(\theta, x_{1:n}) = p(x_{1:n}|\theta)p(\theta) \quad (688)$$

be a probabilistic model with scalar parameter θ and $x_{1:n} = (x_1, \dots, x_n)$ denoting a value of a random sample $X = (X_1, \dots, X_n)$ with $X_i \sim p(x_i|\theta)$ for $i = 1, \dots, n$. Consider the problem of evaluating a posterior distribution expectation of the form

$$\mathbb{E}_{p(\theta|x_{1:n})}(f(\theta)) \text{ for } f : \mathbb{R} \rightarrow \mathbb{R}, \theta \mapsto f(\theta). \quad (689)$$

A reasonable approximation of such a feature is given by

$$\mathbb{E}_{p(\theta|x_{1:n})}(f(\theta)) \approx \sqrt{\frac{\kappa_1}{\kappa_2}} \exp\left(-n\left(h_2(\tilde{\theta}_2) - h_1(\tilde{\theta}_1)\right)\right), \quad (690)$$

where

$$\begin{aligned} h_1 : \mathbb{R} &\rightarrow \mathbb{R}, \theta \mapsto h_1(\theta) := -\ln f(\theta) - \ln p(x_i|\theta) - \ln p(\theta) \\ h_2 : \mathbb{R} &\rightarrow \mathbb{R}, \theta \mapsto h_2(\theta) := -\ln p(x_i|\theta) - \ln p(\theta), \end{aligned} \quad (691)$$

$\tilde{\theta}_1$ and $\tilde{\theta}_2$ are minimum points of h_1 and h_2 , respectively, and

$$\kappa_1 := h_1''(\tilde{\theta}_1) > 0 \text{ and } \kappa_2 := h_2''(\tilde{\theta}_2) > 0. \quad (692)$$

Remarks

- “Reasonable” means that the approximation error decreases for $n \rightarrow \infty$.
- If $\kappa_1, \kappa_2, \tilde{\theta}_1, \tilde{\theta}_2$ are not available analytically, they are evaluated numerically.

Laplace approximation

Motivation of the Laplace approximation

We first note that with the definitions of h_1 and h_2 , we have

$$\begin{aligned}\mathbb{E}_{p(\theta|x_{1:n})}(f(\theta)) &= \int f(\theta)p(\theta|x_{1:n}) d\theta \\&= \int f(\theta) \frac{p(x_{1:n}|\theta)p(\theta)}{\int p(x_{1:n}|\theta)p(\theta) d\theta} d\theta \\&= \frac{\int f(\theta)p(x_{1:n}|\theta)p(\theta) d\theta}{\int p(x_{1:n}|\theta)p(\theta) d\theta} \\&= \frac{\int \exp(\ln(f(\theta)p(x_{1:n}|\theta)p(\theta))) d\theta}{\int \exp(\ln(p(x_{1:n}|\theta)p(\theta))) d\theta} \\&= \frac{\int \exp\left(\ln\left(f(\theta)\prod_{i=1}^n p(x_i|\theta)p(\theta)\right)\right) d\theta}{\int \exp\left(\ln\left(\prod_{i=1}^n p(x_i|\theta)p(\theta)\right)\right) d\theta} \\&= \frac{\int \exp\left(\ln\left(\prod_{i=1}^n f(\theta)p(x_i|\theta)p(\theta)\right)\right) d\theta}{\int \exp\left(\ln\left(\prod_{i=1}^n p(x_i|\theta)p(\theta)\right)\right) d\theta} \\&= \frac{\int \exp(n \ln f(\theta) + n \ln p(x_i|\theta) + n \ln p(\theta)) d\theta}{\int \exp(n \ln p(x_i|\theta) + n \ln p(\theta)) d\theta} \\&= \frac{\int \exp(-nh_2(\theta)) d\theta}{\int \exp(-nh_1(\theta)) d\theta}.\end{aligned}\tag{693}$$

Laplace approximation

Motivation of the Laplace approximation (cont.)

Application of Laplace's integral approximation method to the numerator and denominator of

$$\mathbb{E}_{p(\theta|x_{1:n})}(\theta) = \frac{\int \exp(-nh_2(\theta)) d\theta}{\int \exp(-nh_1(\theta)) d\theta} \quad (694)$$

then yields

$$\begin{aligned}\mathbb{E}_{p(\theta|x_{1:n})}(\theta) &\approx \frac{\exp(-nh_1(\tilde{\theta}_1)) \sqrt{\frac{2\pi}{n\kappa_1}}}{\exp(-nh_2(\tilde{\theta}_2)) \sqrt{\frac{2\pi}{n\kappa_2}}} \\ &= \sqrt{\frac{2\pi}{n\kappa_1} \frac{n\kappa_2}{2\pi}} \exp\left(-nh_1(\tilde{\theta}_1) + nh_2(\tilde{\theta}_2)\right) \\ &= \sqrt{\frac{\kappa_1}{\kappa_2}} \exp\left(-n\left(h_1(\tilde{\theta}_1) - h_2(\tilde{\theta}_2)\right)\right).\end{aligned} \quad (695)$$

□

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- **Monte Carlo integration**
- Importance sampling
- Acceptance-rejection sampling
- Exercises

Monte Carlo integration

Monte Carlo integration

- ... is a means to approximate integrals using random sampling.
- ... was allegedly a code word in the Manhattan project.
- ... is used for high-dimensional integration with strong localization.
- ... replaces deterministic support points with random support points.

Definition (Monte Carlo estimator, Monte Carlo algorithm)

For a univariate continuous random variable X with PDF p and a univariate real-valued function f , let

$$I := \mathbb{E}_{p(x)}(f(X)) := \int_{\mathcal{X}} f(x)p(x) dx. \quad (696)$$

Furthermore, let X_1, \dots, X_n be a sample of independent copies of X . Then

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^n f(X_i) \quad (697)$$

is called the *Monte Carlo estimator* of the integral I . A *Monte Carlo algorithm* to obtain a Monte Carlo estimate of I is given by

- (1) Sample $X_1, \dots, X_n \sim p$
- (2) Evaluate and return $\hat{I}_n = \frac{1}{n} \sum_{i=1}^n f(X_i)$.

Theorem (Unbiasedness and consistency of the Monte Carlo estimator)

The Monte Carlo estimator is unbiased and consistent.

Proof

Unbiasedness

$$\mathbb{E}(\hat{I}_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n f(X_i)\right) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}(f(X_i)) = \frac{1}{n} n \mathbb{E}(f(X)) = \mathbb{E}(f(X)). \quad (698)$$

Consistency

The weak law of large numbers states that

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}(f(X))\right| \geq \epsilon\right) = 0 \quad (699)$$

which implies the consistency of the estimator $\hat{I}_n = \frac{1}{n} \sum_{i=1}^n f(X_i)$

□

Example (Monte Carlo estimator)

Consider evaluating the integral

$$I = \int_0^1 \frac{1}{1-x^2} dx. \quad (700)$$

Then a Monte Carlo estimator of I is

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^n \frac{1}{1-x_i^2}, \quad (701)$$

where x_1, \dots, x_n denote n independent realizations of uniformly distributed random variables $X_1, \dots, X_n \sim U(0, 1)$.

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- **Importance sampling**
- Acceptance-rejection sampling
- Exercises

Importance sampling

- ... is a method to approximate expected values $\mathbb{E}_{p(x)}(f(X))$.
- ... works by sampling $X \sim q$ instead of $X \sim p$.
- ... can also be applied, if the normalizing constants of q and p are unknown.
- ... can be used to reduce the variance of Monte Carlo estimators.

Theorem (Importance sampling identity)

Let p and q be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ such that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Then for any $f : \mathcal{X} \rightarrow \mathbb{R}$, it holds that

$$\mathbb{E}_{p(x)}(f(X)) = \mathbb{E}_{q(x)}(f(X)w(X)), \quad (702)$$

where

$$w : \mathcal{X} \rightarrow \mathbb{R}_{>0}, x \mapsto w(x) := \frac{p(x)}{q(x)} \quad (703)$$

denotes the *importance weight function*.

Proof

$$\mathbb{E}_{p(x)}(f(X)) = \int_{\mathcal{X}} f(x)p(x) dx = \int_{\mathcal{X}} f(x) \frac{p(x)}{q(x)} q(x) dx = \int_{\mathcal{X}} f(x)w(x)q(x) dx = \mathbb{E}_{q(x)}(f(X)w(X)) \quad (704)$$

□

Remarks

- Note that the expectation of f under p is equal to the expectation of fw under q .
- p and q are referred to as *nominal* and *importance distributions*, respectively.

Importance sampling

Theorem (Importance sampling estimator)

Let p and q be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ and let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a function such that

$$I := \mathbb{E}_{p(x)}(f(X)) \quad (705)$$

exists. Assume that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x)f(x) \neq 0$. Finally, assume that $X_1, \dots, X_n \sim q$. Then

$$\hat{I}_n := \frac{1}{n} \sum_{i=1}^n f(X_i)w(X_i) \quad (706)$$

is an unbiased and consistent estimator of I .

Proof

We have

$$\mathbb{E}_{q(x)}(\hat{I}_n) = \mathbb{E}_{q(x)}\left(\frac{1}{n} \sum_{i=1}^n f(X_i)w(X_i)\right) = \frac{1}{n}n \mathbb{E}_{q(x)}(f(X_i)w(X_i)) = \mathbb{E}_{p(x)}(f(X_i)) \quad (707)$$

and hence the estimator is unbiased. Consistency follows with the law of large numbers.

□

Remarks

- $X_1, \dots, X_n \sim q$ and $\hat{I}_n := \frac{1}{n} \sum_{i=1}^n f(X_i) \frac{p(X_i)}{q(X_i)}$ are used to estimate $\mathbb{E}_{p(x)}(f(X))$.

Theorem (Normalized importance sampling identity)

For normalization constants c_p, c_q , let $p := \tilde{p}/c_p$ and $q = \tilde{q}/c_q$ be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ such that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Then for any $f : \mathcal{X} \rightarrow \mathbb{R}$, it holds that

$$\mathbb{E}_{p(x)}(f(X)) = \frac{\mathbb{E}_{q(x)}(f(X)\tilde{w}(X))}{\mathbb{E}_{q(x)}(\tilde{w}(X))}, \quad (708)$$

where \tilde{w} denotes the importance weight function

$$\tilde{w} : \mathcal{X} \rightarrow \mathbb{R}_{>0}, x \mapsto \tilde{w}(x) := \frac{\tilde{p}(x)}{\tilde{q}(x)}. \quad (709)$$

Remark

- $\mathbb{E}_{p(x)}(f(X))$ is here estimated based on an “unnormalized PDF” $\tilde{p} = c_p p$.

Importance sampling

Proof

$$\begin{aligned} \frac{\mathbb{E}_{q(x)}(f(X)\tilde{w}(X))}{\mathbb{E}_{q(x)}(\tilde{w}(X))} &= \frac{\mathbb{E}_{q(x)}\left(f(X)\frac{\tilde{p}(X)}{\tilde{q}(X)}\right)}{\mathbb{E}_{q(x)}\left(\frac{\tilde{p}(X)}{\tilde{q}(X)}\right)} \\ &= \frac{\mathbb{E}_{q(x)}\left(f(X)\frac{c_p c_q \tilde{p}(X)}{c_p c_q \tilde{q}(X)}\right)}{\mathbb{E}_{q(x)}\left(\frac{c_p c_q \tilde{p}(X)}{c_p c_q \tilde{q}(X)}\right)} \\ &= \frac{\mathbb{E}_{q(x)}\left(f(X)\frac{c_q p(X)}{c_p q(X)}\right)}{\mathbb{E}_{q(x)}\left(\frac{c_q p(X)}{c_p q(X)}\right)} \\ &= \frac{\int_{\mathcal{X}} f(x) \frac{c_q p(x)}{c_p q(x)} q(x) dx}{\int_{\mathcal{X}} \frac{c_q p(x)}{c_p q(x)} q(x) dx} \tag{710} \\ &= \frac{\int_{\mathcal{X}} f(x) \frac{c_q p(x)}{c_p} dx}{\int_{\mathcal{X}} \frac{c_q p(x)}{c_p} dx} \\ &= \frac{\frac{c_q}{c_p} \int_{\mathcal{X}} f(x) p(x) dx}{\frac{c_q}{c_p} \int_{\mathcal{X}} p(x) dx} \\ &= \int_{\mathcal{X}} f(x) p(x) dx = \mathbb{E}_{p(x)}(f(X)) \end{aligned}$$

□

Definition (Normalized importance sampling estimator)

For an unknown normalization constant c_p and $c_q := 1$, let $p := \tilde{p}/c_p$ and $q := \tilde{q}/c_q$ be probability density functions on $\mathcal{X} \subseteq \mathbb{R}$ such that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a function such that

$$I := \mathbb{E}_{p(x)}(f(X)) \quad (711)$$

exists. Assume that $q(x) > 0$ for all $x \in \mathcal{X}$ with $p(x)f(x) \neq 0$. Finally, assume that $X_1, \dots, X_n \sim q$. Then

$$\hat{I}_n := \frac{\sum_{i=1}^n f(X_i)\tilde{w}(X_i)}{\sum_{i=1}^n \tilde{w}(X_i)} \quad (712)$$

is called the *normalized importance sampling estimator* of I .

Remarks

- $X_1, \dots, X_n \sim q$ and \hat{I}_n are used to estimate I for the “unnormalized PDF” \tilde{p} .
- The sample mean factor n^{-1} cancels out in the \hat{I}_n fraction.
- It can be shown that \hat{I}_n is biased, but asymptotically unbiased and consistent.

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- **Acceptance-rejection sampling**
- Exercises

Acceptance-rejection sampling

Acceptance-rejection sampling

- ... is a method to obtain samples from $Y \sim p_Y$ by sampling $X \sim p_X$.
- ... an algorithmic alternative to the probability integral transform.
- ... works by strategically rejecting samples from X .
- ... works by inducing a sampling bias to the samples from X .
- ... can be used in MC integration, if sampling from the posterior is difficult.

Acceptance-rejection sampling

Theorem (Acceptance-rejection sampling)

Let p_X and p_Y be two probability density functions, referred to as *proposal density* and *target density*, respectively, such that

- o random samples of $X \sim p_X$ can be obtained,
- o the function p_Y/p_X can be evaluated, and
- o $p_Y \leq cp_X$ for a constant $c \in \mathbb{R}$.

Consider the following *acceptance-rejection algorithm*

1. Draw a realization $X \sim p_X$ of the proposal density.
2. Draw a realization $U \sim U(0, 1)$ of the uniform distribution on $[0, 1]$.
3. If

$$U \leq \frac{p_Y(X)}{cp_X(X)} \quad (713)$$

return $Y = X$. Otherwise, return to step 1.

Then Y is distributed according to the target density, $Y \sim p_Y$.

Remarks

- Note that the algorithm returns $X \sim p_X$ conditional on $U \leq \frac{p_Y(X)}{cp_X(X)}$.
- The claim of the theorem can thus equivalently be expressed as

$$\mathbb{P} \left(X \leq x \middle| U \leq \frac{p_Y(X)}{cp_X(X)} \right) = \int_{-\infty}^x p_Y(\xi) d\xi. \quad (714)$$

Proof

We want to show that the conditional distribution of X given the event $U \leq p_Y(X)/cp_X(X)$ conforms to the target density p_Y . For ease of notation, let

$$f(X) := p_Y(X)/cp_X(X). \quad (715)$$

The conditional probability of the event $X \leq x$ given the event $U \leq f(X)$ can then be written as

$$\mathbb{P}(X \leq x | U \leq f(X)) = \frac{\mathbb{P}(X \leq x, U \leq f(X))}{\mathbb{P}(U \leq f(X))}. \quad (716)$$

To evaluate the numerator of the right-hand side of the above, we first note that

$$\begin{aligned} \mathbb{P}(X \leq x, U \leq f(X)) &= \mathbb{P}(U \leq f(X) | X \leq x) \mathbb{P}(X \leq x) \\ &= \mathbb{P}(U \leq f(x)) \mathbb{P}(X \leq x) \\ &= \int_{-\infty}^x \mathbb{P}(U \leq f(\xi)) p_X(\xi) d\xi. \end{aligned} \quad (717)$$

With the CDF of the continuous uniform distribution on $[0, 1]$ and the definition of $f(X)$, we then obtain

Proof (cont.)

$$\begin{aligned}
 \mathbb{P}(X \leq x, U \leq f(X)) &= \int_{-\infty}^x \mathbb{P}(U \leq f(\xi)) p_X(\xi) d\xi \\
 &= \int_{-\infty}^x f(\xi) p_X(\xi) d\xi \\
 &= \int_{-\infty}^x \frac{p_Y(\xi)}{c p_X(\xi)} p_X(\xi) d\xi \\
 &= \frac{1}{c} \int_{-\infty}^x p_Y(\xi) d\xi.
 \end{aligned} \tag{718}$$

We next evaluate the denominator of the above and obtain

$$\begin{aligned}
 \mathbb{P}(U \leq f(X)) &= \int_{-\infty}^{\infty} \mathbb{P}(U \leq f(X) | X = x) \mathbb{P}(X = x) dx \\
 &= \int_{-\infty}^{\infty} \mathbb{P}(U \leq f(x)) \mathbb{P}(X = x) dx \\
 &= \int_{-\infty}^{\infty} \mathbb{P}(U \leq f(\xi)) p_X(\xi) d\xi \\
 &= \int_{-\infty}^{\infty} \frac{p_Y(\xi)}{c p_X(\xi)} p_X(\xi) d\xi \\
 &= \frac{1}{c}.
 \end{aligned} \tag{719}$$

Proof (cont.)

In summary, we obtain

$$\mathbb{P}(X \leq x | U \leq f(X)) = \frac{\frac{1}{c} \int_{-\infty}^x p_Y(\xi) d\xi}{\frac{1}{c}} d\xi = \int_{-\infty}^x p_Y(\xi) d\xi \quad (720)$$

The conditional distribution of X given $U \leq \frac{p_Y(X)}{cp_X(X)}$ thus has PDF p_Y . Because we call this random variable distributed according to this distribution Y , we have obtained $Y \sim p_Y$.

□

Numerical methods

- Motivation
- Quadrature
- Laplace approximation
- Monte Carlo integration
- Importance sampling
- Acceptance-rejection sampling
- **Exercises**

Study questions

1. Name two quantities in Bayesian inference that often necessitate numerical integration.
2. Name an example for a quadrature rule.
3. What is the difference between the right rule and the midpoint rule in Riemann sum-based numerical integration?
4. State Laplace's integral approximation method.
5. Write down the Laplace approximation of a posterior expectation of the form $\mathbb{E}_{p(\theta|x_{1:n})}(f(\theta))$.
6. Write down the definition of the Monte Carlo estimator of an integral $I = \int_{\mathcal{X}} f(x)p(x) dx$.
7. State the importance sampling identity.
8. Write down the acceptance-rejection algorithm.

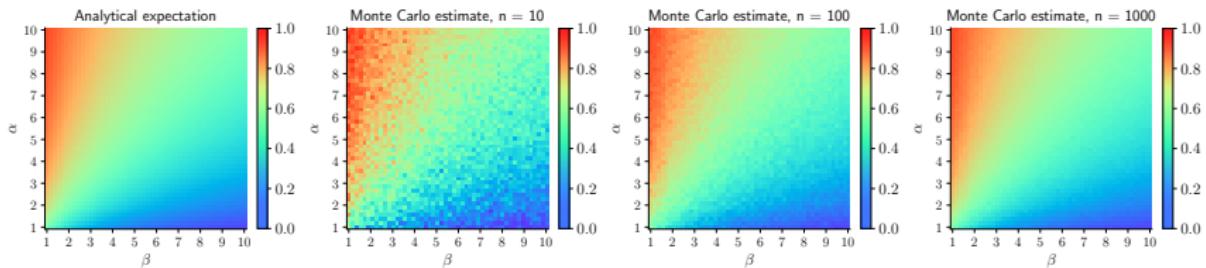
Programming exercises

1. Estimate the expected value of a $\text{Beta}(\alpha, \beta)$ for varying values of α and β by means of Monte Carlo integration by using a Beta distribution random number generator. Compare the results to the true expected values.
2. Estimate the expected value of a $\text{Beta}(\alpha, \beta)$ for varying values of α and β by means of Monte Carlo integration using an importance sampling scheme and a uniform random number generator.
3. Use an acceptance-rejection algorithm to sample random numbers from $\text{Beta}(2, 6)$.

Exercises

Programming Exercise 1

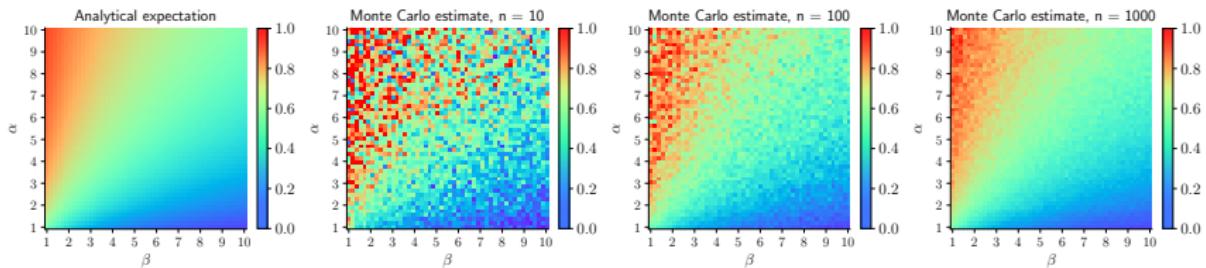
Monte Carlo estimation of $\text{Beta}(\alpha, \beta)$ expectation
using a Beta distribution random number generator



Exercises

Programming Exercise 2

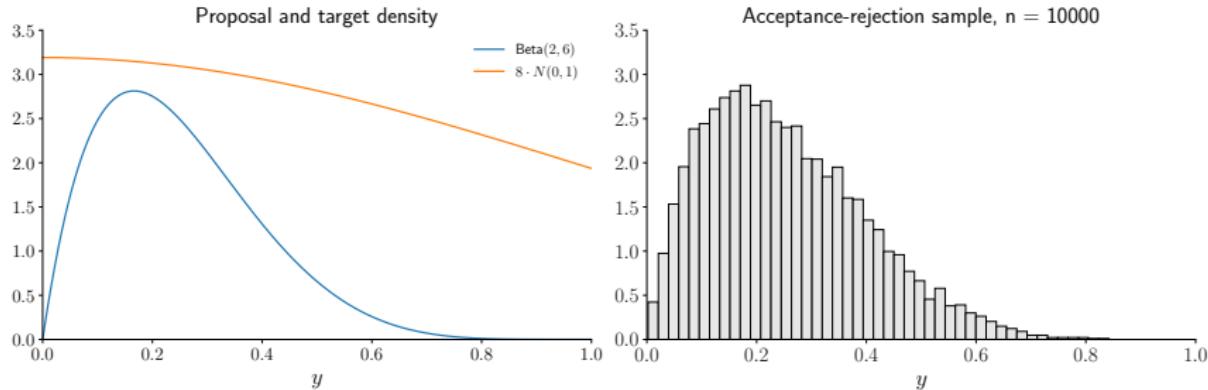
Monte Carlo estimation of $\text{Beta}(\alpha, \beta)$ expectation
using an importance sampling scheme



Exercises

Programming Exercise 3

Acceptance-rejection sampling



(15) Variational inference

Bibliographic remarks

The material presented in this section follows Ostwald et al. (2014). A good starting point for current developments in variational inference is the work Blei et al. (2017). The exemplar applications of variational inference are based on Penny et al. (2003), Friston et al. (2007), and Ostwald and Starke (2016).

Variational inference

- Foundations of variational inference
- Free-form variational inference for a Gaussian-Gamma model
- Fixed-form variational inference for nonlinear Gaussian models
- Exercises

Variational inference

- **Foundations of variational inference**
- Free-form variational inference for a Gaussian-Gamma model
- Fixed-form variational inference for nonlinear Gaussian models
- Exercises

Definition (Variational inference problem)

Let $X = (X_1, \dots, X_n)$ denote a set of observable random variables, and let $Z = (Z_1, \dots, Z_m)$ denote a set of latent random variables. Assume that X and Z form a *probabilistic model*

$$p(x, z) = p(x|z)p(z). \quad (721)$$

Then the *variational inference problem* is to approximate

- (1) the conditional distribution $p(z|x)$, and
- (2) the log model evidence $\ln p(x) = \ln \int p(x, z) dz$.

Remarks

- Conditional densities and marginal likelihoods are central in Bayesian inference.
- Variational inference is also known as variational Bayes.
- The latent variables may represent parameters or latent states.
- It is more common for $p(x, y)$ to be a PDF than a PMF.

Theorem (Log model evidence decomposition)

Let $p(x, z)$ denote a probabilistic model and let $q(z)$ denote a distribution over the latent random variables, referred to as *variational distribution*. Then the following log model evidence decomposition holds

$$\ln p(x) = \text{ELBO}(q(z)) + \text{KL}(q(z)||p(z|x)), \quad (722)$$

where

$$\text{ELBO}(q(z)) := \int q(z) \ln \left(\frac{p(x, z)}{q(z)} \right) dz \quad (723)$$

is referred to as *evidence lower bound*

$$\text{KL}(q(z)||p(z|x)) := \int q(z) \ln \left(\frac{q(z)}{p(z|x)} \right) dz \quad (724)$$

is referred to as *Kullback-Leibler divergence* between $q(z)$ and $p(z|x)$.

Remarks

- Parameters of $q(z)$ are referred to as *variational parameters*.
- The ELBO is sometimes referred to as *variational free energy*.

Proof

With the definitions of the evidence lower bound and the Kullback-Leibler divergence we have

$$\begin{aligned}\text{ELBO}(q(z)) &= \int q(z) \ln \left(\frac{p(x, z)}{q(z)} \right) dz \\ &= \int q(z) \ln \left(\frac{p(x)p(z|x)}{q(z)} \right) dz \\ &= \int q(z) \ln p(x) dz + \int q(z) \ln \left(\frac{p(z|x)}{q(z)} \right) dz \quad (725) \\ &= \ln p(x) \int q(z) dz - \int q(z) \ln \left(\frac{q(z)}{p(z|x)} \right) dz \\ &= \ln p(x) - \text{KL}(q(z)||p(z|x)),\end{aligned}$$

from which the log model evidence decomposition follows immediately. □

Theorem (Kullback-Leibler divergence)

The *Kullback-Leibler (KL) divergence* of two PDFs $q(z)$ and $p(z)$ is defined as

$$\text{KL}(q(z)||p(z)) := \int q(z) \ln \left(\frac{q(z)}{p(z)} \right) dz \quad (726)$$

Theorem (Kullback-Leibler divergence)

The *Kullback-Leibler (KL) divergence* of two PDFs $q(z)$ and $p(z)$ is defined as

$$\text{KL}(q(z)||p(z)) := \int q(z) \ln \left(\frac{q(z)}{p(z)} \right) dz \quad (726)$$

It serves as a dissimilarity measure of $q(z)$ and $p(z)$. Specifically,

$$\text{KL}(q(z)||p(z)) > 0 \text{ for } q(z) \neq p(z). \quad (727)$$

and

$$\text{KL}(q(z)||p(z)) = 0 \text{ for } q(z) = p(z) \quad (728)$$

Theorem (Kullback-Leibler divergence)

The *Kullback-Leibler (KL) divergence* of two PDFs $q(z)$ and $p(z)$ is defined as

$$\text{KL}(q(z)||p(z)) := \int q(z) \ln \left(\frac{q(z)}{p(z)} \right) dz \quad (726)$$

It serves as a dissimilarity measure of $q(z)$ and $p(z)$. Specifically,

$$\text{KL}(q(z)||p(z)) > 0 \text{ for } q(z) \neq p(z). \quad (727)$$

and

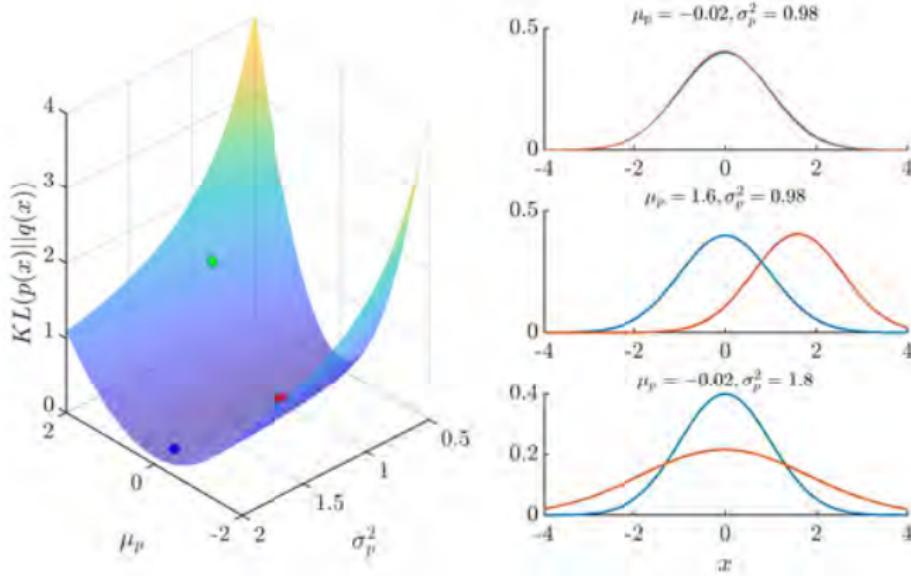
$$\text{KL}(q(z)||p(z)) = 0 \text{ for } q(z) = p(z) \quad (728)$$

Remarks

- The non-negativity of the KL divergence is central for variational inference.
- Proofs of the KL divergence properties will be discussed in the Übung.
- The KL divergence is not a metric, but $\frac{1}{2} (\text{KL}(q(z)||p(z)) + \text{KL}(p(z)||q(z)))$ is.

Example (KL divergences of univariate Gaussians)

$$\text{KL} \left(N(x; \mu_q, \sigma_q^2) || N(x; \mu_p, \sigma_p^2) \right), \mu_q := 0, \sigma_q^2 := 1. \quad (729)$$



Theorem (Evidence lower bound)

The ELBO is a lower bound of the log model evidence,

$$\text{ELBO}(q(x)) \leq \ln p(x). \quad (730)$$

Proof

The lower bound property of the ELBO follows directly from the log model evidence decomposition

$$\ln p(x) = \text{ELBO}(q(z)) + \text{KL}(q(z)||p(z|x)) \quad (731)$$

and the non-negativity of the KL-divergence

$$\text{KL}(q(z)||p(z|x)) \geq 0. \quad (732)$$

□

Foundations of variational inference

Remarks

- For a fixed data set x , $\ln p(x)$ is a fixed scalar quantity.

$$\ln p(x) = \ln \int p(x, z) dz$$

$$\text{ELBO}(q(z))$$

$$\text{KL}(q(z)||p(z|x))$$

- The ELBO property suggests two equivalent options for solving the VI problem.

(1) Minimize $\text{KL}(q(z)||p(z|x))$ with respect to $q(z)$

$$\Rightarrow q(z) \rightarrow p(z|x) \text{ and } \text{ELBO}(q(z)) \rightarrow \ln p(x).$$

(2) Maximize $\text{ELBO}(q(z))$ with respect to $q(z)$

$$\Rightarrow \text{ELBO}(q(z)) \rightarrow \ln p(x) \text{ and } q(z) \rightarrow p(z|x).$$

- Variational inference typically follows the latter approach.

Definition (Mean-field variational approximation, inference)

Let $Z = (Z_1, \dots, Z_m)$ denote the latent random variables of a probabilistic model in the variational inference problem setting and let $\mathcal{S} := (Z_{s_1}, \dots, Z_{s_S})$ denote a partition of Z into S mutually exclusive subsets, such that $\cup_{i=1}^S Z_{s_i} = Z$ and $Z_{s_i} \cap Z_{s_j} = \emptyset$ for $i \neq j$. Then the assumption that Z_{s_1}, \dots, Z_{s_S} form subsets of independent variables, i.e.,

$$q(z) := \prod_{i=1}^S q(z_{s_i}), \quad (733)$$

where $q(z_{s_i})$ denotes the distribution of the elements of $Z_{s_i}, i = 1, \dots, S$ is referred to as *mean-field approximation* of the variational distribution. Using a mean-field approximation for the variational distribution in variational inference is referred to as *mean-field variational inference*.

Remarks

- Common mean-field approximations in variational inference are
 - $q(z) = \prod_{i=1}^m q(z_i)$ (full factorization)
 - $q(z) = q(z_s)q(z_{\setminus s})$, where $Z_s \subset Z$ and $Z_{\setminus s} := Z \setminus Z_s$ (binary factorization)
- Mean-field variational inference neglects potential non-independencies of $p(z|x)$.

Theorem (Free-form mean-field variational inference)

Let $p(x, z)$ denote a probabilistic model comprising observable random variables $X = (X_1, \dots, X_n)$ and latent random variables $Z = (Z_1, \dots, Z_m)$. Let $q(z)$ denote a variational distribution and assume that $q(z)$ factorizes in a binary manner, i.e., $q(z) = q(z_s)q(z_{\setminus s})$ with $Z_s \subset Z$ and $Z_{\setminus s} := Z \setminus Z_s$. Then setting

$$q(z_s) := \frac{1}{\gamma_s} \exp \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right), \quad (734)$$

where γ_s denotes a normalization constant independent of z_s , maximizes the evidence lower bound with respect to $q(z_s)$.

Remarks

- Exchanging the roles of $q(z_s)$ and $q(z_{\setminus s})$ maximizes the ELBO with respect to $q(z_{\setminus s})$, which implies a coordinate-wise ELBO maximization approach.
- The functional form of the maximizing variational distributions depends on the right-hand side of the proportionality statement, hence “free-form”.

Foundations of variational inference

Proof

Below, we show that for a mean-field approximation $q(z) = q(z_s)q(z_{\setminus s})$, the ELBO can be rewritten as

$$\text{ELBO}(q(z_s), q(z_{\setminus s})) = -\text{KL}\left(q(z_s) \middle\| \frac{1}{\gamma_s} \exp\left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s}\right)\right) - c, \quad (735)$$

where $c \in \mathbb{R}$ is a constant not depending on $q(z_s)$. Maximizing this negative KL divergence by setting

$$q(z_s) := \frac{1}{\gamma_s} \exp\left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s}\right) \quad (736)$$

thus maximizes $\text{ELBO}(q(z_s)q(z_{\setminus s}))$ with respect to $q(z_s)$.

ELBO reformulation

$$\begin{aligned} \text{ELBO}(q(z_s), q(z_{\setminus s})) &= \iint q(z_s)q(z_{\setminus s}) \ln \left(\frac{p(x, z)}{q(z_s)q(z_{\setminus s})} \right) dz_s dz_{\setminus s} \\ &= \iint q(z_s)q(z_{\setminus s}) (\ln p(x, z) - \ln q(z_s) - \ln q(z_{\setminus s})) dz_s dz_{\setminus s} \\ &= \iint q(z_s)q(z_{\setminus s}) (\ln p(x, z) - \ln q(z_s)) dz_{\setminus s} dz_s - \iint q(z_s)q(z_{\setminus s}) \ln q(z_{\setminus s}) dz_s dz_{\setminus s} \\ &= \iint q(z_s)q(z_{\setminus s}) (\ln p(x, z) - \ln q(z_s)) dz_{\setminus s} dz_s - \int q(z_{\setminus s}) \ln q(z_{\setminus s}) \left(\int q(z_s) dz_s \right) dz_{\setminus s} \\ &= \iint q(z_s)q(z_{\setminus s}) (\ln p(x, z) - \ln q(z_s)) dz_{\setminus s} dz_s - \int q(z_{\setminus s}) \ln q(z_{\setminus s}) dz_{\setminus s} \\ &= \iint q(z_s)q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} dz_s - \iint q(z_s)q(z_{\setminus s}) \ln q(z_s) dz_{\setminus s} dz_s + c_1 \end{aligned}$$

Foundations of variational inference

Proof (cont.)

$$\begin{aligned}
&= \int q(z_s) \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right) dz_s - \int q(z_s) \ln q(z_s) \left(\int q(z_{\setminus s}) dz_{\setminus s} \right) dz_s + c_1 \\
&= \int q(z_s) \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right) dz_s - \int q(z_s) \ln q(z_s) dz_s + c_1 \\
&= \int q(z_s) \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right) dz_s - \int q(z_s) \ln q(z_s) dz_s + \int q(z_s) \ln \gamma_s dz_s - \int q(z_s) \ln \gamma_s dz_s + c_1 \\
&= \int \left(q(z_s) \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right) - q(z_s) \ln q(z_s) - q(z_s) \ln \gamma_s \right) dz_s + c_2 \\
&= \int q(z_s) \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} - \ln q(z_s) - \ln \gamma_s \right) dz_s + c_2 \\
&= \int q(z_s) \left(\ln \left(\exp \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right) \right) - \ln q(z_s) - \ln \gamma_s \right) dz_s + c_2 \\
&= \int q(z_s) \left(\ln \left(\frac{\frac{1}{\gamma_s} \exp \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right)}{q(z_s)} \right) \right) dz_s + c_2 \\
&= - \int q(z_s) \left(\ln \left(\frac{q(z_s)}{\frac{1}{\gamma_s} \exp \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right)} \right) \right) dz_s + c_2 \\
&= -\text{KL} \left(q(z_s) \middle\| \frac{1}{\gamma_s} \exp \left(\int q(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right), \right) + c_2
\end{aligned}$$

where we defined

$$c_1 := - \int q(z_{\setminus s}) \ln q(z_{\setminus s}) dz_{\setminus s} \text{ and } c_2 := c_1 + \int q(z_s) \ln \gamma_s dz_s = c_1 + \ln \gamma_s. \quad (737)$$

□

Definition (Coordinate-ascent variational inference (CAVI))

Let $p(x, z)$ denote a probabilistic model comprising observable random variables $X = (X_1, \dots, X_n)$ and latent random variables $Z = (Z_1, \dots, Z_m)$. Let $q(z)$ denote a variational distribution and assume that $q(z)$ factorizes in a binary manner. Then the following algorithm, referred to as *coordinate-ascent variational inference*, maximizes the evidence lower bound:

Initialization

- (0) Initialize $q^{(0)}(z_s)$ and $q^{(0)}(z_{\setminus s})$ appropriately, e.g., by setting $q^{(0)}(z_s) := \int p(z) dz_{\setminus s}$ and $q^{(0)}(z_{\setminus s}) := \int p(z) dz_s$, evaluate $\text{ELBO} \left(q^{(0)}(z_s) q^{(0)}(z_{\setminus s}) \right)$, and select a convergence criterion $\delta > 0$.

For $i = 0, 1, 2, \dots$ until $|\text{ELBO} \left(q^{(i+1)}(z_s) q^{(i+1)}(z_{\setminus s}) \right) - \text{ELBO} \left(q^{(i)}(z_s) q^{(i)}(z_{\setminus s}) \right)| < \delta$

(1) Set $q^{(i+1)}(z_s) := \frac{1}{\gamma_s} \exp \left(\int q^{(i)}(z_{\setminus s}) \ln p(x, z) dz_{\setminus s} \right)$

(2) Set $q^{(i+1)}(z_{\setminus s}) := \frac{1}{\gamma_{\setminus s}} \exp \left(\int q^{(i+1)}(z_s) \ln p(x, z) dz_s \right)$

(3) Evaluate $\text{ELBO} \left(q^{(i+1)}(z_s), q^{(i+1)}(z_{\setminus s}) \right)$.

Return $p(z|x) \approx q^{(i+1)}(z_s) q^{(i+1)}(z_{\setminus s})$ and $\ln p(x) \approx \text{ELBO} \left(q^{(i+1)}(z_s) q^{(i+1)}(z_{\setminus s}) \right)$.

Foundations of variational inference

Coordinate-ascent variational inference

Before iteration $i = 1, 2, \dots$

$\text{ELBO}\left(q^{(i)}(z_s), q^{(i)}(z_{\setminus s})\right)$	$\text{KL}(q^{(i)}(z_s)q^{(i)}(z_{\setminus s}) p(z x))$
--	---

On iteration $i = 1, 2, \dots$

$\text{ELBO}\left(q^{(i+1)}(z_s), q^{(i)}(z_{\setminus s})\right)$	$\text{KL}(q^{(i+1)}(z_s)q^{(i)}(z_{\setminus s}) p(z x))$
--	---

$\text{ELBO}\left(q^{(i+1)}(z_s), q^{(i+1)}(z_{\setminus s})\right)$	$\text{KL}(q^{(i+1)}(z_s)q^{(i+1)}(z_{\setminus s}) p(z x))$
--	---

Upon convergence

$$\ln p(x) \approx \text{ELBO}\left(q^{(i+1)}(z_s), q^{(i+1)}(z_{\setminus s})\right) \text{ and } p(z|x) \approx q^{(i+1)}(z_s)q^{(i+1)}(z_{\setminus s})$$

Definition (Fixed-form variational inference)

Let $p(x, z)$ denote a probabilistic model comprising observable random variables $X = (X_1, \dots, X_n)$ and latent random variables $Z = (Z_1, \dots, Z_m)$. Assume further that the resulting evidence lower bound integral

$$\text{ELBO}(q_\theta(z)) = \int q_\theta(z) \ln \left(\frac{p(x, z)}{q_\theta(z)} \right) dz \quad (738)$$

can be analytically evaluated or at least be analytically approximated. Then the evidence lower bound takes the form of a real-valued function

$$f : \Theta \rightarrow \mathbb{R}, \theta \mapsto f(\theta) := \text{ELBO}(q_\theta(z)) \quad (739)$$

Maximizing f with respect to θ thus maximizes the (approximate) evidence lower bound.

Remark

- The functional form of the variational distribution is predefined, for example

$$q(z) := N(z; \theta_q, \Sigma_q), \quad (740)$$

hence “fixed-form”. Using Gaussian distributions as variational distributions is sometimes called “*Variational Laplace*”

Variational inference

- Foundations of variational inference
- **Free-form variational inference for a Gaussian-Gamma model**
- Fixed-form variational inference for nonlinear Gaussian models
- Exercises

Definition (Gaussian-Gamma regression model)

The a *Gaussian-Gamma regression model* is a probabilistic model of the form

$$p(x, \beta, \lambda) = p(x|\beta, \lambda)p(\beta)p(\lambda), \quad (741)$$

where

$$p(x|\beta, \lambda) := N(x; \Phi\beta, \lambda^{-1}I_n), \quad p(\beta) := N(x; 0_p, \alpha^{-1}I_p), \quad p(\lambda) := G(\lambda; \beta_\lambda, \gamma_\lambda), \quad (742)$$

where x is an n -dimensional observed random vector modeling data, β is a p -dimensional unobserved random vector of regression weights, $\lambda > 0$ is an unobserved random variable modeling the data noise precision, $\Phi \in \mathbb{R}^{n \times p}$ is a design matrix, and $\alpha, \beta_\lambda, \gamma_\lambda > 0$ are hyperparameters.

Remarks

- The unobservable random vector takes the form $z = (\beta, \lambda)$.
- The model is an exemplary non-conjugate Bayesian regression model.

Theorem (Free-form VI for the Gaussian-Gamma model)

Application of the free form mean-field variational inference theorem to the Gaussian-Gamma model for the mean-field approximation

$$q(\beta, \lambda) := q(\beta)q(\lambda) \quad (743)$$

yields the following CAVI algorithm: *Initialization*

0. Set

$$q^{(0)}(\beta) := N\left(\beta; m_\beta^{(0)}, S_\beta^{(0)}\right) \text{ and } q^{(0)}(\lambda) := G\left(\lambda; b_\lambda^{(0)}, c_\lambda^{(0)}\right) \quad (744)$$

with variational parameters

$$m_\beta^{(0)} := 0_p, S_\beta^{(0)} := \alpha^{-1} I_p \text{ and } b_\lambda^{(0)} := \beta_\lambda, c_\lambda^{(0)} := \frac{n}{2} + \gamma_\lambda, \quad (745)$$

respectively. Define a convergence criterion $\Delta \text{ELBO} > 0$ and a maximum number of iterations n_i .

Free-form variational inference for a Gaussian-Gamma model

Theorem (Free-form VI for the Gaussian-Gamma model (cont.))

Iterations

For $i = 1, \dots, n_i$ or until convergence is reached

1. $q(\beta)$ update

Set

$$q^{(i)}(\beta) := N\left(\beta; m_{\beta}^{(i)}, S_{\beta}^{(i)}\right), \quad (746)$$

where

$$S_{\beta}^{(i)} := \left(b_{\lambda}^{(i-1)} c_{\lambda}^{(i-1)} \Phi^T \Phi + \alpha I_p\right)^{-1} \text{ and } m_{\beta}^{(i)} := b_{\lambda}^{(i-1)} c_{\lambda}^{(i-1)} S_{\beta}^{(i)} \Phi^T x. \quad (747)$$

2. $q(\lambda)$ update

Set

$$q^{(i)}(\lambda) := G\left(\lambda; b_{\lambda}^{(i)}, c_{\lambda}^{(i)}\right), \quad (748)$$

where

$$b_{\lambda}^{(i)} := \left(\frac{1}{2} \left(\operatorname{tr}\left(S_{\beta}^{(i)} \Phi^T \Phi\right) + \left(x - \Phi m_{\beta}^{(i)}\right)^T \left(x - \Phi m_{\beta}^{(i)}\right)\right) + \frac{1}{\beta_{\lambda}}\right)^{-1} \quad (749)$$

and

$$c_{\lambda}^{(i)} := \frac{n}{2} + \gamma_{\lambda}. \quad (750)$$

Note that $c_{\lambda}^{(i)}$ stays constant throughout.

Free-form variational inference for a Gaussian-Gamma model

Theorem (Free-form VI for the Gaussian-Gamma model (cont.))

3. ELBO ($q(\beta)$, $q(\lambda)$) update

Set

$$\text{ELBO}^{(i)} := \text{ELBO} \left(q^{(i)}(\beta), q^{(i)}(\lambda) \right), \quad (751)$$

where

$$\text{ELBO} \left(q^{(i)}(\beta) q^{(i)}(\lambda) \right) := L_a^{(i)} - \text{KL} \left(q^{(i)}(\beta) || p(\beta) \right) - \text{KL} \left(q^{(i)}(\lambda) || p(\lambda) \right), \quad (752)$$

where with the digamma function ψ , $L_a^{(i)}$ denotes the average likelihood term

$$L_a^{(i)} := -\frac{n}{2} \ln 2\pi - \frac{1}{2} b_\lambda^{(i)} c_\lambda^{(i)} \left(x - \Phi m_\beta^{(i)} \right)^T \left(x - \Phi m_\beta^{(i)} \right) \quad (753)$$

$$- \frac{1}{2} b_\lambda^{(i)} c_\lambda^{(i)} \text{tr} \left(S_\beta^{(i)} \Phi^T \Phi \right) + \frac{n}{2} \psi \left(c_\lambda^{(i)} \right) + \ln b_\lambda^{(i)} \quad (754)$$

and $\text{KL}(q(x)||p(x))$ denotes the KL-divergence between the densities $q(x)$ and $p(x)$.

4. Convergence assessment

If $i > 1$, evaluate $\Delta \text{ELBO} = \text{ELBO}^{(i)} - \text{ELBO}^{(i-1)}$. Then, if $\Delta \text{ELBO} < 0$ issue a warning and end the algorithm, if $0 < \Delta \text{ELBO} < \delta$ end the algorithm and declare convergence, or else go to 1.

Remarks

- For a derivation of the algorithm, see Penny et al. (2003).

Variational inference

- Foundations of variational inference
- Free-form variational inference for a Gaussian-Gamma model
- **Fixed-form variational inference for nonlinear Gaussian models**
- Exercises

Definition (Nonlinear Gaussian model)

Let

$$p(x, \theta) = p(x|\theta)p(\theta) \quad (755)$$

denote the joint probability density function of an n -dimensional observable random vector $X = (X_1, \dots, X_n)$ and an m -dimensional latent random vector $\theta = (\theta_1, \dots, \theta_m)$. We call $p(x, \theta)$ a *nonlinear Gaussian model*, if the prior and likelihood take the forms

$$p(\theta) = N(\theta; \mu_\theta, \Sigma_\theta) \text{ with known } \mu_\theta \in \mathbb{R}^m \text{ and } \Sigma_\theta \in \mathbb{R}^{m \times m} \text{ p.d.} \quad (756)$$

and

$$p(x|\theta) = N(x; f(\theta), \lambda_x^{-1} I_n) \text{ with known } \lambda_x > 0 \text{ and } f(\theta) \in \mathbb{R}^n, \quad (757)$$

respectively, where

$$f : \mathbb{R}^m \rightarrow \mathbb{R}^n, \theta \mapsto f(\theta) \quad (758)$$

is a nonlinear differentiable multivariate vector-valued function.

Remarks

- For linear-affine f Bayesian inference follows the Gaussian theorems.
- The likelihood can be written $X = f(\theta) + \varepsilon, \varepsilon \sim N(0_n, \lambda_x^{-1} I_n)$.
- In this structural likelihood form, $\varepsilon_1, \dots, \varepsilon_n \sim N(0, \lambda_x^{-1})$ are i.i.d. error contributions.
- The model can be conceived as a nonlinear general linear model extension.

Theorem (An ELBO for the nonlinear Gaussian model)

For the variational distribution

$$q(\theta) := N(\theta; m_\theta, S_\theta) \text{ with } m_\theta \in \mathbb{R}^m \text{ and } S_\theta \in \mathbb{R}^{m \times m} \quad (759)$$

the ELBO of the nonlinear Gaussian model can be approximated by the function

$$\phi : \mathbb{R}^m \times \mathbb{R}^{m \times m} \rightarrow \mathbb{R}, (m_\theta, S_\theta) \mapsto \phi(m_\theta, S_\theta), \quad (760)$$

where

$$\begin{aligned} \phi(m_\theta, S_\theta) := & -\frac{n}{2} \ln 2\pi + \frac{n}{2} \ln \lambda_x - \frac{\lambda_x}{2} (x - f(m_\theta))^T (x - f(m_\theta)) - \frac{\lambda_x}{2} \operatorname{tr}(J^f(m_\theta)^T J^f(m_\theta) S_\theta) \\ & - \frac{m}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_\theta| - \frac{1}{2} (m_\theta - \mu_\theta)^T \Sigma_\theta^{-1} (m_\theta - \mu_\theta) - \frac{1}{2} \operatorname{tr}(\Sigma_\theta^{-1} S_\theta) \\ & + \frac{1}{2} \ln |S_\theta| + \frac{m}{2} \ln(2\pi e), \end{aligned} \quad (761)$$

and $J^f(m_\theta)$ denote the Jacobian of f evaluated at m_θ .

Remarks

- The approximation rests on the first-order Taylor approximation

$$f(\theta) \approx f(m_\theta) + J^f(m_\theta)(\theta - \mu_\theta) \quad (762)$$

- For a full derivation of the approximation, see Ostwald and Starke (2016).

Theorem (ELBO maximization with respect to S_θ)

The ELBO approximation function ϕ for the can be maximized with respect to S_θ by setting

$$S_\theta := \left(\lambda_x J^f(m_\theta)^T J^f(m_\theta) + \Sigma_\theta^{-1} \right)^{-1}. \quad (763)$$

Remarks

- The update rule was suggested by Friston et al. (2007).
- Below we provide a heuristic proof to demonstrate the general idea.
- A full proof would require the proof that ϕ is a concave function.
- A full proof would require a development of matrix derivatives.

Proof

For the derivative of ϕ with respect to S_θ , we have

$$\frac{\partial}{\partial S_\theta} \phi(m_\theta, S_\theta) = -\frac{\lambda_x}{2} \frac{\partial}{\partial S_\theta} \text{tr}(J^f(m_\theta)^T J^f(m_\theta) S_\theta) - \frac{1}{2} \frac{\partial}{\partial S_\theta} \text{tr}(\Sigma_\theta^{-1} S_\theta) + \frac{1}{2} \frac{\partial}{\partial S_\theta} \ln |S_\theta|. \quad (764)$$

This yields, using the following rules for matrix derivatives involving the trace operator and logarithmic determinants (cf. equations (103) and (57) in ?)

$$\frac{\partial}{\partial X} \text{tr}(AX^T) = A \text{ and } \frac{\partial}{\partial X} \ln |X| = (X^T)^{-1} \quad (765)$$

with $S_\theta = S_\theta^T$

$$\frac{\partial}{\partial S_\theta} \phi(m_\theta, S_\theta) = -\frac{\lambda_x}{2} J^f(m_\theta)^T J^f(m_\theta) - \frac{1}{2} \Sigma_\theta^{-1} + \frac{1}{2} S_\theta^{-1}. \quad (766)$$

Setting the above to zero then yields the equivalent relations

$$\begin{aligned} & \frac{\partial}{\partial S_\theta} \phi(m_\theta, S_\theta) = 0 \\ \Leftrightarrow & -\frac{\lambda_x}{2} J^f(m_\theta)^T J^f(m_\theta) - \frac{1}{2} \Sigma_\theta^{-1} + \frac{1}{2} S_\theta^{-1} = 0 \\ \Leftrightarrow & S_\theta = (\lambda_x J^f(m_\theta)^T J^f(m_\theta) + \Sigma_\theta^{-1})^{-1}. \end{aligned} \quad (767)$$

□

Theorem (Approximate ELBO gradient for m_θ)

The gradient of the ELBO approximation function ϕ can be approximated by

$$\nabla_{m_\theta} \phi(m_\theta, S_\theta) \approx -\lambda_x J^f(m_\theta)^T (x - f(m_\theta)) - \Sigma_\theta^{-1}(m_\theta - \mu_\theta) \quad (768)$$

Proof

We have

$$\begin{aligned} \nabla_{m_\theta} \phi(m_\theta, S_\theta) &= -\frac{\lambda_x}{2} \frac{\partial}{\partial m_\theta} \left((x - f(m_\theta))^T (x - f(m_\theta)) \right) - \frac{\lambda_x}{2} \frac{\partial}{\partial m_\theta} \text{tr} \left(J^f(m_\theta)^T J^f(m_\theta) S_\theta \right) \\ &\quad - \frac{1}{2} \frac{\partial}{\partial m_\theta} \left((m_\theta - \mu_\theta)^T \Sigma_\theta^{-1} (m_\theta - \mu_\theta) \right). \end{aligned} \quad (769)$$

Notably, the second term above involves second-order derivatives of the function f with respect to m_θ . By neglecting these terms, we obtain

$$\begin{aligned} \nabla_{m_\theta} \phi(m_\theta, S_\theta) &\approx -\frac{\lambda_x}{2} 2 \left(\frac{\partial}{\partial m_\theta} f(m_\theta) \right)^T (x - f(m_\theta)) - \frac{1}{2} 2 \Sigma_\theta^{-1} (m_\theta - \mu_\theta) \\ &= -\lambda_x J^f(m_\theta)^T (x - f(m_\theta)) - \Sigma_\theta^{-1} (m_\theta - \mu_\theta) \end{aligned} \quad (770)$$

□ Remark

- This approximate ELBO gradient was suggested by Friston et al. (2007).

Theorem (Fixed-form VI for nonlinear Gaussian models)

In summary, we have the following CAVI algorithm for $q(\theta) = N(\theta; m_\theta, S_\theta)$: *Initialization*

0. Set $q^{(0)}(\theta) := N(\beta; m_\theta^{(0)}, S_\theta^{(0)})$ with variational parameters $m_\theta^{(0)} := \mu_\theta$, $S_\theta^{(0)} := \Sigma_\theta$, define a convergence criterion $\delta > 0$ and a maximum number of iterations n_i .

Iterations

For $i = 1, \dots, n_i$ or until convergence is reached

1. S_θ update

Set $S_\theta^{(i)} := \left(\lambda_x J^f \left(m_\theta^{(i-1)} \right)^T J^f \left(m_\theta^{(i-1)} \right) + \Sigma_\theta^{-1} \right)^{-1}$.

2. m_θ update

Use a nonlinear optimization approach to evaluate $m_\theta^{(i)}$ based on $S_\theta^{(i)}$ and $m_\theta^{(i-1)}$.

3. ELBO ($q(\theta)$) update

Set $\text{ELBO}^{(i)} := \phi \left(m_\theta^{(i)}, S_\theta^{(i)} \right)$.

4. Convergence assessment

If $i > 1$, evaluate $\Delta \text{ELBO} = |\text{ELBO}^{(i)} - \text{ELBO}^{(i-1)}|$. Then, if $\Delta \text{ELBO} < 0$ issue a warning and end the algorithm, if $0 < \Delta \text{ELBO} < \delta$ end the algorithm and declare convergence, or else go to 1.

Example (An approximate gradient ascent for m_θ)

Initialization

0. Define a starting point $m_\theta^{(0)} \in \mathbb{R}^m$, a step-size $\kappa > 0$, a convergence criterion $\delta > 0$, and set $k := 0$. If $\nabla_{m_\theta} \phi(m_\theta^{(0)}, S_\theta) < \delta$, stop! $m_\theta^{(0)}$ is a zero of $\nabla_{m_\theta} \phi$. If not, proceed to iterations.

Until convergence

1. Set $m_\theta^{(k+1)} := m_\theta^{(k)} + \kappa \nabla_{m_\theta} \phi(m_\theta^{(k)}, S_\theta)$.
2. If $\nabla_{m_\theta} \phi(m_\theta^{(k+1)}, S_\theta) < \delta$, stop! $m_\theta^{(k+1)}$ is a zero of $\nabla_{m_\theta} \phi$. If not, go to 3.
3. Set $k := k + 1$ and go to 1.

Example (An approximate globalized Newton ascent for m_θ)

Initialization

0. Define a starting point $m_\theta^{(0)} \in \mathbb{R}^m$, a convergence criterion $\delta > 0$, and set $k := 0$. If $\nabla_{m_\theta} \phi(m_\theta^{(0)}, S_\theta) < \delta$, stop! $m_\theta^{(0)}$ is a zero of $\nabla_{m_\theta} \phi$. If not, proceed to iterations.

Until convergence

1. Evaluate the Newton search direction $p_k := \left(H_{m_\theta}^\phi \left(m_\theta^{(k)} \right) \right)^{-1} \nabla_{m_\theta} \phi \left(m_\theta^{(k)}, S_\theta \right)$
2. If $p_k^T \nabla_{m_\theta} \phi \left(m_\theta^{(0)}, S_\theta \right) < 0$, p_k is a descent direction. In this case, modify $H_{m_\theta}^\phi \left(m_\theta^{(k)} \right)$ to render it positive definite.
3. Evaluate a step-size t_k fulfilling the sufficient Wolfe-condition:
 - Set $t_k := 1$ and select $\rho \in]0, 1[$, $c \in]0, 1[$
 - Until $\phi \left(m_\theta^{(k)} + t_k p_k, S_\theta \right) \geq \phi \left(m_\theta^{(k)}, S_\theta \right) + c_1 t_k \nabla \phi(m_\theta^{(k)}, S_\theta)^T p_k$ set $t_k := \rho t_k$
4. Set $m_\theta^{(k+1)} := m_\theta^{(k)} + t_k (H_{m_\theta}^\phi \left(m_\theta^{(k)} \right)^{-1} \nabla_{m_\theta} \phi \left(m_\theta^{(k)}, S_\theta \right))$
5. If $\nabla_{m_\theta} \phi(m_\theta^{(k+1)}, S_\theta) < \delta$, stop! $m_\theta^{(k+1)}$ is a zero of $\nabla_{m_\theta} \phi$. If not, go to 3.
6. Set $k := k + 1$ and go to 1.

Remarks

- This algorithm was suggested by Ostwald and Starke (2016).

Current developments in variational inference

- Variational autoencoders (e.g. Kingma and Welling, 2019)

A combination of variational inference and neural network likelihood models.

- Stochastic variational inference (e.g. Hoffman, 2013)

A combination of variational inference with stochastic optimization, a technique that uses noisy estimates of a gradient to optimize an objective function.

- Quality assessment for variational estimators (e.g. Wang and Blei, 2019)

The study of how good variational inference actually is, often involving frequentist properties of variational inference point estimators.

Variational inference

- Foundations of variational inference
- Free-form variational inference for a Gaussian-Gamma model
- Fixed-form variational inference for nonlinear Gaussian models
- **Exercises**

Study questions

1. Define the variational inference problem.
2. Write down the log model evidence decomposition.
3. Write down the definition of the evidence lower bound.
4. Write down the definition and two properties of the Kullback-Leibler divergence.
5. State the evidence lower bound theorem.
6. Describe two approaches of using the evidence lower bound theorem for solving the variational inference problem.
7. Define the concept of a mean-field approximation in variational inference.
8. State the free-form mean-field variational inference theorem.
9. Write down the general CAVI algorithm.
10. Define the concept of fixed-form variational inference.

Theoretical exercises

1. Prove Jensen's inequality for the discrete case by induction.
2. Show that $\text{KL}(q(x)||p(x)) \geq 0$ with equality if and only if $q(x) = p(x)$.

Theoretical Exercise 1

Theorem (Jensen's inequality, discrete case)

Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a convex function on $]a, b[$, i.e., for $0 \leq c \leq 1$ and $x_1, x_2 \in]a, b[$ it holds that

$$cf(x_1) + (1 - c)f(x_2) \geq f(cx_1 + (1 - c)x_2), \quad (771)$$

and let $x_1, \dots, x_n \in]a, b[$ and c_1, \dots, c_n with $c_i \geq 0$ and $\sum_{i=1}^n c_i = 1$. Then

$$\sum_{i=1}^n c_i f(x_i) \geq f\left(\sum_{i=1}^n c_i x_i\right) \quad (772)$$

Exercises

Theoretical Exercise 1

Proof

We prove the theorem by induction with respect to n .

Base case

Let $n = 2$. Then the convexity of f implies that for $c_1 := c$ and $c_2 := 1 - c$ with $0 \leq c \leq 1$ and thus $c_i \geq 0, i = 1, 2$ and $\sum_{i=1}^2 c_i = 1$ it holds that

$$\begin{aligned} \sum_{i=1}^2 c_i f(x_i) &= c_1 f(x_1) + c_2 f(x_2) = cf(x_1) + (1 - c)f(x_2) \\ &\geq f(cx_1 + (1 - c)x_2) = f(c_1 x_1 + c_2 x_2) = f\left(\sum_{i=1}^2 c_i x_i\right) \end{aligned}$$

Inductive step

We assume that the induction hypothesis holds for some $n = k$, i.e.

$$\sum_{i=1}^k c_i f(x_i) \geq f\left(\sum_{i=1}^k c_i x_i\right) \quad (773)$$

and that for $c_i \geq 0, i = 1, \dots, k+1$ it holds that

$$\sum_{i=1}^{k+1} c_i = 1 \Leftrightarrow \sum_{i=1}^k c_i = 1 - c_{k+1} \Leftrightarrow \frac{1}{1 - c_{k+1}} \sum_{i=1}^k c_i = \frac{1 - c_{k+1}}{1 - c_{k+1}} \Leftrightarrow \sum_{i=1}^k \frac{c_i}{1 - c_{k+1}} = 1. \quad (774)$$

Exercises

Theoretical Exercise 1

Proof

We then have by using the convexity of f twice

$$\begin{aligned} f\left(\sum_{i=1}^{k+1} c_i x_i\right) &= f\left(\sum_{i=1}^k c_i x_i + c_{k+1} x_{k+1}\right) \\ &= f\left((1 - c_{k+1}) \sum_{i=1}^k \frac{c_i}{1 - c_{k+1}} x_i + c_{k+1} x_{k+1}\right) \\ &\leq (1 - c_{k+1}) f\left(\sum_{i=1}^k \frac{c_i}{1 - c_{k+1}} x_i\right) + c_{k+1} f(x_{k+1}) \\ &\leq (1 - c_{k+1}) \sum_{i=1}^k \frac{c_i}{1 - c_{k+1}} f(x_i) + c_{k+1} f(x_{k+1}) \\ &= \frac{1 - c_{k+1}}{1 - c_{k+1}} \sum_{i=1}^k c_i f(x_i) + c_{k+1} f(x_{k+1}) \\ &= \sum_{i=1}^{k+1} c_i f(x_i), \end{aligned} \tag{775}$$

such that the inequality also holds for $n := k + 1$.

□

Exercises

Theoretical Exercise 2

Theorem

Let $\text{KL}(q(z)||p(z))$ denote the Kullback-Leibler divergence of the PDFs $q(z)$ and $p(z)$. Then

$$\text{KL}(q(z)||p(z)) \geq 0 \text{ for } q(z) \neq p(z) \text{ and } \text{KL}(q(z)||p(z)) = 0 \text{ for } q(z) = p(z). \quad (776)$$

Proof

We first note that for a PDF $q(z)$ and a convex function f , Jensen's inequality applies, i.e.

$$\int q(z)f(z) dz \geq f\left(\int q(z)z dz\right). \quad (777)$$

Because \ln is a concave function and thus $-\ln$ is a convex function, we have

$$\begin{aligned} \text{KL}(q(z)||p(z)) &:= \int q(z) \ln\left(\frac{q(z)}{p(z)}\right) dz \\ &= - \int q(z) \ln\left(\frac{p(z)}{q(z)}\right) dz \\ &\geq - \ln\left(\int q(z) \frac{p(z)}{q(z)} dz\right) = - \ln\left(\int p(z) dz\right) = - \ln 1 = 0. \end{aligned} \quad (778)$$

Finally, for $q(z) = p(z)$, we have

$$\text{KL}(q(z)||p(z)) = \int q(z) \ln\left(\frac{q(z)}{p(z)}\right) dz = \int q(z) \cdot \ln 1 dz = \int q(z) \cdot 0 dz = 0. \quad (779)$$

Programming exercises

1. Evaluate and visualize the KL divergence between two Gaussian PDFs as well as between two Gamma PDFs for varying parameter settings of the distributions. Closed-form solutions for these KL divergences are available from the literature.
2. Implement the free-form CAVI algorithm for the Gaussian-Gamma model as introduced in the lecture. Visualize the likelihood, prior, approximate posterior, iterative variational distributions, as well as the ELBO.

Exercises

Programming Exercise 1

KL divergence for univariate Gaussians

Let

$$q(z) := N(z; \mu_q, \sigma_q^2) \text{ and } p(z) := N(z; \mu_p, \sigma_p^2) \quad (780)$$

Then

$$\text{KL}(q(z))||p(z)) = \frac{1}{2} \left(\ln \left(\frac{\sigma_p^2}{\sigma_q^2} \right) + \frac{\sigma_q^2}{\sigma_p^2} + \frac{1}{\sigma_p^2} (\mu_q - \mu_p)^2 - 1 \right). \quad (781)$$

KL divergence for Gamma distributions

Let

$$G(z; a, b) := \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \quad (782)$$

denote the Gamma distribution PDF in its shape and rate parameterization and let

$$q(z) := G(z; a_q, b_q) \text{ and } p(z) := G(z; a_p, b_p) \quad (783)$$

Then

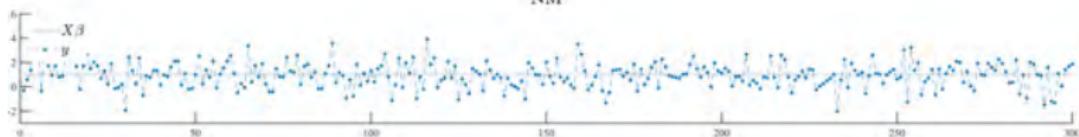
$$\text{KL}(q(z))||p(z)) = a_p \ln \left(\frac{b_q}{b_p} \right) - \ln \left(\frac{\Gamma(a_q)}{\Gamma(a_p)} \right) + (a_q - a_p)\psi(a_q) - (b_q - b_p) \frac{a_q}{b_q}, \quad (784)$$

where ψ denotes the digamma function.

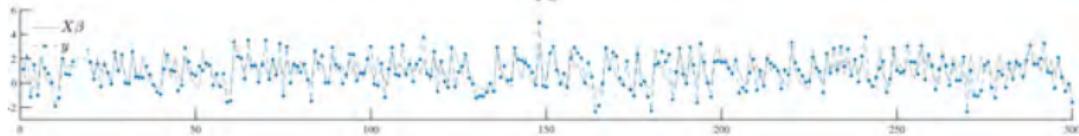
Exercises

Programming Exercise 2 | Model formulation and sampling

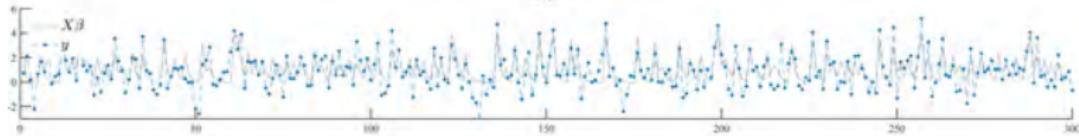
NM



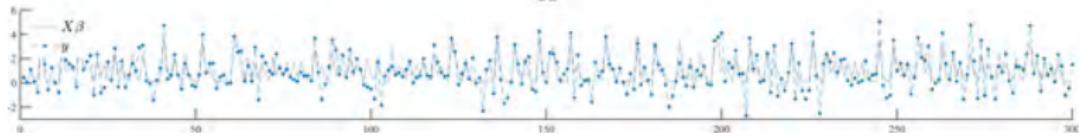
PS



BS



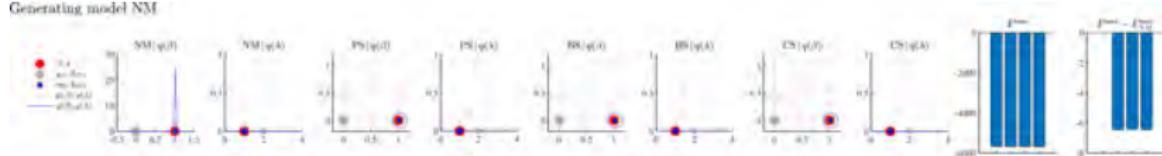
CS



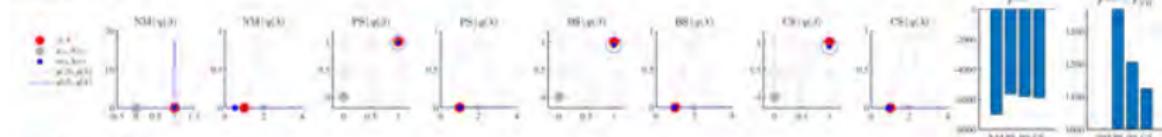
Exercises

Programming Exercise 2 | Model evaluation

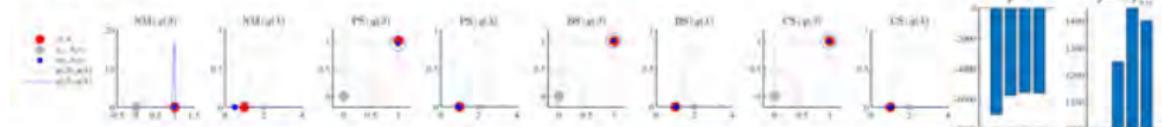
Generating model NM



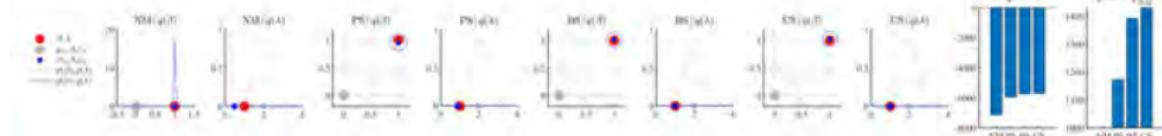
Generating model PS



Generating model BS



Generating model CS



References

- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, N.J, 3rd ed edition.
- Andrieu, C. (2003). An Introduction to MCMC for Machine Learning. page 39.
- Billingsley, P. (1995). *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 3rd ed edition.
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877.
- Brunner, E., Bathke, A. C., and Konietzschke, F. (2018). *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS*. Springer Series in Statistics. Springer International Publishing, Cham.
- Casella, G. and Berger, R. (2012). *Statistical Inference*. Duxbury.
- Czado, C. and Schmidt, T. (2011). *Mathematische Statistik. Statistik und ihre Anwendungen*. Springer, Berlin.
- DeGroot, M. H. and Schervish, M. J. (2012). *Probability and Statistics*. Addison-Wesley, Boston, 4th ed edition.
- Fristedt, B. E., Gray, L. F., and Birkhäuser Publishing Ltd (1998). *A Modern Approach to Probability Theory*.
- Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., and Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234.

-
- Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Hoffman, M. D. (2013). Stochastic Variational Inference. page 45.
- Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.
- Lange, T. and Mosler, K. (2017). *Statistik kompakt*. Springer-Lehrbuch. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Wiley Series in Probability and Statistics.
- Mann, H. B. and Wald, A. (1943). On Stochastic Limit and Order Relationships. *The Annals of Mathematical Statistics*, 14(3):217–226.
- Moeschlin, O. (2000a). *Angewandte Statistik*.
- Moeschlin, O. (2000b). *Wahrscheinlichkeitstheorie I*.
- Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, 2nd ed edition.
- Osborne, M. R. (1992). Fisher's Method of Scoring. *International Statistical Review / Revue Internationale de Statistique*, 60(1):99.
- Ostwald, D., Kirilina, E., Starke, L., and Blankenburg, F. (2014). A tutorial on variational Bayes for latent linear stochastic time-series models. *Journal of Mathematical Psychology*, 60:1–19.

- Ostwald, D., Schneider, S., Bruckner, R., and Horvath, L. (2019). Power, positive predictive value, and sample size calculations for random field theory-based fMRI inference. *BioRxiv*: doi.org/10.1101/613331.
- Ostwald, D. and Starke, L. (2016). Probabilistic delay differential equation modeling of event-related potentials. *NeuroImage*, 136:227–257.
- Penny, W., Kiebel, S., and Friston, K. (2003). Variational Bayesian inference for fMRI time series. *NeuroImage*, 19(3):727–741.
- Rosenthal, J. S. (2006). *A First Look at Rigorous Probability Theory*. World Scientific, Singapore ; Hackensack, N.J, 2nd ed edition.
- Slutsky, E. (1925). über stochastische Asymptoten und Grenzwerte. *Metron*, 5(3):3–89.
- Student (1908). The Probable Error of a Mean. *Biometrika*, 6(1):1–25.
- Wang, Y. and Blei, D. M. (2019). Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.
- Wasserman, L. (2004). *All of Statistics*.
- Zabell, S. L. (2008). On Student's 1908 Article "The Probable Error of a Mean". *Journal of the American Statistical Association*, 103(481):1–7.