

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

First lecture, October 14, 2024

Practical matters

- Lectures on Mondays at 10:15-11:45 in A6/032 by Vesa Kaarnioja.
- Exercises on Tuesdays at 10:15-11:45 in A7/031 by Vesa Kaarnioja starting **October 29**.
 - There will be 13 exercise sheets in total. The first exercise sheet will be published October 21.
 - You are allowed to submit your solutions in groups of 3–4 members.
- Weekly exercises will be published on Mondays. Please complete the tasks before the exercise session of the following week.
- The conditions for completing this course are
 - (1) *successfully completing at least 60% of the course's exercises* (active participation + regular attendance), and
 - (2) *successfully passing the course exam*.
- The course exam will be held **February 10, 2025**, starting at 10:00 in room A6/032.
- The make-up exam will be held **February 24, 2025**, starting at 10:00 in room A6/032.
- Grading: pass/fail for the exercises, pass/fail for the exam.

Exercise guidelines

- You can hand in your solutions at the beginning of the corresponding exercise session or submit your solutions online. Late submissions will **not** be considered.
- Please present your calculations clearly and neatly, providing explanation for all steps.
- Ensure that your arguments are coherent and presented in an orderly fashion. Organize your solutions logically, starting from the problem statement and proceeding step-by-step to the solution.
- Typeset or write your solutions in clear handwriting for easy readability.
- Avoid ambiguity in your solutions: consider the perspective or the reader and ensure that your solutions are understandable from their point of view (i.e., the reader should not have to guess what you have written).
- Use appropriate mathematical notation and terminology.
- Double-check your solutions for errors and correctness before submission. Aim for precision and accuracy in your mathematical expressions and calculations.
- In programming tasks, ensure that your program executes successfully. Include the source code as well as the output of the program as part of your submission.

Course contents

This course will consist of three main parts:

- Probability foundations
 - probability spaces, random variables, distribution of a random variable, expectation and covariance, main limit theorems and inequalities
- Frequentist inference
 - point estimators, confidence intervals, hypothesis testing
- Bayesian inference
 - conjugate inference, numerical models, data assimilation

Introduction

Statistics have been used to organize and interpret data for centuries. In modern statistics, we use various statistical methods to make predictions, provide classifications, derive estimations, etc. The problem set-up for these different problems is usually the same: assume that there is some process generating data. Given the observed data, what can we infer about the process that generated the data? How can we control the uncertainty in our results?

Several theorems of probability (the Law of Large Numbers, the Central Limit Theorem, Hoeffding's inequality, . . .) play a key role in statistics.

In **frequentist inference**, probability is interpreted as an approximate empirical mean observed when running some random experiment a large number N of times. Assume that we are measuring a random quantity X , and let x_i , $1 \leq i \leq N$ be the observed results. Then the probability $\mathbb{P}(X \in E)$ of an outcome E for this experiment is approximately the value, when N is very large, of the ratio of the number of experiments with outcome E with the total number N of experiments.

Using probabilistic notations,

$$\mathbb{P}(X \in E) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N 1_E(x_i).$$

The above equality is justified by the Law of Large Numbers (LLN), one of the cornerstones of the theory of probability.

In some cases, the frequentist interpretation of probability is not meaningful. One example is in weather forecast: the probability of the event “it will rain tomorrow” cannot be thought of as the limit of the empirical mean of some experiment repeated several times.

An alternative way to interpret probability is in terms of a (subjective) degree of belief: the higher the probability of an event, the more likely this event is to happen. This interpretation is the basis of **Bayesian inference**.

Probability foundations

We will need to introduce some terminology to talk about outcomes and their probabilities in order to carry out probability calculations consistently. A **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ is a structured framework that allows us to consistently measure uncertainty. The probability space is comprised of the following components:

- **Sample space** Ω : the set of all possible outcomes of an experiment. For example, the possible space of outcomes for a die toss is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **σ -algebra** \mathcal{F} : a collection of **events** (subsets) of Ω . These are the things we care about when it comes to outcomes – like whether the die shows an even number or a number less than 4. An intuitive way of thinking about σ -algebras is that they contain information: the subsets contained in a σ -algebra represent events for which we can decide, after observation, whether they happened or not. Hence, \mathcal{F} represents all the information we can get from an experiment.
- **Probability measure** \mathbb{P} : the probability measure assigns probabilities to each event. For example, like saying that there is a $\frac{1}{6}$ chance of each specific die face showing up, or a $\frac{1}{2}$ chance of an even number appearing.

Sample space

The fundamental object in probability theory is a **nonempty sample space** Ω . This set encodes all possible outcomes of an experiment. For example, when throwing a dice, a natural choice is $\Omega = \{1, \dots, 6\}$.

An **event** is a subset $A \subset \Omega$. An event represents a collection of outcomes of the experiment we are interested in.

Example

We throw a dice. To model this experiment, we choose $\Omega = \{1, \dots, 6\}$ as our sample space. An event is any subset A of $\{1, \dots, 6\}$. For instance, $A = \{1, 3, 5\}$ represents the event that the result of the throw is an odd number.

Given two events $A \subset \Omega$ and $B \subset \Omega$, we may consider their union $A \cup B$ which represents the event that A or B (or both) occur.

Likewise, the intersection $A \cap B$ represents the event that *both A and B* occur simultaneously.

If $A \cap B = \emptyset$, then we say that A and B are **incompatible** (or **mutually exclusive**).

Example

We throw a coin three times. To model this experiment, we consider $\Omega = \{H, T\}^3$, i.e., the set of all vectors with 3 entries, with each entry taking value H or T . Here, H stands for "Heads" and T for "Tails" (choosing, e.g., $\Omega = \{0, 1\}^3$, with 0 and 1 representing Heads and Tails, respectively, would be equally valid). An event is any subset of $\{H, T\}^3$. For instance, we may consider the events

$$A = \{(H, H, T), (H, T, H), (T, H, H)\} \quad (\text{get Heads exactly twice})$$

$$B = \{(H, H, H), (T, T, T)\}. \quad (\text{get 3 times the same result})$$

Note that $A \cap B = \emptyset$, so A and B are incompatible. Consider now the event

$$C = \{\omega \in \Omega \mid \omega_i = H \text{ for some } i = 1, 2, 3\} \quad (\text{get Heads at least once})$$

Then the joint occurrence of B and C is

$$B \cap C = \{(H, H, H)\}. \quad (\text{get Heads 3 times})$$

Probability measure

Given a random experiment and a nonempty sample space Ω encoding all possible outcomes, we wish to assign to each event of Ω a number known as its **probability**. Let $\mathcal{F} \subset \mathcal{P}(\Omega) := \{A \mid A \subset \Omega\}$ denote the collection of all events of Ω . In what follows, we always assume that \mathcal{F} is a **σ -algebra**:

- $\emptyset \in \mathcal{F}$;
- If $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$;
- If $\{A_n\}_{n \geq 1}$ is a countable sequence with $A_n \in \mathcal{F}$ for all $n \geq 1$, then $\bigcup_{n \geq 1} A_n \in \mathcal{F}$.

A mapping $\mathbb{P}: \mathcal{F} \rightarrow [0, 1]$ is a **probability measure**, if

- ① $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}$;
- ② $\mathbb{P}(\Omega) = 1$;
- ③ (σ -additivity) If $\{A_n\}_{n \geq 1}$ is a countable collection of events that are pairwise disjoint, i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$, then there holds

$$\mathbb{P}\left(\bigcup_{n \geq 1} A_n\right) = \sum_{n \geq 1} \mathbb{P}(A_n).$$

The tuple $(\Omega, \mathbb{P}) = (\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

Definition

If Ω is a finite, non-empty set, the **uniform probability measure** \mathbb{P} on Ω is the probability measure defined by

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \quad \text{for all events } A \subset \Omega.$$

Here, $|\cdot|$ denotes the cardinality (i.e., the number of elements) of a set.

The uniform probability measure is often used to model random experiments where the different possible outcomes happen equally often, or are deemed equally likely to happen.

Example

We throw a fair die. As outcome space we set $\Omega = \{1, \dots, 6\}$, and since the die is fair it is reasonable to consider the uniform probability measure \mathbb{P} on it. With this probability space, for all $i = 1, \dots, 6$, the event “the outcome is i ” is represented by the event $\{i\}$ and its probability is $\mathbb{P}(\{i\}) = \frac{1}{6}$. This probability does not depend on i . As an example of an event, consider $A = \{1, 3, 5\}$, which represents the event that the result of the throw is an odd number. This event has probability

$$\mathbb{P}(A) = \frac{3}{6} = \frac{1}{2}.$$

Example

Consider rolling two fair dice. The corresponding sample space is $\Omega = \{(1, 1), (1, 2), \dots, (6, 6)\}$ endowed with the uniform probability measure \mathbb{P} .

The event “both dice > 2 ”, is

$$A = \{\omega = (\omega_1, \omega_2) \in \Omega \mid \omega_1 > 2 \text{ and } \omega_2 > 2\}.$$

In this example, $\mathbb{P}(\{\omega\}) = \frac{1}{36}$ for all $\omega \in \Omega$ and $\mathbb{P}(A) = \frac{4}{9}$.

Example

We throw a fair coin three consecutive times. As outcome space, we set $\Omega = \{H, T\}^3$, interpreting H as heads and T as tails. For instance, the element $\omega = (H, H, T)$ represents the outcome “Heads, Heads, Tails”. Since the coin is fair, it is reasonable to consider the uniform probability measure \mathbb{P} on Ω . Under this measure, the event $A = \{(H, H, H), (T, T, T)\}$, which represents the event that three tosses yield the same outcome, has probability

$$\frac{|A|}{|\Omega|} = \frac{2}{2^3} = \frac{1}{4}.$$

Corollaries

Proposition

Let (Ω, \mathbb{P}) be a probability space.

- ① $\mathbb{P}(\emptyset) = 0$.
- ② If A, B are two events and $A \subset B$, then $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$.
- ③ If A, B are two events and $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- ④ For any event A , we have $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$, where $A^C := \Omega \setminus A$.
- ⑤ For any two events A and B (not necessarily disjoint), we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

- ⑥ For any countable sequence of events $\{A_n\}_{n \geq 1}$, not necessarily pairwise disjoint, we have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

Proof.

1. Let $A_n = \emptyset$, $n \geq 1$. Clearly these sets satisfy $A_i \cap A_j = \emptyset$ whenever $i \neq j$, and by σ -additivity

$$\underbrace{\mathbb{P}\left(\bigcup_{n \geq 1} \emptyset\right)}_{=\mathbb{P}(\emptyset)} = \sum_{n \geq 1} \mathbb{P}(\emptyset) \quad \Rightarrow \quad \sum_{n \geq 2} \underbrace{\mathbb{P}(\emptyset)}_{\geq 0} = 0 \quad \Rightarrow \quad \mathbb{P}(\emptyset) = 0.$$

2. Since $A \subset B$ and $B = A \cup (B \setminus A)$, where $A \cap (B \setminus A) = \emptyset$, we obtain by σ -additivity

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \quad \Rightarrow \quad \mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A).$$

3. Since $A \subset B$, by part 2 we get $\mathbb{P}(A) = \mathbb{P}(B) - \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0} \leq \mathbb{P}(B)$.
4. Apply part 2 with $B = \Omega$ to get $\mathbb{P}(\Omega \setminus A) = \mathbb{P}(\Omega) - \mathbb{P}(A) = 1 - \mathbb{P}(A)$.

5. Define $E_1 = A \cap B^C$, $E_2 = A \cap B$, and $E_3 = A^C \cap B$. These are pairwise disjoint with $E_1 \cup E_2 \cup E_3 = A \cup B$. Moreover, $A = E_1 \cup E_2$ and $B = E_2 \cup E_3$. Hence

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) \\ \Rightarrow \mathbb{P}(A \cup B) + \mathbb{P}(E_2) &= (\mathbb{P}(E_1) + \mathbb{P}(E_2)) + (\mathbb{P}(E_2) + \mathbb{P}(E_3)) \\ &= \mathbb{P}(A) + \mathbb{P}(B).\end{aligned}$$

Recalling that $\mathbb{P}(E_2) = \mathbb{P}(A \cap B)$ yields the assertion.

6. We define $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup \dots \cup A_{n-1})$ for $n > 1$. Then the B_n are pairwise disjoint and $\bigcup_n B_n = \bigcup_n A_n$. Therefore

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n).$$

Since $B_n \subset A_n$ for all n , we have $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$ by part 3, and the claim follows. □

Definition (Conditional probability)

Let A and B be two events. We assume that $\mathbb{P}(B) > 0$. The **conditional probability** of A given B is

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

The probability $\mathbb{P}(A|B)$ is the probability of A under the assumption that B has already occurred.

Remarks.

- Given an event B such that $\mathbb{P}(B) > 0$, the map $A \mapsto \mathbb{P}(A|B)$ defines a probability measure on Ω . That probability measure is supported on B , i.e., $\mathbb{P}(B|B) = 1$.
- The quantities $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ are **not** the same!

Example

Consider again rolling two fair dice, where the corresponding sample space is $\Omega = \{(1, 1), (1, 2), \dots, (6, 6)\}$ endowed with the uniform probability measure \mathbb{P} .

The probability of getting 3 (event A) when rolling the first dice, given that the other dice gave 4 (event B):

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\frac{1}{36}}{6 \cdot \frac{1}{36}} = \frac{1}{6}.$$

Independence of events

Let (Ω, \mathbb{P}) be a probability space.

Definition

Two events A and B are said to be **independent** if

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

This notion can be expressed in terms of conditional probability.

Lemma

Assume $\mathbb{P}(B) > 0$. Then

- ① $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$.
- ② the events A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Proof. The first point follows from the definition of conditional probability
 $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$. For the second point, note that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \stackrel{\text{divide by } \mathbb{P}(B)}{\Leftrightarrow} \mathbb{P}(A|B) = \mathbb{P}(A). \quad \square$$

Remark. The independence of A and B means that the *a priori* knowledge that B occurs does not change the probability that A occurs.

Example

We throw a fair coin twice. To model this experiment, we consider the probability space (Ω, \mathbb{P}) , where $\Omega = \{H, T\}^2$ and \mathbb{P} is the uniform probability measure on Ω . Let

$$A = \{(H, H), (H, T)\} \quad (1^{\text{st}} \text{ toss gives Heads})$$
$$B = \{(H, H), (T, H)\}. \quad (2^{\text{nd}} \text{ toss gives Heads})$$

Then

$$\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2},$$

while

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(H, H)\}) = \frac{1}{4}.$$

Thus $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ so A and B are independent.

Theorem (Law of total probability)

Let A_1, \dots, A_k be events that form a partition of Ω , i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$ and $\Omega = \bigcup_{i=1}^k A_i$. Then, for any event B , there holds

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

Proof. We have

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^k A_i \right) = \bigcup_{i=1}^k (B \cap A_i),$$

where we used the fact that $\Omega = \bigcup_{i=1}^k A_i$ in the last equality. Since the events A_i are pairwise disjoint, so are the events $B \cap A_i$, and we obtain by σ -additivity that

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

The claim follows by noting that $\mathbb{P}(B \cap A_i) = \mathbb{P}(B|A_i)\mathbb{P}(A_i)$. □

Theorem (Bayes' theorem)

Let A and B be events and assume that $\mathbb{P}(B) > 0$. Then

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

Proof. By definition,

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

On the other hand,

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A) \quad \text{if } \mathbb{P}(A) > 0,$$

which yields the assertion. □



Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Second lecture, October 21, 2024

Random variables

Random variables

Let (Ω, \mathbb{P}) be a probability space and let E be a set.

Definition

A **random variable (RV)** X with values in E is a function $X: \Omega \rightarrow E$.

Remark. The set E is called the **outcome** or **target space**.

- When $E \subset \mathbb{R}$, we say that X is a **real-valued random variable**.
- When $E \subset \mathbb{R}^n$, $n \geq 2$, we call X a **vector-valued random variable**.
- When E is countable, we call X a **discrete random variable**.

In practice, ω is usually not observed directly and analysis is based on the observed random variable $X(\omega)$. Physically, one can think of a realization $X(\omega)$ of a random variable for some $\omega \in \Omega$ as some measurement, or observation performed on a system.

Statistical analysis is based on the *pushforward measure* $B \mapsto \mathbb{P}(X^{-1}(B))$, also called the *probability distribution* or *law* of X , not on \mathbb{P} . Note that here $X^{-1}(B) := \{\omega \in \Omega \mid X(\omega) \in B\}$ is the preimage of B under the mapping X .

Example, two dice

As an example of a random variable, consider the sum:

$$X: \{(1, 1), (1, 2), \dots, (6, 6)\} \rightarrow \{2, \dots, 12\}, \quad X(\omega) = \omega_1 + \omega_2.$$

The identity function $Y(\omega_1, \omega_2) = (\omega_1, \omega_2)$ also defines a random variable. Since $Y: \Omega \rightarrow \mathbb{R}^2$, this random variable is vector-valued.

Let (Ω, \mathbb{P}) be a probability space and E a set. A random variable $X: \Omega \rightarrow E$ induces a probability measure P_X on E , defined by

$$P_X(B) := \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) \quad \text{for } B \subset E,$$

which is called the **probability distribution** (or **law**) of X .

In other words, a random variable X connects an event $B \subset E$ with a corresponding event $X^{-1}(B) \subset \Omega$ and assigns the probability of $X^{-1}(B)$ to B .

Often, we shall simply denote

$$\{X \in B\} := \{\omega \in \Omega \mid X(\omega) \in B\},$$

and write

$$P_X(B) = \mathbb{P}(X \in B).$$

Two random variables X and Y with the same target space E are said to be **equal in law** if they have the same probability function, i.e.,

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \quad \text{for all subsets } B \subset E.$$

Usually, we are ultimately interested in the laws of random variables, rather than the random variables *per se*.

Example

Two players play Heads and Tails on a fair coin. The coin is thrown 10 times, the gain of player 1 is the total number of Heads, while the gain of player 2 is the total number of Tails. This situation is modeled by introducing $\Omega = \{H, T\}^{10}$ endowed with the uniform distribution, and defining random variables X and Y by

$$X(\omega) = \#\{i = 1, \dots, 10 \mid \omega_i = H\}, \quad Y(\omega) = \#\{i = 1, \dots, 10 \mid \omega_i = T\}$$

for all $\omega \in \{H, T\}^{10}$. Then $X + Y = 10$. Clearly X and Y are not equal, however they have equal distribution: for all k ,

$$\mathbb{P}(X = k) = \frac{1}{2^{10}} \binom{10}{k} = \frac{1}{2^{10}} \binom{10}{10 - k} = \mathbb{P}(X = 10 - k) = \mathbb{P}(Y = k).$$

This implies that X and Y are equal in distribution.

Probability mass function

Let (Ω, \mathbb{P}) be a probability space. Let $X: \Omega \rightarrow E$ be a discrete random variable (recall that this means that E is countable). Then, for all $B \subset E$, we can write

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x), \quad (1)$$

where $p_X(x) := \mathbb{P}(X = x)$, $x \in E$. We call p_X the **probability mass function (PMF)** of X .

Properties. The PMF p_X of a discrete random variable X is

- non-negative $p_X(x) \geq 0$ for all $x \in E$;
- normalized $\sum_{x \in E} p_X(x) = 1$.

In consequence, $0 \leq p_X(x) \leq 1$ for all $x \in E$.

- The law of a discrete random variable X with countable target space E is uniquely determined by its PMF. This is a consequence of the fact that, by (1),

$$P_X(B) := \mathbb{P}(X \in B) = \sum_{x \in B} p_X(x),$$

meaning that the PMF *determines* the law of X completely.

Probability density function

Definition

A function $f: \mathbb{R} \rightarrow \mathbb{R}$ is called a **probability density function (PDF)** if the following conditions hold:

- $f(x) \geq 0$ for all $x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} f(x) dx = 1$.

A real-valued random variable X is said to be a **continuous random variable** if there exists a PDF $f_X: \mathbb{R} \rightarrow \mathbb{R}$ such that, for all $a \leq b$, there holds

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx. \quad (2)$$

Then we call f_X the **probability density function (PDF)** of X .

Equation (2) implies for any (measurable) subset $A \subset \mathbb{R}$ that

$$P_X(A) := \mathbb{P}(X \in A) = \int_A f_X(x) dx,$$

meaning that the PDF f_X determines the law of X completely.

Remark. One may think of the PDF as a “continuous” version of the PMF. However, the PMF and PDF are two quite different types of functions.

- The PMF of a *discrete random variable* X can take values between $[0, 1]$, i.e.,

$$\mathbb{P}(X = x) = p_X(x) \in [0, 1].$$

- For a *continuous random variable* X , there *always* holds

$$\mathbb{P}(X = x) = \int_x^x f_X(y) dy = 0.$$

Examples of discrete random variables

Example

Let $p \in (0, 1)$. Let X be a random variable with values in $E = \{0, 1\}$ and with PMF given by

$$p_X(x) = \begin{cases} 1 - p & \text{if } x = 0, \\ p & \text{if } x = 1. \end{cases}$$

Then we say that X is a **Bernoulli random variable** with parameter p , and we write

$$X \sim \text{Ber}(p).$$

A Bernoulli random variable with parameter p represents the result of throwing a coin that falls on Heads with probability p and Tails with probability $1 - p$ ($p = 1/2$ if the coin is fair).

Example

Let $p \in (0, 1)$ and $n \geq 1$ an integer. Let X be a random variable with values in $\{0, \dots, n\}$ and with PMF given by

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in \{0, \dots, n\}.$$

Then we say that X is a **binomial random variable** with parameters n and p , and we write

$$X \sim \text{Bin}(n, p).$$

This corresponds to the probability of the number of times a coin lands on Heads in n tosses of a coin, with p denoting the probability of a coin landing on Heads.

Example

Let $p \in (0, 1)$. Let X be a random variable with values in \mathbb{N} and with PMF given by

$$p_X(x) = (1 - p)^{x-1} p, \quad x \geq 1.$$

Then we say that X is a **geometric random variable** with parameter p , and we write

$$X \sim \text{Geo}(p).$$

This corresponds with the probability of hitting Heads for the first time, when the probability of hitting Heads is equal to p .

That is,

$$\mathbb{P}(X = k) = p_X(k) = (1 - p)^{k-1} p$$

denotes the probability of hitting Tails for the first $k - 1$ rounds and hitting heads on the k^{th} round.

Example

Let $\lambda > 0$. Let X be a random variable with values in \mathbb{N}_0 and with PMF given by

$$p_X(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x \geq 0.$$

We then say that X is a **Poisson random variable** with parameter λ , and we write

$$X \sim \text{Poisson}(\lambda).$$

Poisson random variables can be used to model the count of rare events such as nuclei decaying in a radioactive sample.

Examples of continuous real-valued random variables

Definition

Let $a < b$. Let X be a real-valued continuous random variable with PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise,} \end{cases} \quad x \in \mathbb{R}.$$

We then say that X is a **uniform random variable** in $[a, b]$, and we write

$$X \sim \mathcal{U}(a, b).$$

Definition

Let $\lambda > 0$. Let X be a real-valued continuous random variable with PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \quad x \in \mathbb{R}.$$

We then say that X is an **exponential random variable** with parameter λ , and we write

$$X \sim \text{Exp}(\lambda).$$

Example

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. Let X be a real-valued continuous random variable with PDF given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

We then say that X is a **Gaussian random variable** with parameters μ and σ^2 , and we write

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

The parameter μ is called the mean and σ is called the standard deviation of X .

Cumulative distribution function

The **cumulative distribution function (CDF)** of a real-valued random variable X is the function $F_X : \mathbb{R} \rightarrow [0, 1]$ given by

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \leq x\}) . \quad (\text{or shortly } = \mathbb{P}(X \leq x))$$

Note that the CDF is defined for any random variable taking values in \mathbb{R} , whether discrete or continuous.

Proposition

Let $F_X : \mathbb{R} \rightarrow [0, 1]$ be the CDF of a real-valued random variable X . Then

- F_X is non-decreasing: if $a \leq b$, then $F_X(a) \leq F_X(b)$.
- F_X is right-continuous: for all $a \in \mathbb{R}$,

$$F_X(a) = \lim_{x \rightarrow a+} F_X(x).$$

- $F_X(-\infty) := \lim_{x \rightarrow -\infty} F_X(x) = 0$ and $F_X(\infty) := \lim_{x \rightarrow \infty} F_X(x) = 1$.

One can read off relevant information on the distribution of X from its CDF.

Lemma

Let $F_X : \mathbb{R} \rightarrow [0, 1]$ be the CDF of a real-valued random variable X . Then

- For any real numbers $a < b$,

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

- For any $a \in \mathbb{R}$,

$$\mathbb{P}(X > a) = 1 - F_X(a).$$

- For any $x \in \mathbb{R}$,

$$\mathbb{P}(X = x) = F_X(x) - \lim_{y \rightarrow x^-} F_X(y).$$

Remark. In particular, if X is a continuous random variable, we have $F_X(x) = \lim_{y \rightarrow x^-} F_X(y)$ for all $x \in \mathbb{R}$; no jumps occur. For a discrete random variable, the situation is different: F_X is then a pure-jump function, meaning that it increases purely through jumps.

Relationship between the CDF and PMF (discrete case)

Proposition

Let X be a discrete random variable taking values in a countable subset E of \mathbb{R} . Denoting the PMF of X by p_X and its CDF by F_X , we have

$$F_X(a) = \sum_{\substack{x \in E \\ x \leq a}} p_X(x) \quad \text{for all } a \in \mathbb{R},$$

$$p_X(x) = F_X(x) - \lim_{y \rightarrow x^-} F_X(y).$$

Proof. By the definition of the PMF, there holds

$$\mathbb{P}(X \in B) = \sum_{x \in B} p_X(x) \quad \text{for all subsets } B \subset E.$$

Setting $B = \{x \in E \mid x \leq a\}$ yields the first relation.

For the second relation, we note that

$$\{X = x\} = \bigcap_{n \geq 1} E_n,$$

where the sets $E_n := \left\{X \in \left(x - \frac{1}{n}, x\right]\right\}$ form a decreasing sequence of events $E_{n+1} \subset E_n$ for $n \geq 1$. In this case, there holds

$$\begin{aligned}\mathbb{P}\left(\bigcap_{n \geq 1} E_n\right) &= \lim_{n \rightarrow \infty} \mathbb{P}(E_n) \\ &= \lim_{n \rightarrow \infty} (F_X(x) - F_X(x - \frac{1}{n})) \\ &= F_X(x) - \lim_{y \rightarrow x-} F_X(y),\end{aligned}$$

as desired. □

Relationship between the CDF and PDF (continuous case)

Proposition

Let X be a continuous real-valued random variable. Denoting the PDF of X by f_X , and its CDF by F_X , we have

$$F_X(a) = \int_{-\infty}^a f_X(y) dy \quad \text{for all } a \in \mathbb{R}.$$

In addition, if F_X is differentiable at $x \in E$, we have

$$f_X(x) = F'_X(x).$$

Proof. For the first statement, note that for all $u < a$ there holds

$$F_X(a) - F_X(u) = \mathbb{P}(X \in (u, a]) = \mathbb{P}(X \in [u, a]) = \int_u^a f_X(y) dy,$$

where we used the fact that $\mathbb{P}(X = u) = 0$ since X is a continuous random variable. Letting $u \rightarrow -\infty$ and recalling $F_X(-\infty) = 0$, we obtain $F_X(a) = \int_{-\infty}^a f_X(y) dy$. The second statement follows from the fundamental theorem of calculus (F_X is the antiderivative of f_X). □



Proposition

The probability distribution of a real-valued random variable is uniquely determined by its CDF.

Proof. We give a proof in the discrete case. Let X and Y be two real-valued random variables with the same CDF:

$$F_X(x) = F_Y(x) \quad \text{for all } x \in \mathbb{R}.$$

Then by the previous discussion,

$$p_X(x) = F_X(x) - \lim_{y \rightarrow x^-} F_X(y) = F_Y(x) - \lim_{y \rightarrow x^-} F_Y(y) = p_Y(x).$$

Thus X and Y have the same PMF, meaning that X and Y are equal in law. □

Quantile function

Definition

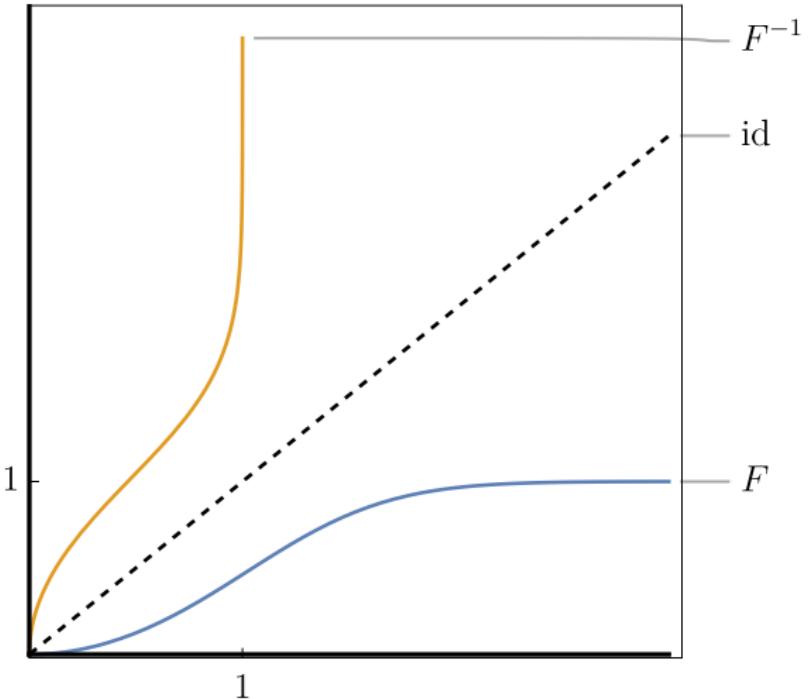
Let X be a real-valued random variable with CDF F . The generalized inverse $F^{-1}: (0, 1) \rightarrow \mathbb{R}$,

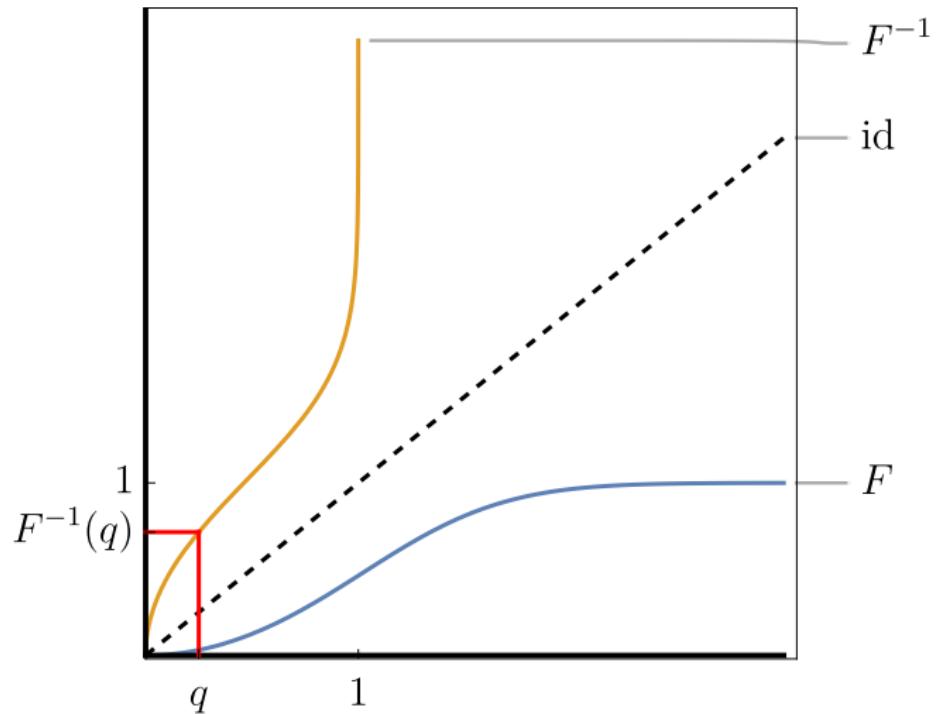
$$F^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \geq q\}, \quad q \in (0, 1),$$

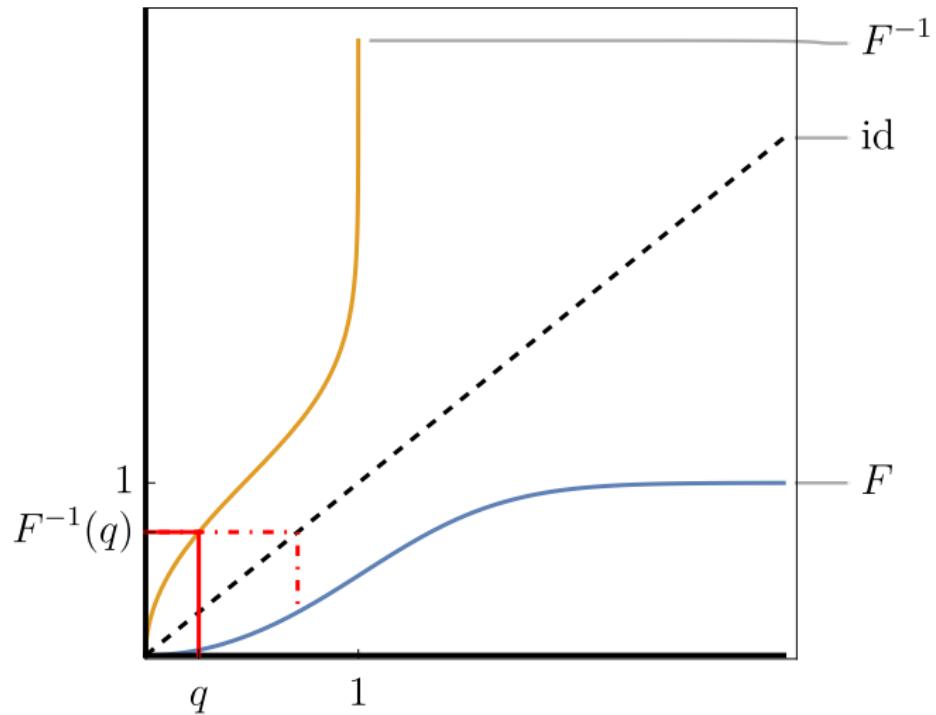
is called the **quantile function** of X .

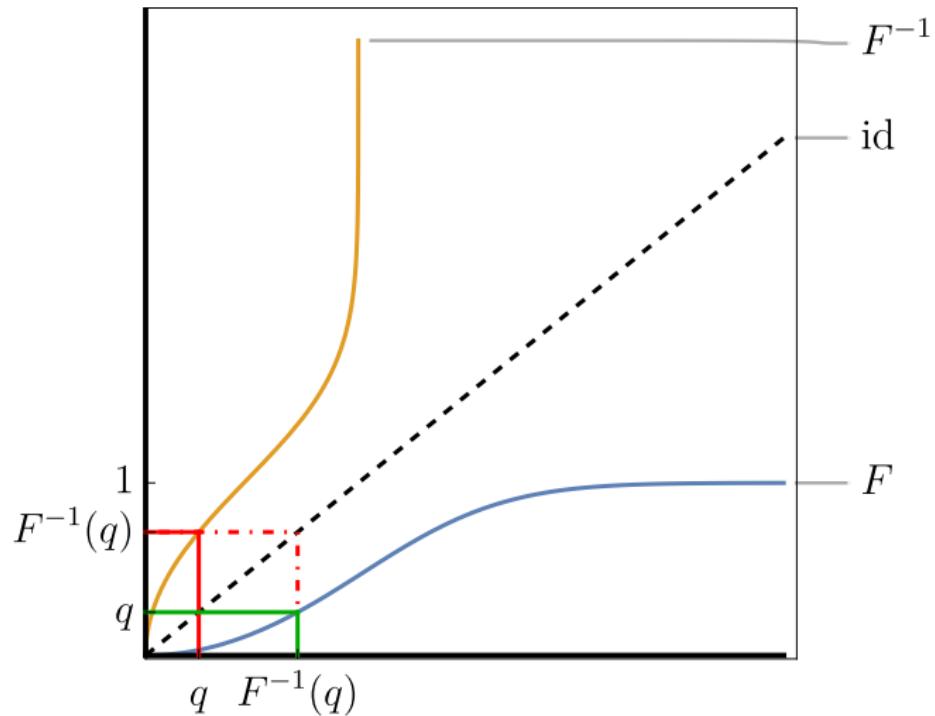
- If F is strictly increasing, then the quantile function is the inverse function of F .
- For example, the CDF and inverse CDF of a Bernoulli random variable $X \sim \text{Ber}(\frac{1}{2})$ are

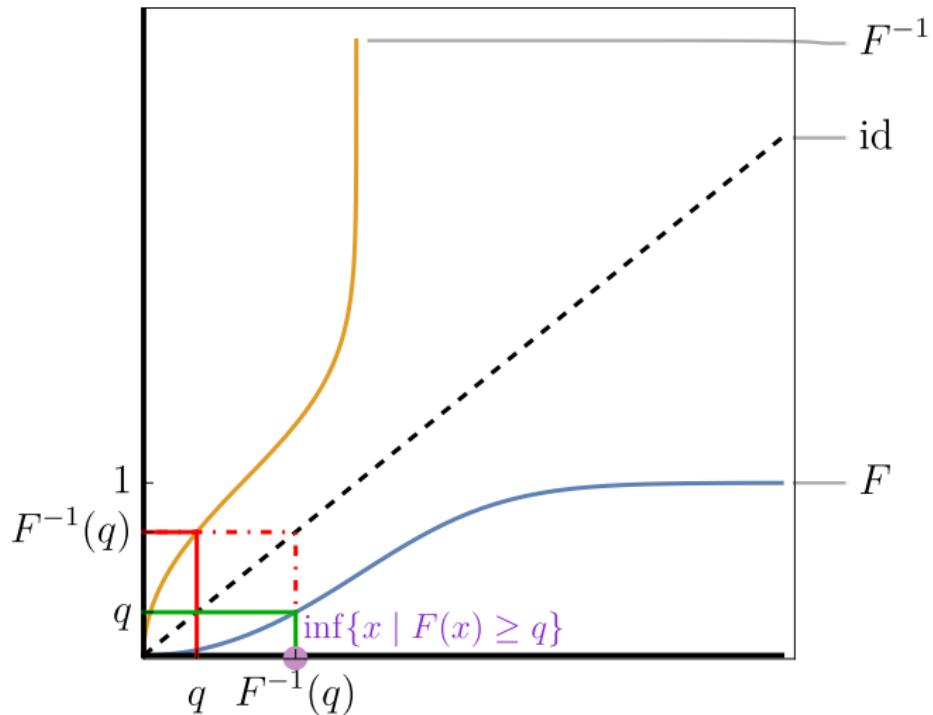
$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases} \quad \text{and} \quad F^{-1}(q) = \begin{cases} 0 & \text{if } 0 < q \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} < q < 1. \end{cases}$$











“Find the smallest value of x such that $F(x) \geq q$.”

Proposition

Let X be a real-valued random variable with CDF F_X . Then

- ① For all $q \in (0, 1)$, $F_X(F_X^{-1}(q)) \geq q$.
- ② If X is a continuous random variable, then $F_X(F_X^{-1}(q)) = q$ for all $q \in (0, 1)$.

Proof. (1) Let $q \in (0, 1)$. Since $F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F(x) \geq q\}$ by definition, we can find a sequence $(a_n)_{n \geq 1}$ of real numbers such that $F_X(a_n) \geq q$ and $a_n \searrow F_X^{-1}(q)$. By the right-continuity of F_X , there holds

$$F_X(F_X^{-1}(q)) = \lim_{n \rightarrow \infty} F_X(a_n) \geq q.$$

(2) It suffices to prove the inequality $F_X(F_X^{-1}(q)) \leq q$ by (1). Assume to the contrary that $F_X(F_X^{-1}(q)) > q$. Since F_X is the CDF of a continuous random variable, it is continuous. By continuity of F_X , there exists $a \in (-\infty, F_X^{-1}(q))$ such that $F_X(a) > q$, which contradicts the definition of F_X^{-1} . □

CDF of a normal random variable

Example

The CDF of a normal random variable $X \sim \mathcal{N}(0, 1)$ is often denoted by Φ ,

$$\Phi(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt, \quad x \in \mathbb{R}.$$

Typical values to remember:

$$\Phi(1.645) = \mathbb{P}(X \leq 1.645) \approx 0.95,$$

$$\Phi(1.960) = \mathbb{P}(X \leq 1.960) \approx 0.975.$$

In this case the CDF Φ is injective and the quantile function, denoted by Φ^{-1} , coincides with its inverse. The above equalities can be recast as

$$\Phi^{-1}(0.95) \approx 1.645,$$

$$\Phi^{-1}(0.975) \approx 1.960.$$

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Third lecture, October 28, 2024

Joint distributions

Often, instead of dealing with one random variable only, we are interested in several random variables X_1, \dots, X_n .

Let (Ω, \mathbb{P}) be a probability space and let $X_j: \Omega \rightarrow E_j$ be random variables with target spaces E_j , $j = 1, \dots, n$. One can view the map

$$X := (X_1, \dots, X_n): \Omega \rightarrow E_1 \times \dots \times E_n, \quad \omega \mapsto (X_1(\omega), \dots, X_n(\omega))$$

as a single, multivariate random variable.

In analogy to the univariate case, the **joint probability distribution** of X_1, \dots, X_n is

$$P_{X_1, \dots, X_n}(C) = \mathbb{P}((X_1, \dots, X_n) \in C) \quad \text{for } C \subset E_1 \times \dots \times E_n.$$

Informally speaking, the **marginal distribution** of X_i is obtained by “integrating out” (continuous RVs) / “summation over” (discrete RVs) all variables except the i^{th} one. The precise definition is

$$\begin{aligned} P_{X_i}(A) &= P_{X_1, \dots, X_n}(E_1 \times \dots \times E_{i-1} \times A \times E_{i+1} \times \dots \times E_n) \\ &= \mathbb{P}(X_1 \in E_1, \dots, X_{i-1} \in E_{i-1}, X_i \in A, X_{i+1} \in E_{i+1}, \dots, X_n \in E_n) \end{aligned}$$

for all events $A \subset E_i$.

Joint PMF (discrete RVs)

Assume that $X_j: \Omega \rightarrow E_j$ are discrete random variables (recall that this means each E_j is countable). This means that $E_1 \times \cdots \times E_n$ is also countable. The **joint PMF** $p_{X_1, \dots, X_n}: E_1 \times \cdots \times E_n \rightarrow [0, 1]$ is defined as

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n), \quad (x_1, \dots, x_n) \in E_1 \times \cdots \times E_n.$$

The probability distribution can be expressed as follows in the discrete case.

Proposition

For all events $C \subset E_1 \times \cdots \times E_n$, there holds

$$P_{X_1, \dots, X_n}(C) = \sum_{(x_1, \dots, x_n) \in C} p_{X_1, \dots, X_n}(x_1, \dots, x_n).$$

Proof. The claim is an immediate consequence of σ -additivity of disjoint events

$$\{(X_1, \dots, X_n) \in C\} = \bigcup_{(x_1, \dots, x_n) \in C} \{X_1 = x_1, \dots, X_n = x_n\}. \quad \square$$

The **marginal PMF** of a discrete RV X_i can be obtained from the joint PMF by summation over all the other RVs:

$$p_{X_i}(x) = \sum_{\substack{x_1 \in E_1, \dots, \\ x_{i-1} \in E_{i-1}, \\ x_{i+1} \in E_{i+1}, \dots \\ x_n \in E_n}} p_{X_1, \dots, X_n}(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n).$$

More generally, for any subset of indices $\mathcal{I} \subset \{1, \dots, n\}$, we can recover the joint PMF of the random variables $(X_i)_{i \in \mathcal{I}}$ from the joint PMF of X_1, \dots, X_n by summing up p_{X_1, \dots, X_n} over all possible values in the coordinates $j \notin \mathcal{I}$.

For example, if $n = 4$, we can recover the joint PMF of (X_2, X_3) via

$$p_{X_2, X_3}(x, y) = \sum_{x_1 \in E_1, x_4 \in E_4} p_{X_1, X_2, X_3, X_4}(x_1, x, y, x_4).$$

Example (Bivariate case $n = 2$)

If (X, Y) is a bivariate discrete RV with PMF $p_{X,Y}$, then the PMFs of X and Y are respectively given by

$$p_X(x) = \sum_{y \in E_2} p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \sum_{x \in E_1} p_{X,Y}(x,y).$$

Example

Let (X, Y) be a bivariate RV taking values in $\{1, 2\} \times \{1, 2, 3\}$ and with joint PMF p given as below

$p(x,y)$	$y = 1$	$y = 2$	$y = 3$
$x = 1$	0.1	0.3	0.2
$x = 2$	0.2	0.2	0

The values of the marginal PMF $p_X(x)$, $x = 1, 2$, are obtained by summing up the probabilities in each of the corresponding rows

$$p_X(1) = 0.1 + 0.3 + 0.2 = 0.6$$

$$p_X(2) = 0.2 + 0.2 + 0 = 0.4.$$

Similarly, the values of the marginal PMF $p_Y(y)$, $y = 1, 2, 3$, are obtained by summing up the probabilities in each of the corresponding columns:

$$p_Y(1) = 0.1 + 0.2 = 0.3, \quad p_Y(2) = 0.3 + 0.2 = 0.5, \quad p_Y(3) = 0.2 + 0 = 0.2.$$

Joint PDF (continuous RVs)

Definition

A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is called a **probability density function (PDF)** if the following conditions hold:

- $f(x_1, \dots, x_n) \geq 0$ for all $(x_1, \dots, x_n) \in \mathbb{R}^n$;
- $\int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(x_1, \dots, x_n) dx_1 \cdots dx_n = 1$.

The real-valued random variables X_1, \dots, X_n admit a **continuous joint distribution** (resp. admit a **joint density**) if there exists a PDF $f_{X_1, \dots, X_n}: \mathbb{R}^n \rightarrow \mathbb{R}$ such that, for all subsets $A \subset \mathbb{R}^n$, there holds

$$\mathbb{P}((X_1, \dots, X_n) \in A) = \int_A f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \cdots dx_n.$$

Then we call f_{X_1, \dots, X_n} the **probability density function (PDF)** of X .

Lemma

If X_1, \dots, X_n admit a joint density f_{X_1, \dots, X_n} , then X_1, \dots, X_n are continuous RVs with PDF given by

$$f_{X_i}(x) = \int_{\mathbb{R}^{n-1}} f_{X_1, \dots, X_n}(x_1, \dots, x_{i-1}, x, x_{i+1}, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n$$

for $x \in \mathbb{R}$. We call f_{X_i} the marginal PDF of X_i .

More generally, for any subset of indices $\mathcal{I} \subset \{1, \dots, n\}$ we can recover the joint PDF of the random variables $(X_i)_{i \in \mathcal{I}}$ from the joint PDF of X_1, \dots, X_n by integrating over all possible values in the coordinates $j \notin \mathcal{I}$.

For example, if $n = 4$, we can recover the joint PDF of (X_2, X_3) via

$$f_{X_2, X_3}(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1, X_2, X_3, X_4}(x_1, x, y, x_4) dx_1 dx_4.$$

Example

Let $a, b, c, d \in \mathbb{R}$ be such that $a < b$ and $c < d$. Then the function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(z) = \frac{1}{(b-a)(d-c)} \mathbf{1}_{[a,b] \times [c,d]}(z), \quad z \in \mathbb{R}^2,$$

is a PDF. It corresponds to the **uniform distribution** on the rectangle $[a, b] \times [c, d]$. The marginal distributions are univariate distributions on the $[a, b]$ and $[c, d]$, respectively:

$$X \sim \mathcal{U}(a, b), \quad Y \sim \mathcal{U}(c, d).$$

Example (Bivariate Gaussian distribution)

Let $\mu \in \mathbb{R}^2$ and let $C \in \mathbb{R}^{2 \times 2}$ be a symmetric, positive definite 2×2 matrix. The function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ given by

$$f(z) = \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2}(z - \mu)^T C^{-1}(z - \mu)\right), \quad z \in \mathbb{R}^2,$$

is a PDF. A random vector $Z = (X, Y)$ with PDF f is said to have Gaussian distribution with mean μ and covariance matrix C . Denoting

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \quad C = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix},$$

then the marginal PDFs are given by

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

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(y - \mu_Y)^2}{2\sigma_Y^2}\right).$$

Thus $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.

In the special case $\mu = 0$ and $C = I_2$, i.e., $\mu_X = \mu_Y = 0$, $\sigma_{XY} = 0$, and $\sigma_X^2 = \sigma_Y^2 = 1$:

$$f(z) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\|z\|^2\right), \quad z \in \mathbb{R}^2,$$

where $\|z\| = \sqrt{x^2 + y^2}$ denotes the Euclidean norm of $z = (x, y)$.

Independence of random variables

Definition

The random variables X_1, \dots, X_n are said to be independent if, for any subsets $A_1 \subset E_1, \dots, A_n \subset E_n$, there holds

$$\mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n).$$

Theorem (Independence of discrete RVs)

Assume that X_1, \dots, X_n are discrete random variables with joint PMF p_{X_1, \dots, X_n} and marginal PMFs p_{X_1}, \dots, p_{X_n} . Then X_1, \dots, X_n are independent if and only if

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n), \quad (x_1, \dots, x_n) \in E_1 \times \cdots \times E_n.$$

Theorem (Independence of continuous RVs)

Assume that X_1, \dots, X_n are continuous random variables with joint PDF f_{X_1, \dots, X_n} and marginal PDFs f_{X_1}, \dots, f_{X_n} . Then X_1, \dots, X_n are independent if and only if

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n), \quad (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Example, independence

Let X and Y have the joint PDF

$$f(x, y) = \begin{cases} x + y & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Are the variables X and Y independent?

Now

$$f(x) = \int_0^1 (x + y) dy = x + \frac{1}{2}, \quad 0 < x < 1$$

and

$$f(y) = \int_0^1 (x + y) dx = y + \frac{1}{2}, \quad 0 < y < 1.$$

If the random variables are independent, then $f(x, y) = f(x) \cdot f(y)$. Let $x = 1/3$ and $y = 1/3$. Now

$$f(x, y) = x + y = \frac{1}{3} + \frac{1}{3} = \frac{2}{3},$$

$$f(x) \cdot f(y) = \left(\frac{1}{3} + \frac{1}{2}\right)\left(\frac{1}{3} + \frac{1}{2}\right) = \frac{5}{6} \cdot \frac{5}{6} = \frac{25}{36} \neq \frac{2}{3}.$$

Thus X and Y are not independent.

Example, independence

Let X and Y have the joint PMF

$$p(x, y) = \begin{cases} \frac{1}{4} & \text{if } x \in \{1, 2\}, y \in \{1, 2\}, \\ 0 & \text{otherwise.} \end{cases}$$

Now

$$p(x) = \sum_{y \in \{1, 2\}} p(x, y) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad x \in \{1, 2\},$$

and otherwise $p(x) = 0$,

and

$$p(y) = \sum_{x \in \{1, 2\}} p(x, y) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad y \in \{1, 2\},$$

and otherwise $p(y) = 0$.

Therefore $p(x, y) = p(x)p(y)$ for all x and y , meaning that X and Y are independent.

Conditional distribution

Definition

Let (X, Y) be a discrete random variable in $E_1 \times E_2$ with joint PMF $p_{X,Y}$ and marginal PMFs p_X and p_Y . The **conditional PMF** $p_{X|Y}$ of X given Y is defined by

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

for all $x \in E_1$ and $y \in E_2$ such that $p_Y(y) > 0$.

Definition

Let (X, Y) be a continuous random variable in $\mathbb{R}^n \times \mathbb{R}^k$ with joint PDF $f_{X,Y}$ and marginal PMFs f_X and f_Y . The **conditional PDF** $f_{X|Y}$ of X given Y is defined by

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^k$ such that $f_Y(y) > 0$.

Transformations of random variables

When we perform arithmetic with random variables, it is natural to ask

- if X and Y are random variables, what is the distribution of $Z = X + Y$?
- if X is an \mathbb{R}^k -valued random variable with known distribution and $g: \mathbb{R}^k \rightarrow \mathbb{R}^k$ is a function, what is the distribution of the transformed random variable $Y = g(X)$?

Theorem

Let X be a continuous real-valued random variable with CDF F_X and quantile function F_X^{-1} .

- ① The random variable $U = F_X(X) \sim \mathcal{U}(0, 1)$.
- ② If $U \sim \mathcal{U}(0, 1)$, then $F_X^{-1}(U)$ has the same distribution as X (they are equal in law).

Proof. (1) Note that $\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t))$.[†] We observe that for all $t \in (0, 1)$,

$$\mathbb{P}(U \leq t) = \mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t)) = F_X(F_X^{-1}(t)) = t.$$

Therefore $\mathbb{P}(U \leq t) = t$, meaning that $U \sim \mathcal{U}(0, 1)$.

(2) $\mathbb{P}(F_X^{-1}(U) \leq t) = \mathbb{P}(U \leq F_X(t)) = F_X(t)$. □

[†]If $F_X(X) < t$, then $X < F_X^{-1}(t)$, which implies (since X is a continuous RV) that $\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(F_X(X) < t) \leq \mathbb{P}(X < F_X^{-1}(t)) = \mathbb{P}(X \leq F_X^{-1}(t))$.

On the other hand, $X \leq F_X^{-1}(t)$ implies $F_X(X) \leq F_X(F_X^{-1}(t)) = t$, so

$\mathbb{P}(X \leq F_X^{-1}(t)) \leq \mathbb{P}(F_X(X) \leq t)$. Therefore $\mathbb{P}(F_X(X) \leq t) = \mathbb{P}(X \leq F_X^{-1}(t))$.

The previous theorem is very useful for simulations: if we have a uniform random number generator, we can generate samples from any distribution provided that we have access to its quantile function.

Algorithm (Inverse transform sampling)

1. Draw $U \sim \mathcal{U}(0, 1)$.
2. Calculate $X = F_X^{-1}(U)$.

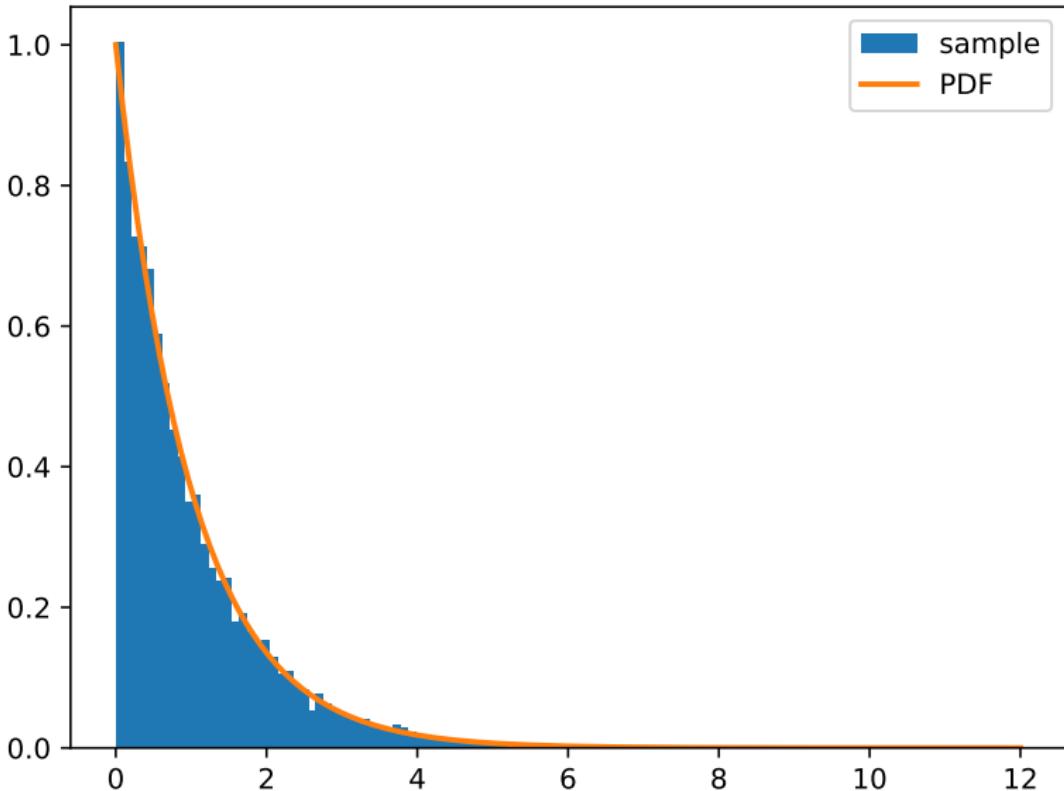
If a closed form expression for the inverse CDF is not available, then a computationally attractive formula for approximating the value $F_X^{-1}(U)$ is given by the generalized inverse:

$$F_X^{-1}(q) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq q\}.$$

Example (Exponential distribution)

Let $X \sim \text{Exp}(\lambda)$, $\lambda > 0$, with the PDF $f_X(x) = \lambda e^{-\lambda x} \mathbf{1}_{[0,\infty)}(x)$. In this case, $F_X(a) = \mathbf{1}_{[0,\infty)}(a)(1 - e^{-\lambda a})$ and $F_X^{-1}(q) = -\frac{1}{\lambda} \log(1 - q)$, $q \in (0, 1)$. We implement inverse transform sampling to draw a sample $X \sim \text{Exp}(1)$.

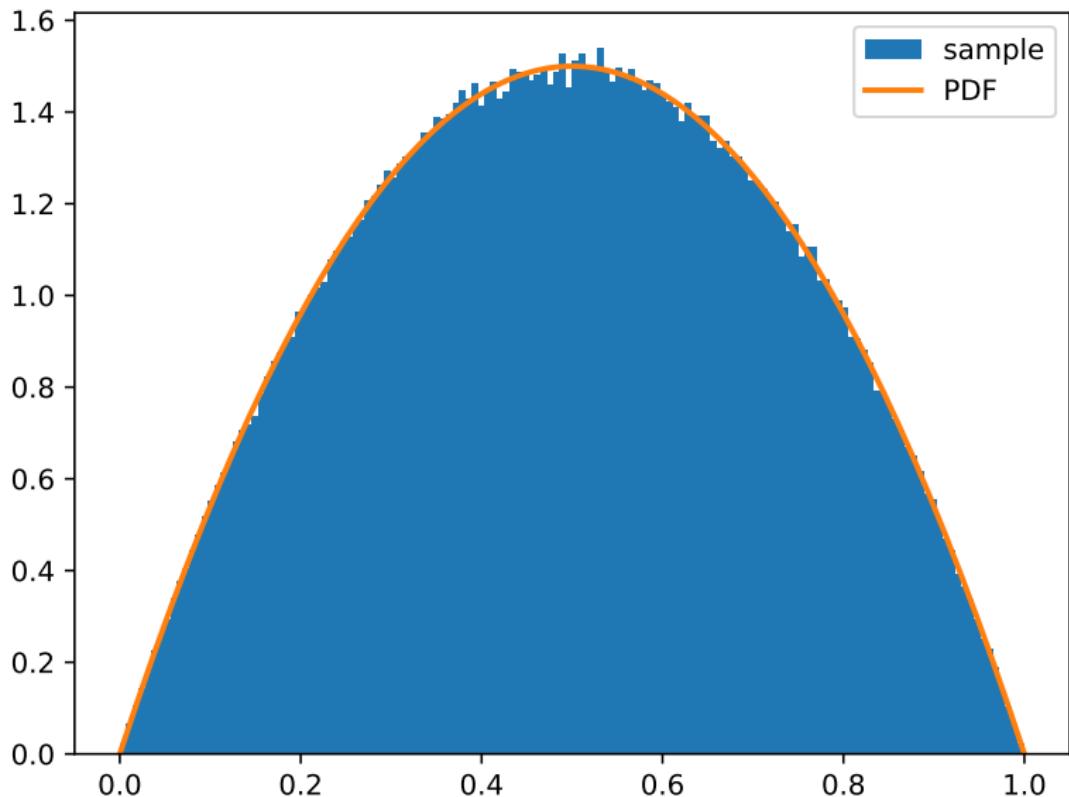
```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e5) # sample size
x = np.linspace(0,12,1000)
lam = 1 # lambda parameter of Exp distribution
p = lambda x: lam * np.exp(-lam*x) # PDF
invF = lambda q: -1/lam * np.log(1-q) # quantile function
u = np.random.uniform(size=n) # i.i.d. sample from U(0,1)
sample = invF(u) # inverse transform
plt.hist(sample,bins='auto',
         density=True,label='sample') # draw histogram
plt.plot(x,p(x),linewidth=2,label='PDF') # plot the PDF
plt.legend()
plt.show()
```



Example

Let the random variable X have the PDF $f_X(x) = (6x - 6x^2)\mathbf{1}_{[0,1]}(x)$. In this case, the quantile function is difficult to write down, but we can still implement inverse transform sampling numerically.

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e6) # sample size
x = np.linspace(0,1,10000)
p = lambda x: 6*x-6*x**2 # PDF
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF of p
sample = []
for _ in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse transform
    sample.append(x[ind]) # store sample
plt.hist(sample,bins='auto',
         density=True,label='sample') # draw histogram
plt.plot(x,p(x),linewidth=2,label='PDF') # plot the PDF
plt.legend(); plt.show()
```



Change of variables formula (discrete RVs)

Proposition

Let $X: \Omega \rightarrow E$ and $Y: \Omega \rightarrow F$ be discrete random variables such that $Y = g(X)$, where $g: E \rightarrow F$. Then the PMF of Y is given by

$$p_Y(y) = \sum_{x \in g^{-1}(\{y\})} p_X(x) = \sum_{\substack{x \in E \\ g(x)=y}} p_X(x).$$

In other words, the PMF of Y at point y is obtained by summing up the PMF of X over the preimage $g^{-1}(\{y\})$.

Proof. Recall that $g^{-1}(\{y\}) = \{x \in E \mid g(x) = y\}$. Thus

$$\begin{aligned} p_Y(y) &= \mathbb{P}(Y = y) = \mathbb{P}(g(X) = y) = \mathbb{P}(X = g^{-1}(\{y\})) \\ &= \mathbb{P}\left(\bigcup_{x \in g^{-1}(\{y\})} \{X = x\}\right) = \sum_{x \in g^{-1}(\{y\})} \mathbb{P}(X = x) = \sum_{x \in g^{-1}(\{y\})} p_X(x), \end{aligned}$$

where we used the σ -additivity of the disjoint sets $(\{X = x\})_{x \in g^{-1}(y)}$. □

Change of variables formula (continuous, univariate case)

Let X and Y be real-valued random variables such that $Y = g(X)$, where $g: \mathbb{R} \rightarrow \mathbb{R}$. By noting that the CDF of Y satisfies

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y),$$

one can use the following method to obtain the PDF of Y given the PDF of X :

- Compute the CDF of Y using

$$F_Y(y) = \mathbb{P}(g(X) \leq y) \quad \text{for } y \in \mathbb{R}.$$

- If F_Y is differentiable, then Y has the PDF $f_Y = F'_Y$.

Example

Let $X \sim \mathcal{U}(0, 1)$, $g(x) = x^2$, and define $Y = g(X)$. We wish to find $f_Y(y)$. We begin by noting that

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X^2 \leq y) = \begin{cases} \mathbb{P}(\emptyset) & \text{if } y < 0, \\ \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) & \text{if } y \geq 0. \end{cases}$$

Here, $\mathbb{P}(\emptyset) = 0$ and

$$\mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \mathbf{1}_{[0,1]}(x) dx = \begin{cases} \sqrt{y} & \text{if } y \in [0, 1], \\ 1 & \text{if } y > 1. \end{cases}$$

Hence

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ \sqrt{y} & \text{if } y \in [0, 1], \\ 1 & \text{if } y > 1 \end{cases} \quad \Rightarrow \quad f_Y(y) = \frac{\mathbf{1}_{[0,1]}(y)}{2\sqrt{y}}, \quad y \in \mathbb{R}.$$

In the special case where $g: \mathbb{R} \rightarrow \mathbb{R}$ is a strictly monotonic, continuously differentiable function, one has the following formula.

Theorem

Let $g: \mathbb{R} \rightarrow \mathbb{R}$ be a continuously differentiable and strictly monotonic function. Let X and Y be continuous, real-valued random variables satisfying $Y = g(X)$. Then we have the following:

$$f_X(x) = f_Y(g(x))|g'(x)|, \quad x \in \mathbb{R},$$

and

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1})'(y)| = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y))|}, \quad y \in \mathbb{R}.$$

Proof. For each (measurable) subset $B \subset \mathbb{R}$, there holds

$$\mathbb{P}(X \in B) = \mathbb{P}(Y \in g(B)) = \int_{g(B)} f_Y(y) dy = \int_B f_Y(g(x))|g'(x)| dx.$$

Since B is arbitrary, we conclude that $f_X(x) = f_Y(g(x))|g'(x)|$.

The second claim follows from the first one by writing $X = g^{-1}(Y)$. □

Change of variables formula (continuous, multivariate case)

The change of variables formulae can be generalized to higher dimensions. For example, let X_1, \dots, X_k be real-valued random variables and let $g: \mathbb{R}^k \rightarrow \mathbb{R}$. We wish to derive the PDF of the real-valued random variable $Z = g(X_1, \dots, X_k)$.

One can proceed as follows:

- ① Compute the CDF F_Z of Z by

$$F_Z(z) = \mathbb{P}(g(X_1, \dots, X_k) \leq z).$$

- ② If F_Z is differentiable, then its PDF is given by $f_Z = F'_Z$.

Example

Let $X, Y \sim \mathcal{U}(0, 1)$ be independent random variables and define $Z = \max(X, Y)$. Now[†]

$$F_Z(z) = \mathbb{P}(\max(X, Y) \leq z) = \mathbb{P}(X \leq z, Y \leq z).$$

Since X and Y were assumed to be independent, and both X and Y are uniformly distributed in $[0, 1]$, we get

$$F_Z(z) = \mathbb{P}(X \leq z)\mathbb{P}(Y \leq z) = \left(\int_{-\infty}^z \mathbf{1}_{[0,1]}(t) dt \right)^2 = \begin{cases} 0 & \text{if } z < 0, \\ z^2 & \text{if } z \in [0, 1], \\ 1 & \text{if } z > 1. \end{cases}$$

Differentiating the above yields

$$f_Z(z) = 2z \mathbf{1}_{[0,1]}(z), \quad z \in \mathbb{R}.$$

[†]Note that $\max(X, Y) \leq z \Leftrightarrow X \leq z \text{ and } Y \leq z$. Recall also the notation $\mathbb{P}(X \leq z, Y \leq z) = \mathbb{P}(X \leq z \text{ and } Y \leq z)$.

The following change of variable formula works in the case where X, Y are \mathbb{R}^n -valued random variables and $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is C^1 -diffeomorphism (i.e., g is a bijection and both g and its inverse g^{-1} are continuously differentiable). The **Jacobian matrix** of a vector field

$F(x) = [F_1(x), \dots, F_n(x)]^T$, where $F_j: \mathbb{R}^n \rightarrow \mathbb{R}$ for $j = 1, \dots, n$, is

$$DF(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} F_1(x) & \cdots & \frac{\partial}{\partial x_n} F_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} F_n(x) & \cdots & \frac{\partial}{\partial x_n} F_n(x) \end{bmatrix}.$$

Theorem

Let $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a C^1 -diffeomorphism and let X and Y be \mathbb{R}^n -valued random variables such that $Y = g(X)$. Then

$$f_X(x) = f_Y(g(x)) |\det Dg(x)|, \quad x \in \mathbb{R}^n,$$

and

$$f_Y(y) = f_X(g^{-1}(y)) |\det Dg^{-1}(y)|, \quad y \in \mathbb{R}^n.$$

Proof. The argument is exactly the same as the univariate version (use the multivariate change of variables formula for integration). □

Example

Assume that g is an affine transformation

$$g(x) = Ax + b, \quad x \in \mathbb{R}^n,$$

for some fixed vector $b \in \mathbb{R}^n$ and invertible matrix $A \in \mathbb{R}^{n \times n}$. Suppose that X has the PDF f_X and $Y = g(X)$. We wish to find the PDF f_Y of Y .

The Jacobian matrix of g is given by

$$Dg(x) = A, \quad x \in \mathbb{R}^n,$$

and we have

$$g^{-1}(y) = A^{-1}(y - b).$$

Therefore the change of variables formula yields

$$f_Y(y) = f_X(A^{-1}(y - b)) |\det A^{-1}| = f_X(A^{-1}(y - b)) \frac{1}{|\det A|}, \quad y \in \mathbb{R}^n.$$

Sums of independent random variables

Theorem

Let X and Y be independent, real-valued discrete random variables with PMFs p_X and p_Y , respectively. Then the random variable $Z = X + Y$ has the PMF

$$p_Z(z) = \sum_{x \in E} p_X(x)p_Y(z - x).$$

Example

Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be two independent Poisson random variables with parameters $\lambda, \mu > 0$. Then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Theorem

Let X and Y be independent, real-valued continuous random variables with PDFs f_X and f_Y , respectively. Then the random variable $Z = X + Y$ has the PDF

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x) dx, \quad z \in \mathbb{R}.$$

This is the **convolution** of f_X and f_Y and denoted $f_Z(z) = (f_X * f_Y)(z)$.

Positive definite matrices

Definition

Let $A \in \mathbb{R}^{d \times d}$ be a *symmetric matrix*. We call A a **positive definite matrix** if

$$x^T A x > 0 \quad \text{for all } x \in \mathbb{R}^d \setminus \{0\}.$$

This implies that A is invertible and that A^{-1} is positive definite if A is.

Characterization

Let $A \in \mathbb{R}^{d \times d}$ be a *symmetric matrix*. Then the following are equivalent:

- The matrix A is positive definite.
- The eigenvalues of A are positive.
- The matrix A has a **Cholesky decomposition**: there exists an upper triangular matrix $R \in \mathbb{R}^{d \times d}$ such that

$$A = R^T R.$$

- The matrix A has a **matrix square root**, denoted by $A^{1/2}$, which satisfies

$$A = A^{1/2} A^{1/2}.$$

Note that the matrix square root $A^{1/2}$ is also positive definite.

Multivariate Gaussian random variables

Definition

Let $\mu \in \mathbb{R}^d$ and let $C \in \mathbb{R}^{d \times d}$ be a positive definite matrix. We call a random variable X on \mathbb{R}^d a **multivariate Gaussian random variable** with mean μ and covariance C if it has the PDF

$$f_X(x) = \left(\frac{1}{(2\pi)^d \det C} \right)^{1/2} \exp \left(-\frac{1}{2}(x - \mu)^T C^{-1}(x - \mu) \right), \quad x \in \mathbb{R}^d.$$

In this case, we denote $X \sim \mathcal{N}(\mu, C)$.

Remark. There exists a concept of Gaussian random variable even in the case where the matrix C is positive semi-definite, i.e., at least one of its eigenvalues is 0, but such a random variable does not have a well-defined PDF (it is a “degenerate” random variable). The definition uses the so-called characteristic function. We omit the details.

The inverse of the covariance matrix is sometimes called a *precision matrix*. An often used notation is $\|x\|_C = \sqrt{x^T C^{-1} x}$ for $x \in \mathbb{R}^d$.

Transformations of Gaussian random variables

Gaussian random variables behave predictably under affine transformations:

- Multiplying a Gaussian RV with a (deterministic) scalar number yields another Gaussian RV with an updated mean and variance.
- Translating a Gaussian RV yields another Gaussian RV with an updated mean, but the same variance.
- An affine transformation of a Gaussian RV yields another Gaussian RVs with an updated mean and variance.
- Nonlinear transformations of Gaussian RVs are typically no longer Gaussian RVs!
 - For example, the Euclidean norm $Y = \|X\|$ of a Gaussian RV is not Gaussian (it follows a so-called “folded normal distribution”).
 - The sum of squares of independent Gaussian RVs $Z = X_1^2 + \dots + X_k^2$, where X_i are assumed to be independent Gaussian RVs, has the $\chi^2(k)$ distribution.

Proposition (ZCA transform, univariate version)

Let $\mu \in \mathbb{R}$ and $\sigma > 0$. The univariate Gaussian distribution satisfies the following properties:

- ① If $X \sim \mathcal{N}(0, 1)$, then $Y := \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$.
- ② If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $X := \frac{1}{\sigma}(Y - \mu) \sim \mathcal{N}(0, 1)$.

Proposition (ZCA transform, multivariate version)

Let $\mu \in \mathbb{R}^d$ and let $C \in \mathbb{R}^{d \times d}$ be a symmetric positive definite covariance matrix. The multivariate Gaussian distribution satisfies the following properties:

- ① If $X \sim \mathcal{N}(0, I_d)$, then $Y := \mu + C^{1/2}X \sim \mathcal{N}(\mu, C)$.
- ② If $Y \sim \mathcal{N}(\mu, C)$, then $X := C^{-1/2}(Y - \mu) \sim \mathcal{N}(0, I_d)$.

(Here, $C^{-1/2} := (C^{1/2})^{-1}$ is the inverse of the matrix square root of C .)

Remark. (1) is called a **Mahalanobis** or **ZCA[†] coloring transform**: it turns a *standard* Gaussian RV into a Gaussian RV with specified mean and covariance. (2) is called a **Mahalanobis** or **ZCA[†] whitening transform**: it turns a Gaussian RV with a specified mean and covariance into a *standard Gaussian* RV.

[†]Zero-phase component analysis

Proof. Let us prove claim (1) of the multivariate version. Let $X \sim \mathcal{N}(0, I_d)$ and define $Y = \mu + C^{1/2}x$. By defining $g(x) = \mu + C^{1/2}x$, we can write

$$Y = g(X) \quad \Rightarrow \quad f_Y(y) = f_X(g^{-1}(y)) |\det Dg^{-1}(y)|.$$

In this case, we have

$$g^{-1}(y) = C^{-1/2}(y - \mu) \quad \text{and} \quad |\det Dg^{-1}(y)| = |\det C^{-1/2}| = \frac{1}{\sqrt{\det C}}.$$

Therefore

$$\begin{aligned} f_Y(y) &= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|C^{-1/2}(y - \mu)\|^2\right) \frac{1}{\sqrt{\det C}} \\ &= \left(\frac{1}{(2\pi)^d \det C}\right)^{1/2} \exp\left(-\frac{1}{2}(x - \mu)^T C^{-1}(x - \mu)\right), \end{aligned}$$

which implies that $Y \sim \mathcal{N}(\mu, C)$.

The proof for (2) follows by writing $X = g^{-1}(Y)$ and using the change of variables formula $f_X(x) = f_Y(g(x)) |\det Dg(x)|$. □

Linear transformation of a Gaussian random variable

Proposition

Let $\mu \in \mathbb{R}^d$ and let $C \in \mathbb{R}^{d \times d}$ be a symmetric, positive definite matrix. Let $X \sim \mathcal{N}(\mu, C)$. If $k \leq d$ and $L \in \mathbb{R}^{k \times d}$ is a matrix with full rank, then

$$Y = LX \sim \mathcal{N}(L\mu, LCL^T).$$

Different coloring transforms

Let $\mu \in \mathbb{R}^d$, let $C \in \mathbb{R}^{d \times d}$ be a symmetric positive covariance matrix, and let $X \sim \mathcal{N}(0, I_d)$.

- The Mahalanobis or ZCA coloring transform uses the matrix square root factorization $C = C^{1/2}C^{1/2}$ to write a standard Gaussian RV as

$$Y = \mu + C^{1/2}X \sim \mathcal{N}(\mu, C).$$

- One could alternatively use the Cholesky decomposition $C = R^T R$ to obtain the Cholesky coloring transform

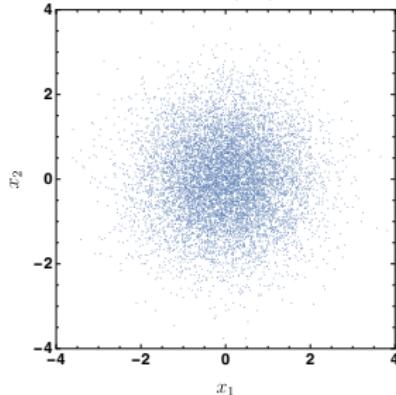
$$Y = \mu + R^T X \sim \mathcal{N}(\mu, C).$$

- Finally, one could use the eigendecomposition $C = U\Lambda U^T = (U\Lambda^{1/2})(U\Lambda^{1/2})^T$, where $UU^T = I = U^T U$ and Λ is a diagonal matrix containing the eigenvalues of C , to obtain the PCA[†] coloring transform

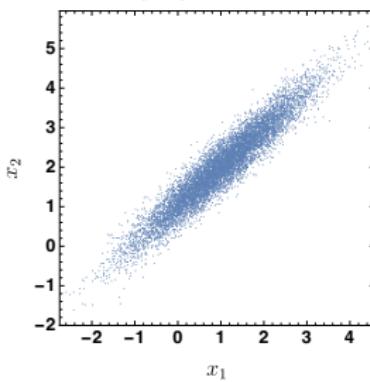
$$Y = \mu + U\Lambda^{1/2}X \sim \mathcal{N}(\mu, C).$$

[†]Principal component analysis

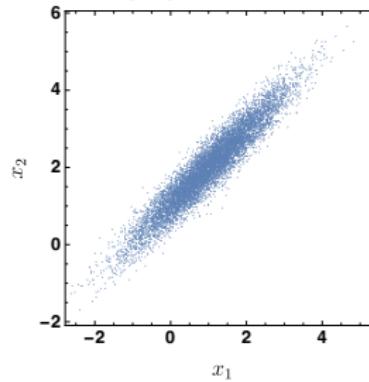
$$X \sim \mathcal{N}(0, I_2)$$



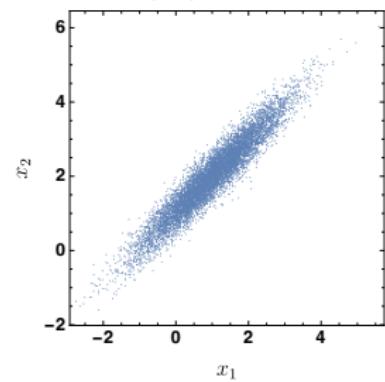
$Y \sim \mathcal{N}(\mu, C)$ using ZCA coloring



$Y \sim \mathcal{N}(\mu, C)$ using Cholesky coloring



$Y \sim \mathcal{N}(\mu, C)$ using PCA coloring



Coloring transforms with $\mu = [1, 2]^T$ and $C = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix}$.

Different whitening transforms

Let $\mu \in \mathbb{R}^d$, let $C \in \mathbb{R}^{d \times d}$ be a symmetric positive covariance matrix, and let $Y \sim \mathcal{N}(\mu, C)$.

- The **Mahalanobis or ZCA whitening transform** uses the matrix square root factorization $C = C^{1/2}C^{1/2}$ to write a standard Gaussian RV as

$$X = C^{-1/2}(Y - \mu) \sim \mathcal{N}(0, I_d).$$

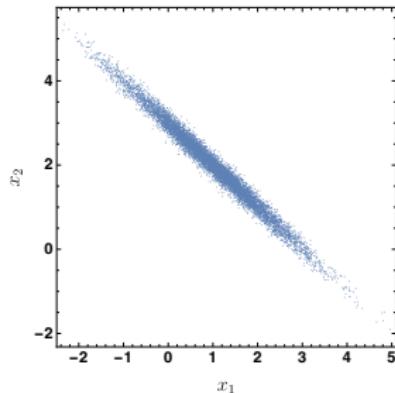
- One could alternatively use the Cholesky decomposition $C = R^T R$ to obtain the **Cholesky whitening transform**

$$X = R^{-T}(Y - \mu) \sim \mathcal{N}(0, I_d).$$

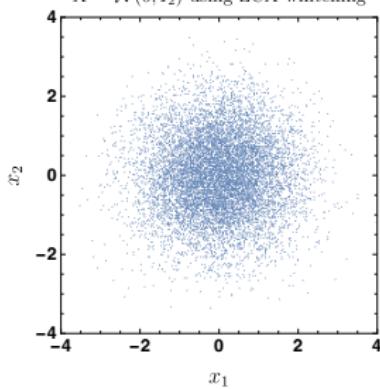
- Finally, one could use the eigendecomposition $C = U\Lambda U^T = (U\Lambda^{1/2})(U\Lambda^{1/2})^T$, where $UU^T = I = U^T U$ and Λ is a diagonal matrix containing the eigenvalues of C , to obtain the **PCA whitening transform**

$$X = \Lambda^{-1/2}U^T(Y - \mu) \sim \mathcal{N}(0, I_d).$$

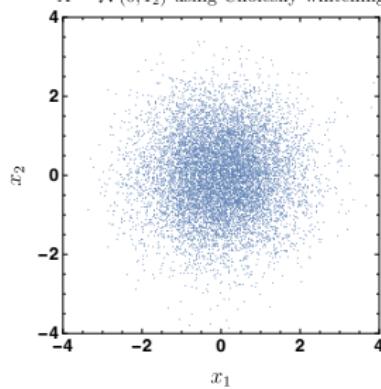
$$Y \sim \mathcal{N}(\mu, C)$$



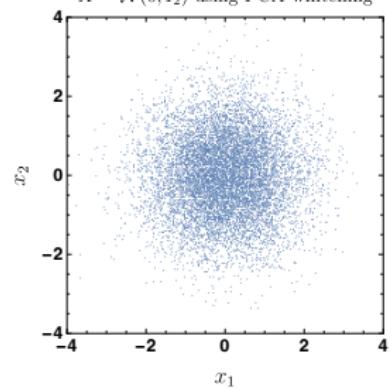
$X \sim \mathcal{N}(0, I_2)$ using ZCA whitening



$X \sim \mathcal{N}(0, I_2)$ using Cholesky whitening



$X \sim \mathcal{N}(0, I_2)$ using PCA whitening



Whitening transforms with $\mu = [1, 2]^T$ and $C = \begin{bmatrix} 1 & -0.99 \\ -0.99 & 1 \end{bmatrix}$.

By inductive reasoning, one can deduce that any finite linear combination of Gaussian RVs is a Gaussian RV.

Proposition (Univariate version)

Let $X_j \sim \mathcal{N}(\mu_i, \sigma_i^2)$ be independent Gaussian random variables with $\mu_i \in \mathbb{R}$ and $\sigma_i > 0$ for $i = 1, \dots, n$. Then

$$X := \sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

Proposition (Multivariate version)

Let $X_j \sim \mathcal{N}(\mu_i, C_i)$ be independent Gaussian random variables with $\mu_i \in \mathbb{R}^d$ and symmetric, positive definite $C_i \in \mathbb{R}^{d \times d}$ for $i = 1, \dots, n$. Then

$$X := \sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n C_i\right).$$

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fourth lecture, November 4, 2024

Expected value and covariance

Example

If a random variable X takes finitely many values x_1, \dots, x_n with equal probability, it is natural to define the *average* of X as the arithmetic average $\frac{1}{n} \sum_{i=1}^n x_i$.

More generally, if X takes the value x_i with probability p_i , then it is natural to define the average of X as the weighted average $\sum_{i=1}^n p_i x_i$, i.e., values x_i which are more likely to be realized are assigned a larger weight and *vice versa* for values x_i which are less likely to occur.

The *expected value* of a random variable is used to formalize the notion of “mean” or “average” of a real-valued random variable X .

Definition (Expected value of a discrete, real-valued RV)

Let X be a discrete, real-valued random variable with target space $E \subset \mathbb{R}$ and PMF p_X . The **expected value** (also called **mean**) of X is

$$\mathbb{E}[X] = \sum_{x \in E} x p_X(x). \quad (1)$$

Definition (Expected value of a continuous, real-valued RV)

Let X be a continuous, real-valued random variable with PDF f_X . The **expected value** (also called **mean**) of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx. \quad (2)$$

A random variable X is called **integrable** if

- X is a discrete, real-valued random variable and the series (1) is absolutely convergent.
- X is a continuous, real-valued random variable and the integral (2) is absolutely convergent.

Example

The expected value of X can be interpreted as the value that X will take on average. If we observe realizations x_1, \dots, x_n of X , then for large n , the empirical mean should be close to $\mathbb{E}[X]$:

$$\frac{1}{n} \sum_{i=1}^n x_i \approx \mathbb{E}[X].$$

Example

Assume that X is **deterministic**, i.e., there exists $x \in \mathbb{R}$ such that $X = x$ almost surely[†]. Then $\mathbb{E}[X] = x$.

Example

Let X be a discrete random variable with a **finite** target space $E \subset \mathbb{R}$. Suppose that X is uniformly distributed in E . Then

$$\mathbb{E}[X] = \frac{1}{|E|} \sum_{x \in E} x,$$

so the expected value of X coincides with the algebraic average of the values $x \in E$.

[†]The term “almost surely”, abbreviated “a.s.”, means that the probability of this outcome is 1.

Example

Let $a < b$ and assume that $X \sim \mathcal{U}(a, b)$. Then $f_X(x) = \frac{\mathbf{1}_{(a,b)}(x)}{b-a}$, and

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{\mathbf{1}_{(a,b)}(x)}{b-a} dx = \int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}.$$

Example

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ and consider $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx.$$

Performing the change of variables $y = x - \mu$, we obtain

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} (y + \mu) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}y^2} dy \\ &= \frac{1}{2\pi\sigma^2} \underbrace{\int_{-\infty}^{\infty} ye^{-\frac{1}{2\sigma^2}y^2} dy}_{= 0 \text{ as an odd function of } y} + \mu \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}y^2} dy}_{= 1 \text{ (PDF integrates to 1 over } \mathbb{R})} \\ &= \mu.\end{aligned}$$

This justifies calling the parameter μ the **mean** of the Gaussian RV X .

In many cases, one is interested in the expected value of some derived quantity of the random variable X . The following result makes this simple.

Theorem (Law of the unconscious statistician)

- If X is a discrete random variable with PMF p_X and $g: E \rightarrow \mathbb{R}$, then

$$\mathbb{E}[g(X)] = \sum_{x \in E} g(x)p_X(x).$$

- If X is a continuous RV with PDF f_X and $g: \mathbb{R} \rightarrow \mathbb{R}$ continuous,

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

- If X is a continuous \mathbb{R}^k -valued RV with PDF f_X and $g: \mathbb{R}^k \rightarrow \mathbb{R}^k$ continuous,

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^k} g(x)f_X(x) dx.$$

In other words, it is enough to know the distribution of X in order to be able to compute $\mathbb{E}[g(X)]$ for any continuous function g . It is not necessary to solve the distribution of $g(X)$.

Example

A stick of length 1 is broken into two pieces at a uniformly random point between 0 and 1. Let Y denote the length of the larger piece and we wish to know $\mathbb{E}[Y]$.

Let $X \sim \mathcal{U}(0, 1)$ denote the position of the breaking point. Then $Y = \max(X, 1 - X)$. By the law of the unconscious statistician, we obtain

$$\begin{aligned}\mathbb{E}[Y] &= \int_{-\infty}^{\infty} \max(x, 1 - x) \mathbf{1}_{(0,1)}(x) dx = \int_0^1 \max(x, 1 - x) dx \\ &= \int_0^{1/2} (1 - x) dx + \int_{1/2}^1 x dx = \frac{1}{2} - \frac{1}{8} + \frac{1}{2} - \frac{1}{8} = \frac{3}{4}.\end{aligned}$$

Example (Moments)

An important class of maps g are given by $g(x) = x^k$. Then

$$\mathbb{E}[X^k] = \begin{cases} \sum_{x \in E} x^k p_X(x) & \text{if } X \text{ is a discrete RV with target space } E \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x^k f_X(x) dx & \text{if } X \text{ is a continuous, real-valued RV} \end{cases}$$

is the k^{th} moment of X . (If $\mathbb{E}[|X|^k] = \infty$, the moment is said not to exist.)

If this expression is finite for $k = 2$, then X is called **square-integrable**.

Example

Let $a < b$ and assume that $X \sim \mathcal{U}(a, b)$. Then

$$\begin{aligned}\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 \frac{\mathbf{1}_{(a,b)}(x)}{b-a} dx = \int_a^b x^2 \frac{1}{b-a} dx \\ &= \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}.\end{aligned}$$

The probability of an event A of a probability space (Ω, \mathbb{P}) can be written as the expected value of the indicator function for set A .

Proposition

Let (Ω, \mathbb{P}) be a probability space and let $A \subset \Omega$ be an event. Define the random variable $\mathbf{1}_A: \Omega \rightarrow \mathbb{R}$,

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Then

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

Proof. Since $X = \mathbf{1}_A$ is a discrete random variable taking values in $E = \{0, 1\}$, its PMF satisfies

$$p_X(0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A), \quad p_X(1) = \mathbb{P}(A).$$

Hence

$$\mathbb{E}[X] = 0 \cdot p_X(0) + 1 \cdot p_X(1) = \mathbb{P}(A). \quad \square$$

Properties of the expected value

Proposition

Let X be a real-valued random variable and $a, b \in \mathbb{R}$. Then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Proof. For continuous random variables: $\mathbb{E}[aX + b] = \int_{\mathbb{R}} (ax + b)f_X(x) dx$
 $= a \underbrace{\int_{\mathbb{R}} xf_X(x) dx}_{= \mathbb{E}[X]} + b \underbrace{\int_{\mathbb{R}} f_X(x) dx}_{= 1}$. The proof is similar for discrete RVs. \square

Theorem

- ① If $X \geq 0$ almost surely, then $\mathbb{E}[X] \geq 0$. (Similarly, if $X \leq 0$ almost surely, then $\mathbb{E}[X] \leq 0$.)
- ② If X_1, \dots, X_n are real-valued random variables and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, then

$$\mathbb{E}\left[\sum_{i=1}^n \alpha_i X_i\right] = \sum_{i=1}^n \alpha_i \mathbb{E}[X_i].$$

- ③ If $X \leq Y$ almost surely, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Finally, the expected value of a product of **independent** random variables is the product of the expected values.

Theorem

Let X_1, \dots, X_n be **independent** real-valued random variables. Then

$$\mathbb{E}\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n \mathbb{E}[X_i].$$

Variance

Definition

Let X be a real-valued random variable with mean $\mu = \mathbb{E}[X]$. The **variance** of X is defined as

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

Note that this quantity is well-defined provided that $\mathbb{E}[X^2] < \infty$.

The **standard deviation** of X is defined as

$$\sigma_X = \sqrt{\text{Var}(X)}.$$

- Note that $\text{Var}(X) = \sum_{x \in E} (x - \mu)^2 p_X(x)$ if X is a discrete random variable with PMF p_X , and $\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx$ if X is a continuous random variable with PDF f_X .
- The variance $\text{Var}(X)$ is always *nonnegative*. While $\mathbb{E}[X]$ represents the *average value* of X , $\text{Var}(X)$ quantifies how far realizations of X can spread away from this average value.

Theorem (Variance translation)

Let $\mu = \mathbb{E}[X]$ denote the mean of random variable X . Then

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

Proof.

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu \underbrace{\mathbb{E}[X]}_{=\mu} + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2.\end{aligned}\quad \square$$

Remark. If the random variable X satisfies $\mathbb{E}[X] = 0$, then we say that X is **centered**. In this case, we simply have $\text{Var}(X) = \mathbb{E}[X^2]$.

Example

Let $a < b$ and suppose that $X \sim \mathcal{U}(a, b)$. We have already computed that

$$\mathbb{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \mathbb{E}[X^2] = \frac{a^2 + ab + b^2}{3}.$$

Therefore

$$\text{Var}(X) = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12},$$

and the standard deviation $\sigma_X = \frac{b-a}{2\sqrt{3}}$. Hence, the larger the interval $[a, b]$ for the uniform distribution, the larger the standard deviation.

Example

Let $\mu \in \mathbb{R}$ and $\sigma > 0$ and suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx.$$

Carrying out the change of variables $y = \frac{x-\mu}{\sigma}$, where $dx = \sigma dy$, we get

$$\text{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{1}{2}y^2} dy.$$

Since

$$\int_{-\infty}^{\infty} y^2 e^{-\frac{1}{2}y^2} dy = \sqrt{2\pi} \tag{3}$$

(see the following slide for an argument), we conclude that

$$\text{Var}(X) = \sigma^2.$$

This justifies calling the parameter σ^2 the variance of the Gaussian RV X .

Intermezzo – computing the value of the integral (3)

Let $a > 0$ be a parameter and consider the following *parametric integral*:

$$\begin{aligned} I(a) &:= \int_{-\infty}^{\infty} y^2 e^{-\frac{1}{2}ay^2} dy = -2 \int_{-\infty}^{\infty} \frac{\partial}{\partial a} e^{-\frac{1}{2}ay^2} dy \\ &\stackrel{(*)}{=} -2 \frac{d}{da} \int_{-\infty}^{\infty} e^{-\frac{1}{2}ay^2} dy. \end{aligned}$$

Applying $\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}x^2} dx = 1 \Leftrightarrow \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}x^2} dx = \sqrt{2\pi}\sigma$
with $\sigma = \frac{1}{\sqrt{a}}$ yields

$$I(a) = -2 \frac{d}{da} \frac{\sqrt{2\pi}}{\sqrt{a}} = \frac{\sqrt{2\pi}}{a^{3/2}}.$$

The value of the integral (3) corresponds to $I(1) = \sqrt{2\pi}$.

This technique is known as the “Leibniz integral rule”, or “Feynman’s differentiation under the integral sign”. The difficult part is verifying that the order of integration and differentiation can be switched in (*). This is allowed, e.g., when the integrand $f(a, y)$ is continuously differentiable.

Theorem

- ① If X is a real-valued random variable and $a, b \in \mathbb{R}$, then

$$\text{Var}(aX + b) = a^2\text{Var}(X).$$

- ② If X_1, \dots, X_n are *independent* real-valued random variables and $a_1, \dots, a_n \in \mathbb{R}$, then

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i).$$

Covariance and correlation

Definition

Let X and Y be two real-valued random variables with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then the **covariance** of X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)].$$

If $\sigma_X^2 = \text{Var}(X)$ and $\sigma_Y^2 = \text{Var}(Y)$ are the variances, then the **correlation** of X and Y is

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Remark. The correlation always satisfies

$$-1 \leq \rho_{X,Y} \leq 1$$

as a consequence of the *Cauchy–Schwarz inequality*.

Theorem

Let X and Y be two real-valued random variables with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. Then

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X\mu_Y.$$

Proof.

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbb{E}[XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \underbrace{\mu_Y \mathbb{E}[X]}_{=\mu_X} - \underbrace{\mu_X \mathbb{E}[Y]}_{=\mu_Y} + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mu_X\mu_Y. \quad \square\end{aligned}$$

The random variables X and Y are said to be **uncorrelated** if $\text{Cov}(X, Y) = 0$.

Theorem

If X and Y are independent, then X and Y are uncorrelated.

Proof. Since $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for independent X and Y , there holds

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X\mu_Y = \underbrace{\mathbb{E}[X]\mathbb{E}[Y]}_{=\mu_X = \mu_Y} - \mu_X\mu_Y = 0. \quad \square$$

Note that, in general, X, Y are uncorrelated $\not\Rightarrow X, Y$ are independent!
(However, this converse statement does hold for *jointly Gaussian distributions* – we will formulate a special case of this in a moment.)

Theorem

$$\text{Var}(X + Y) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y),$$

$$\text{Var}(X - Y) = \text{Var}(X) - 2\text{Cov}(X, Y) + \text{Var}(Y).$$

Joint random variables

Definition

Let $X = (X_1, \dots, X_d)$, $d \in \mathbb{N}$, be a joint random variable. We define the **mean** $\mu = (\mu_i)_{i=1}^d \in \mathbb{R}^d$ and the **covariance matrix** $C = (C_{i,j})_{i,j=1}^d \in \mathbb{R}^{d \times d}$ of X by

$$\mu_i = \mathbb{E}[X_i] \quad \text{for } i = 1, \dots, d,$$

$$C_{i,j} = \text{Cov}(X_i, X_j) \quad \text{for } i, j = 1, \dots, d.$$

Example

Let $X = (X_1, \dots, X_d)$ be a d -dimensional Gaussian random variable $X \sim \mathcal{N}(\mu, C)$, where $\mu = (\mu_i)_{i=1}^d \in \mathbb{R}^d$ and $C = (C_{i,j})_{i,j=1}^n \in \mathbb{R}^{d \times d}$ is a symmetric, positive definite matrix. Then

$$\mu_i = \mathbb{E}[X_i] \quad \text{for } i = 1, \dots, d,$$

$$C_{i,j} = \text{Cov}(X_i, X_j) \quad \text{for } i, j = 1, \dots, d,$$

meaning that μ is the mean of X and C is the covariance matrix of X .

Corollary (Independence of jointly Gaussian random variables)

Let $X = (X_1, \dots, X_d) \sim \mathcal{N}(\mu, C)$ for $\mu = (\mu_j)_{j=1}^d \in \mathbb{R}^d$ and symmetric, positive definite $C = (C_{i,j})_{i,j=1}^d \in \mathbb{R}^{d \times d}$. Then X_1, \dots, X_d are independent if and only if C is a diagonal matrix, i.e., $C_{i,j} = 0$ whenever $i \neq j$.

Proof. “ \Rightarrow ” If X_1, \dots, X_d are independent, then X_i and X_j are independent for all $i \neq j$. Independent random variables are uncorrelated, so the covariance

$$C_{i,j} = \text{Cov}(X_i, X_j) = 0 \quad \text{whenever } i \neq j.$$

“ \Leftarrow ” Let $C = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$. Then the marginal distribution of X_j is Gaussian, with PDF $f_{X_j}(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2\sigma_j^2}(x-\mu_j)^2}$. Hence,

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det C}} e^{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)} = \prod_{j=1}^d \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2\sigma_j^2}(x_j-\mu_j)^2},$$

i.e., $f_X(x) = f_{X_1}(x_1) \cdots f_{X_d}(x_d)$, meaning that X_1, \dots, X_d are independent. □

Sample mean and sample variance

In practice, the random variables are not observed directly: we observe realizations, or a **sample**, thereof. It is useful to define notions of *sample mean* and *sample variance*, which are quantities that can be computed directly from the observed realizations.

Definition

Let X_1, \dots, X_n be real-valued random variables[†]. The **sample mean** of is defined as the arithmetic average

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

The **sample variance** is defined as

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

and the **sample standard deviation** is defined as $s_n = \sqrt{s_n^2}$.

Remark. Note that the sample mean \bar{X}_n and the sample variance s_n^2 are themselves **random variables**. As we shall see, if X_1, \dots, X_n are *independent and identically distributed* provided some integrability conditions are satisfied, then there holds for large n that

$$\bar{X}_n \approx \mathbb{E}[X_1] \quad \text{and} \quad s_n^2 \approx \text{Var}(X_1).$$

[†]One may think of X_1, \dots, X_n as representing a sample from some random variable X .

Sample covariance of vector-valued random variables

If X_1, \dots, X_n are vector-valued random variables taking values in \mathbb{R}^d , then their sample covariance matrix $\mathbf{Q} = (Q_{j,k})_{j,k=1}^n$ is defined as

$$Q_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (X_{i,j} - \mu_j)(X_{i,k} - \mu_k), \quad j, k = 1, \dots, d,$$

where $\mu = \bar{\mathbf{X}} = (\bar{X}_1, \dots, \bar{X}_d)$ is the mean.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fifth lecture, November 11, 2024

Inequalities and limits

Random sample / i.i.d. random variables

Let X_1, \dots, X_n be random variables. We call X_1, \dots, X_n a **random sample** if the random variables are **independent** and **identically distributed** (i.i.d.).

- **Independent** means that X_1, \dots, X_n are mutually independent random variables.
- **Identically distributed** means that X_1, \dots, X_n all have the same **law**.

Often, we specify the law (probability distribution) of a random variable X and say that X_1, \dots, X_n are i.i.d. copies of X .

Example

Let $X \sim \mathcal{N}(0, 1)$. Suppose that X_1, \dots, X_n are i.i.d. copies of X . This means that

- $X_i \sim \mathcal{N}(0, 1)$ for all $i = 1, \dots, n$ ("identically distributed").
- X_i are mutually independent:
 $p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_X(x_1) \cdots p_X(x_n)$, where $p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ is the PDF of $X \sim \mathcal{N}(0, 1)$ ("independence").

In practice, the terms "random sample" and "i.i.d." are interchangeable.

- We begin by deriving bounds on the probabilities that a random variable X stays away from its mean by a certain distance $t > 0$:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t).$$

- Then we discuss two results, which lie at the heart of statistical inference: the **Law of Large Numbers (LLN)** and the **Central Limit Theorem (CLT)**. The LLN states that, if X_1, X_2, \dots are i.i.d. random variables with finite mean, then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \rightarrow \infty} \mathbb{E}[X]$$

where the convergence happens in a sense to be specified. The CLT states that, if the i.i.d. random variables X_1, X_2, \dots have finite variance, then this convergence happens at rate $\mathcal{O}(n^{-1/2})$.

- Together, these two results can be used to obtain *approximate* bounds on the probability that the empirical sum remains away from its mean:

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}[X]\right| > \frac{t}{\sqrt{n}}\right)$$

for fixed $t > 0$ and n large.

Inequalities for expected values

Theorem (Cauchy–Schwarz inequality)

Let X and Y be two square-integrable, real-valued random variables.[†]

Then

$$\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]} \sqrt{\mathbb{E}[Y^2]}.$$

Proof. If $X = 0$ or $Y = 0$ almost surely, then the claim is trivial. Suppose that $X \neq 0$ and $Y \neq 0$ almost surely. Let $t \in \mathbb{R}$ and note that

$$0 \leq \mathbb{E}[(X + tY)^2] = \mathbb{E}[X^2] + 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2]$$

is a second degree polynomial with respect to t which has at most one real root. Therefore its discriminant must be nonpositive:

$$\begin{aligned} \text{discriminant} \leq 0 &\Leftrightarrow (2\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0 \\ &\Leftrightarrow \mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \quad \square \end{aligned}$$

[†]Recall that square-integrability implies that $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ are well-defined and finite.

Let X be a real-valued random variable. A fundamental problem in statistics is to be able to bound from above the probability $\mathbb{P}(X > t)$ for fixed $t > 0$. Bounds of the following kind are known as “tail bounds”.

Theorem (Markov's inequality)

If X is an integrable[†], non-negative real-valued random variable and $t > 0$, then

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Let us consider the case of X being a continuous RV (the discrete case is similar). There holds

$$\mathbb{P}(X > t) = \int_t^\infty f_X(x) dx \stackrel{(*)}{\leq} \frac{1}{t} \int_t^\infty xf_X(x) dx \leq \frac{1}{t} \int_0^\infty xf_X(x) dx,$$

where $(*)$ follows from $x \geq t \Leftrightarrow 1 \leq \frac{x}{t}$. Since we assumed that X is non-negative, $f_X(x) = 0$ for $x < 0$, and thus

$$\mathbb{P}(X > t) \leq \frac{1}{t} \int_0^\infty xf_X(x) dx = \frac{1}{t} \int_{-\infty}^\infty xf_X(x) dx = \frac{\mathbb{E}[X]}{t}. \quad \square$$

[†]Recall that this means $\mathbb{E}[|X|] < \infty$.

If X is a square-integrable random variable, then we can bound the probability that X is at least distance $t > 0$ away from the average.

Theorem (Chebyshev's inequality)

Let X be a square-integrable random variable. For all $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq \frac{\text{Var}(X)}{t^2}.$$

Proof. By Markov's inequality,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}(|X - \mathbb{E}[X]|^2 > t^2) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{t^2} = \frac{\text{Var}(X)}{t^2},$$

where we applied Markov's inequality $\mathbb{P}(Y > t') \leq \frac{\mathbb{E}[Y]}{t'}$ to the non-negative random variable $Y = |X - \mathbb{E}[X]|^2$ and $t' = t^2$. □

Let $\sigma = \sqrt{\text{Var}(X)}$. It is sometimes useful to rewrite the Chebyshev inequality in the form (set $t = k\sigma$)

$$\mathbb{P}(|X - \mathbb{E}[X]| > k\sigma) \leq \frac{1}{k^2}.$$

If $k = 2$, then $1 - \frac{1}{k^2} = 75\%$.

If $k = 3$, then $1 - \frac{1}{k^3} \approx 88.9\%$.

In practice, expected value and variance must be estimated. Chebyshev's inequality can be used to evaluate the rareness of a single observation.

The Chebyshev inequality can be useful in situations where we **only** know the mean and variance of X . On the other hand, it is quite a rough bound. If we know the distribution of X , the probability $\mathbb{P}(|X - \mathbb{E}[X]| > t)$ can be computed more precisely, typically leading to much better bounds.

Example

Let X be a random variable with mean $\mathbb{E}[X] = 0$ and variance $\text{Var}(X) = 1$. Suppose that we wish to estimate $\mathbb{P}(|X| > 2)$.

If the mean and variance is all we know about the random variable, then Chebyshev's inequality gives a very rough bound:

$$\mathbb{P}(|X| > 2) \leq \frac{1}{2^2} = \frac{1}{4} = 0.25.$$

If X is a Gaussian random variable, i.e., in this case we would have $X \sim \mathcal{N}(0, 1)$, then we know precisely

$$\mathbb{P}(|X| > 2) = 2\Phi(-2) = 0.04550\dots$$

If X is a Gaussian random variable, then the probabilities $\mathbb{P}(|X| > t)$, $t > 0$, can be computed numerically using the CDF. Unfortunately, the CDF does not have a closed form expression. Sometimes the following bound is useful.

Theorem (Mill's inequality)

Let $X \sim \mathcal{N}(0, 1)$. Then for all $t > 0$,

$$\mathbb{P}(|X| > t) \leq \sqrt{\frac{2}{\pi}} \frac{\exp(-\frac{1}{2}t^2)}{t}.$$

Proof. Let $X \sim \mathcal{N}(0, 1)$ and $t > 0$. Then

$$\begin{aligned}\mathbb{P}(|X| > t) &= \int_{-\infty}^{-t} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds + \int_t^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds \\ &= 2 \int_t^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds.\end{aligned}$$

It is enough to bound this last integral.

Arguing similarly as in the proof of Markov's inequality,

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds \leq \int_t^\infty \frac{s}{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds = \frac{1}{t\sqrt{2\pi}} \int_t^\infty s e^{-\frac{s^2}{2}} ds.$$

For this integral, we have

$$\int_t^\infty s e^{-\frac{s^2}{2}} ds = - \left[e^{-\frac{s^2}{2}} \right] \Big|_{s=t}^{s=\infty} = e^{-\frac{t^2}{2}},$$

which yields the assertion. □

The previous result can be generalized to Gaussian random variables with arbitrary variance via the whitening transform.

Theorem

Let $\mu \in \mathbb{R}$, $\sigma > 0$, and let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then for all $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq \sqrt{\frac{2\sigma^2}{\pi}} \frac{e^{-\frac{t^2}{2\sigma^2}}}{t}.$$

Proof. By the whitening transform, the random variable $Y = \frac{1}{\sigma}(X - \mu) \sim \mathcal{N}(0, 1)$, so

$$\mathbb{P}(|X - \mu| > t) = \mathbb{P}(|Y| > \sigma^{-1}t),$$

and the result follows from Mill's inequality with t replaced by $\sigma^{-1}t$. □



Limit theorems

We will state two fundamental limit theorems for sums of i.i.d. random variables. To do so, we will first need to define what we mean by convergence of a sequence of random variables.

Definition (Convergence in probability)

Let X be a real-valued random variable and let $(X_n)_{n \geq 0}$ be a sequence of real-valued random variables. We say that X_n converges to X in probability, and write $X_n \xrightarrow{P} X$, if for any $\varepsilon > 0$, there holds

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

In other words, X_n converges to X in probability if the probability that X_n is separated from X by any (even very small) non-zero distance vanishes as n grows.

Another, weaker form of convergence involves the CDFs F_{X_n} and F_X of the RVs X_n and X , respectively.

Definition

Let X be a real-valued random variable and let $(X_n)_{n \geq 0}$ be a sequence of real-valued random variables. We say that X_n converges to X in distribution (or in law), and write $X_n \xrightarrow{d} X$, if for any $x \in \mathbb{R}$ where F_X is continuous, there holds

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x).$$

- If X is a continuous random variable, then F_X is everywhere continuous, and the above condition simply means that F_{X_n} converges pointwise to F_X .
- If X is discrete, then F_X will be discontinuous at every point x such that $\mathbb{P}(X = x) > 0$. The above definition says that, when checking whether X_n converges in distribution to X , we do not need to look at these points of discontinuities.
- That X_n converges in distribution to X means that $\mathbb{P}(X_n \leq x) \xrightarrow{n \rightarrow \infty} \mathbb{P}(X \leq x)$ for all points x where F_X does not jump. It is only a statement about the probability distributions of X_n and X . In particular, it does not say at all that X_n is close to X when n is large.

Proposition

- ① If X_n and X are square-integrable and

$$\mathbb{E}[|X_n - X|^2] \xrightarrow{n \rightarrow \infty} 0, \quad (1)$$

then $X_n \xrightarrow{P} X$. The converse is false in general.

- ② If X_n converges to X in probability, then X_n also converges to X in law. The converse is false in general.
- ③ If X is constant, i.e., there exists $a \in \mathbb{R}$ such that $X = a$ almost surely, then

$$X_n \xrightarrow{P} X \Leftrightarrow X_n \xrightarrow{d} X.$$

The convergence (1) is called “convergence in quadratic mean”, and written $X_n \xrightarrow{q.m.} X$. By the above proposition, convergence in quadratic mean is strictly stronger than convergence in probability, and convergence in probability is strictly stronger than convergence in distribution.

Proof. We only prove the first claim. Assume that $X_n \xrightarrow{q.m.} X$. Then for all $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^2 > \varepsilon^2) \leq \frac{\mathbb{E}[|X_n - X|^2]}{\varepsilon^2},$$

where the last inequality is a consequence of Markov's inequality. By assumption, $\mathbb{E}[|X_n - X|^2] \xrightarrow{n \rightarrow \infty} 0$. Hence, by the above inequality, we get $\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0$. This proves that $X_n \xrightarrow{P} X$. That the converse implication is false in general can be shown by counterexample (left as an exercise). □

The Law of Large Numbers (LLN)

Before stating the LLN, we need a technical, but intuitive, lemma.

Lemma

Let X and Y be real-valued random variables which are equal in law. Then, for any real-valued map such that $f(X)$ is integrable, we have $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$.

Proof. Let us prove the claim for discrete RVs (the continuous case is similar just by replacing PMFs with PDFs and sums by integrals). Let X and Y be discrete. Then $p_X = p_Y$, so for all integrable functions f , by the law of the unconscious statistician, there holds

$$\mathbb{E}[f(X)] = \sum_{x \in E} p_X(x)f(x) = \sum_{y \in E} p_Y(y)f(y) = \mathbb{E}[f(Y)]. \quad \square$$

The above result implies also that if X and Y are equal in law, then

$$\mathbb{E}[X] = \mathbb{E}[Y], \quad \mathbb{E}[X^2] = \mathbb{E}[Y^2], \quad \text{Var}(X) = \text{Var}(Y),$$

provided that these quantities are well-defined.

Let $(X_n)_{n \geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X . By this we mean that $(X_n)_{n \geq 1}$ is a sequence of i.i.d. real-valued random variables having the same law as X . For all $n \geq 1$, let \bar{X}_n denote the sample mean of X_1, \dots, X_n :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

If X_i are integrable, then by linearity of the expected value, there holds

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mathbb{E}[X].$$

Heuristically, we expect \bar{X}_n to converge to $\mathbb{E}[X]$ when $n \rightarrow \infty$. This is made precise by the following theorem.

Theorem (Weak Law of Large Numbers)

Let $(X_i)_{i \geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X . If X_i are integrable, then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}[X].$$

Proof. For simplicity, we provide a proof in the special case where the X_n are also square-integrable. Then

$$\mathbb{E}[(\bar{X}_n - \mathbb{E}[X])^2] = \mathbb{E}[(\bar{X}_n - \mathbb{E}[\bar{X}_n])^2] = \text{Var}(\bar{X}_n).$$

Now

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \sum_{i=1}^n \frac{1}{n^2} \text{Var}(X_i),$$

where the second equality holds since the X_i are independent. Now, by the technical lemma we proved prior to this result, $\text{Var}(X_i) = \text{Var}(X)$ for all i , so we get

$$\text{Var}(\bar{X}_n) = \frac{\text{Var}(X)}{n} \xrightarrow{n \rightarrow \infty} 0.$$

Hence $\mathbb{E}[(\bar{X}_n - \mathbb{E}[X])^2] \xrightarrow{n \rightarrow \infty} 0$, therefore \bar{X}_n converges to $\mathbb{E}[X]$ in quadratic mean, and hence also in probability. □



A stronger statement holds with the same assumptions.

Theorem (Strong Law of Large Numbers)

Let $(X_i)_{i \geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X . If X_i are integrable, then

$$\mathbb{P}(\{\omega \in \Omega \mid \overline{X}_n(\omega) \xrightarrow{n \rightarrow \infty} \mathbb{E}[X]\}) = 1.$$

That is, $\overline{X}_n \rightarrow \mathbb{E}[X]$ almost surely.

Remark. The significance of the LLN is that it provides a concrete way of approximating the value of $\mathbb{E}[X]$ by sampling values of X a large number of times and taking the sample average.

Example

Let $X_1, \dots, X_n \sim \text{Ber}(p)$ be independent for some $p \in (0, 1)$. Then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}[X_1] = p.$$

In other words, if we keep throwing a coin with parameter p a large number of times, the rate of success will converge in probability to p . If the coin is fair, i.e., $p = 1/2$, then the rate of success approaches $1/2$ for n large.

Example

Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be independent for some $\mu \in \mathbb{R}$ and $\sigma > 0$.

Then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}[X_1] = \mu.$$

The LLN implies that $\bar{X}_n = \mathbb{E}[X_1] + \varepsilon_n$, where ε_n is some remainder satisfying $\varepsilon_n \xrightarrow{P} 0$. The obvious question to consider is to ask, **how fast does ε_n converge to 0?**

The Central Limit Theorem

Let X_1, X_2, \dots be a sequence i.i.d. real-valued random variables. We assume that the X_i are square-integrable and denote by μ and σ^2 their mean and variance, respectively. Thus, for all i ,

$$\mathbb{E}[X_i] = \mu, \quad \text{Var}(X_i) = \sigma^2.$$

As we saw in the previous section,

$$\mathbb{E}[\bar{X}_n] = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

We can perform an affine transformation on \bar{X}_n in order to set its expectation and variance to 0 and 1, respectively. This can be achieved as follows:

- ① We **center** it by subtracting its mean $\mathbb{E}[\bar{X}_n]$,
- ② We **normalize** it by dividing it by its standard deviation $\sqrt{\text{Var}(\bar{X}_n)}$.

In other words, we set

$$Y_n = \frac{1}{\sqrt{\text{Var}(\bar{X}_n)}} (\bar{X}_n - \mathbb{E}[\bar{X}_n]) = \sqrt{\frac{n}{\sigma^2}} (\bar{X}_n - \mu).$$

With this procedure, we obtain a random variable Y_n which is centered and normalized, i.e., which satisfies

$$\mathbb{E}[Y_n] = 0, \quad \text{Var}(Y_n) = 1.$$

The following theorem shows that, for n large, the distribution of Y_n is actually close to $\mathcal{N}(0, 1)$.

Theorem (Central Limit Theorem)

Let X_1, X_2, \dots be a sequence of i.i.d. real-valued, square-integrable random variables with mean μ and variance σ^2 . Then

$$\sqrt{\frac{n}{\sigma^2}} (\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, 1).$$

Remark. The CLT implies that, for all $a \in \mathbb{R}$,

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}(\bar{X}_n - \mu) \leq a\right) \xrightarrow{n \rightarrow \infty} \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx.$$

Remark. One may loosely formulate the CLT as saying that

$$\sqrt{\frac{n}{\sigma^2}}(\bar{X}_n - \mu) \stackrel{d}{\approx} \mathcal{N}(0, 1)$$

for n large. In other words,

$$\bar{X}_n \stackrel{d}{\approx} \mu + \sqrt{\frac{\sigma^2}{n}} Z,$$

where $Z \sim \mathcal{N}(0, 1)$. Thus, by the coloring transform,

$$\bar{X}_n \stackrel{d}{\approx} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Example

Let $X_1, \dots, X_n \sim \text{Ber}(p)$ be independent for some $p \in (0, 1)$. We know from the LLN that

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}[X] = p.$$

Since $\text{Var}(X) = p(1 - p)$, the CLT further implies that

$$\sqrt{\frac{n}{p(1-p)}}(\bar{X}_n - p) \xrightarrow{d} \mathcal{N}(0, 1),$$

or, loosely speaking,

$$\bar{X}_n \xrightarrow{d} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \quad \text{for } n \text{ large.}$$

Example (Continued)

By approximating \bar{X}_n using the Gaussian distribution, we can make inferences about the spread of \bar{X}_n . For example, if $p = \frac{1}{2}$ and $n = 10^4$, we can use the Gaussian approximation to derive a confidence interval \mathcal{I} such that $\mathbb{P}(\bar{X}_n \in \mathcal{I}) \approx 0.95$. Since n is large, we can use the Gaussian approximation

$$\bar{X}_n \approx \mu + \sigma_n Z, \quad Z \sim \mathcal{N}(0, 1),$$

where $\mu = p$ and $\sigma_n = \sqrt{\frac{p(1-p)}{n}}$. We wish to find $a > 0$ such that

$$\int_{\mu-a}^{\mu+a} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{1}{2\sigma_n^2}(x-\mu)^2} dx = 0.95.$$

Using the change of variables $z = \frac{x-\mu}{\sigma_n}$, where $dx = \sigma_n dz$, we obtain

$$\begin{aligned} \int_{-a/\sigma_n}^{a/\sigma_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0.95 &\Leftrightarrow 2 \int_0^{a/\sigma_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = 0.95 \\ &\Leftrightarrow \int_{-\infty}^{a/\sigma_n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{2} + \frac{0.95}{2}. \end{aligned}$$

Example (Continued)

Using the CDF $\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$, we obtain

$$\Phi\left(\frac{a}{\sigma_n}\right) = \frac{1}{2} + \frac{0.95}{2} \quad \Leftrightarrow \quad a = \sigma_n \Phi^{-1}\left(\frac{1}{2} + \frac{0.95}{2}\right).$$

Plugging in the values $\mu = p = \frac{1}{2}$ and $\sigma_n = \sqrt{\frac{p(1-p)}{n}} = \frac{1}{200}$ yields the interval

$$\mathcal{I} = (\mu - a, \mu + a) = (0.4902, 0.5098).$$

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Sixth lecture, November 18, 2024

Statistical testing

Statistical research is collecting, organizing, analyzing, and interpreting data.

Statistical models are mathematical and are based on probability theory.

- In probability theory, if we know the law of a random variable, then we are easily able to draw an i.i.d. sample from the distribution, compute the probabilities of different events, compute the expected value, variance, higher moments, etc.
- In statistics, we are usually given a finite sample, and we are interested in making inferences about the distribution and population parameters such as the expected value, variance, higher moments, etc. We are also interested in assessing the *uncertainty* of the population parameters (confidence interval).
 - If the data (approximately) follows a distribution which we are able to identify, we can make use of the properties of that distribution from probability theory to assess the uncertainty (parametric tests).
 - It is also important to discuss statistical methods for data which does not clearly follow a known distribution (non-parametric tests).
- Correlations between variables, regression models, . . .

Population and sample

- In statistical analysis, a **population** is a collection of all the people, items or events about which one wants to make inferences. (For example, university students in Germany.)
- In statistical analysis, a **sample** is a subset of the population (i.e., the people, items or events) that one collects and analyzes to make inferences. (For example, 200 randomly chosen university students.)
- In statistical analysis, an **observation** is an element of the sample. (For example Helen, a student at FU Berlin.)

In statistical research, **data** consists of the values of selected **variables** that describe the observations. The data points (the values of the selected variables) can also be called **observations**.

Examples:

- temperature, height, blood pressure (continuous quantitative variables)
- gender, eye color (categorical qualitative variables)
- clothing size (s,m,l) (ordinal quantitative variable)

Statistical research projects

Statistical research projects can usually be conducted in the following steps:

- ① Setting of the research topic and the relevant research questions.
Research questions should be defined precisely.
- ② Defining of the population and interesting variables.
- ③ Planning of the sample collection. Collected sample must represent the population!
- ④ Collection of the sample.
- ⑤ Organization of the sample.
- ⑥ Description of the variables and the sample, descriptive statistics and visualization.
- ⑦ Inference based on statistical analysis. Model assumptions have to be tested separately!
- ⑧ Critical evaluation of the analysis. Possible errors and weaknesses have to be reported.
- ⑨ Communication of the research and findings.

Different statistical studies

Statistical research projects can be conducted in several different ways. Research questions, population, goals, and resources all have an effect on the choice of the methods.

- In **observational research**, observations are made without changing any existing conditions. For example, temperature is measured or the lung cancer risk of smokers is compared to the lung cancer risk on non-smokers.
- In **controlled experiments**, the effect of one variable to another is examined by controlling existing conditions. For example, the effect of allergy medicine is compared to the effect of placebo by randomizing patients to two groups.

Different statistical studies

- In **simulations**, mathematical modeling is used to mimic natural conditions or processes. For example, the spread of the Ebola virus is predicted by applying computer simulations or the safety of a new car model is tested using crash test dummies.
- In **surveys**, the goal is to find a representative sample of the population and get answers to some particular questions. For example, opinion polls are used in order to predict election results, or health related questionnaires are used to assess the health of university students.

Descriptive statistics

Descriptive statistics and inference

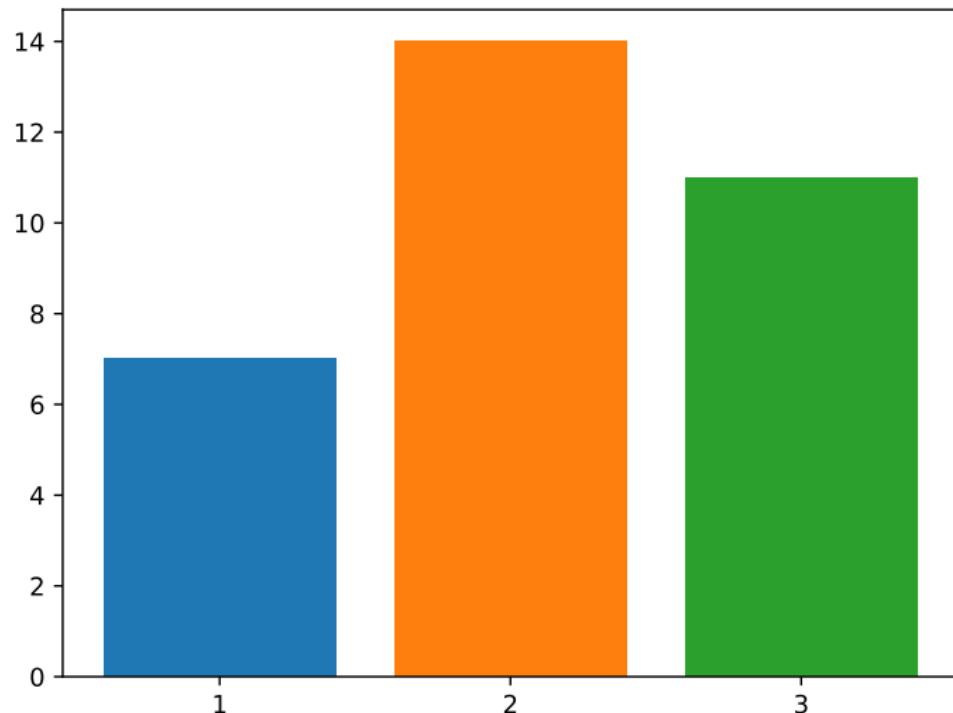
Descriptive statistics provide a concise summary of the data. The summary may either be numerical or graphical or both. Descriptive statistics may consist of, for example, numerical tables, average values, deviations, summaries and visualizations.

Statistical inference draws conclusions about the population using data. Statistical inference is based on mathematical modeling and probabilities. Inferential statistical analysis includes, for example, estimation and statistical testing.

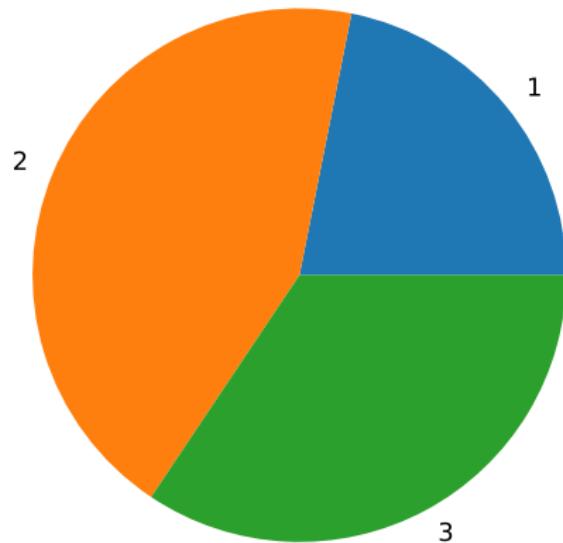
Visualization

- Discrete variable: bar plot, pie chart
- Continuous variable: box plot, histogram
- Bivariate: scatter plot

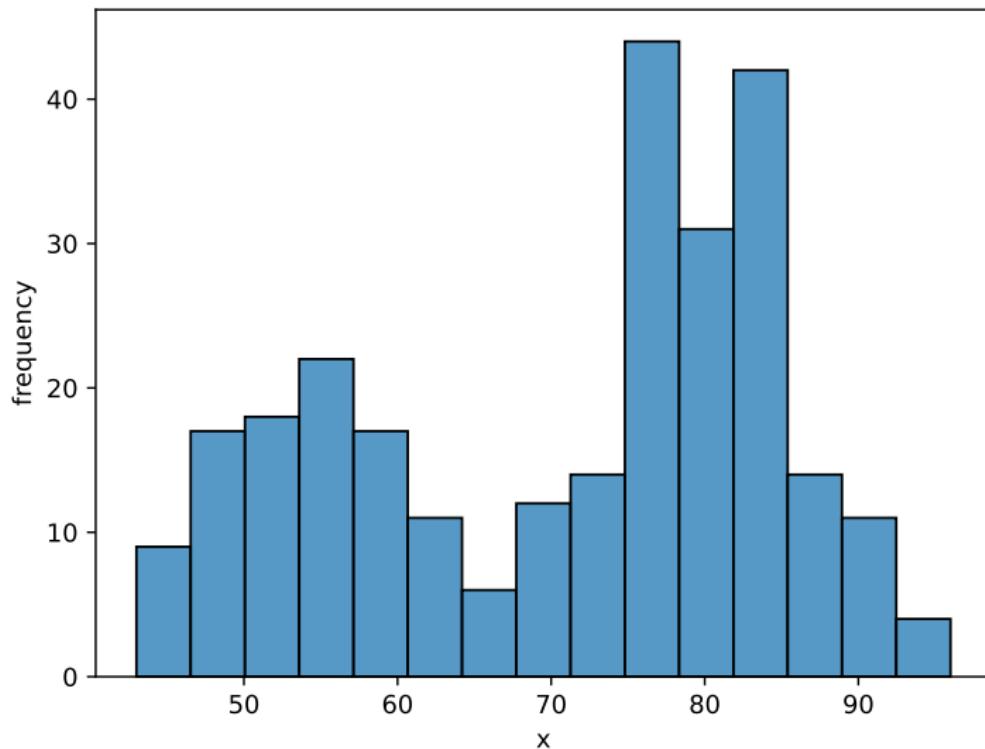
Bar plot



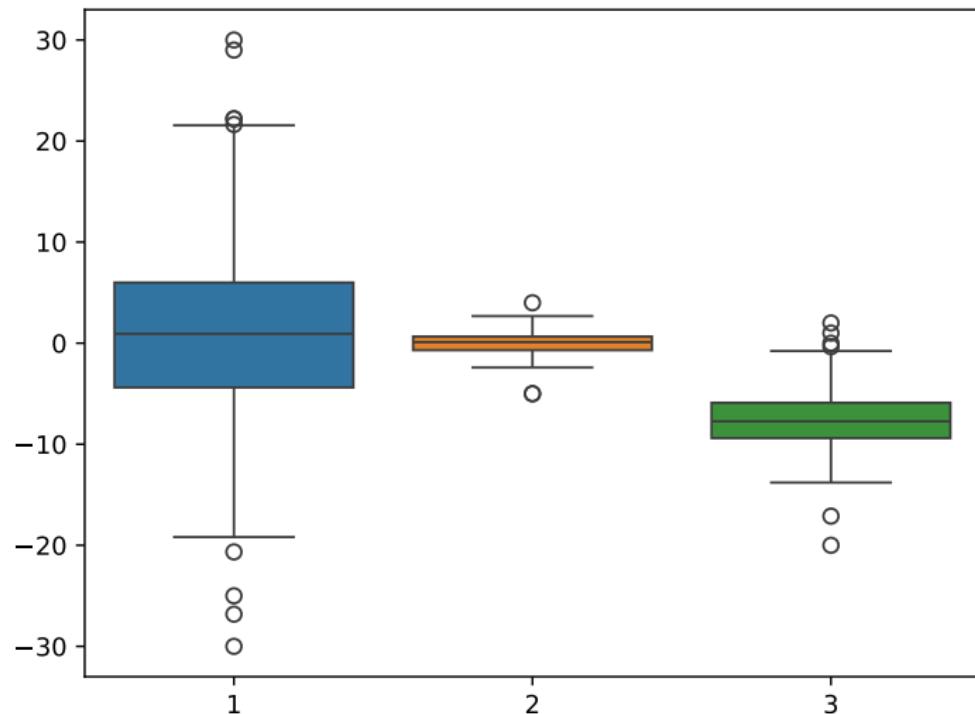
Pie chart



Histogram



Box plot



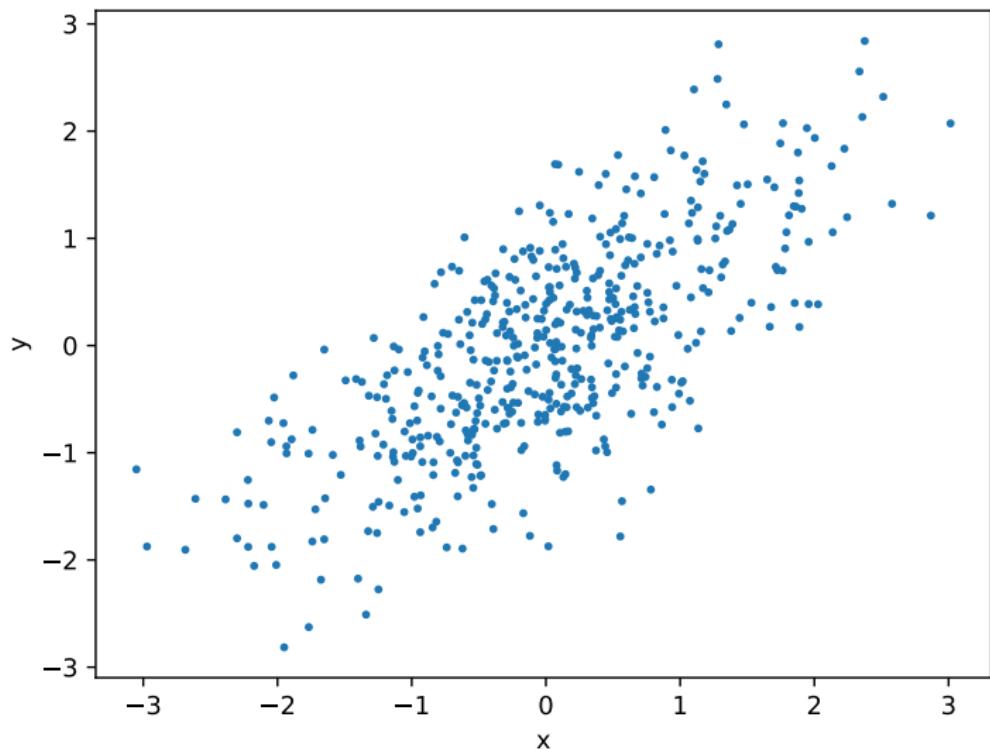
Box plot

In a box plot (sometimes also called a “box-and-whisker plot”), the box contains 50% of the data. The line in the middle is the sample median.

Let Q_1 and Q_3 denote the 25 and 75 sample percentiles. By default, the lower whisker is at the lowest data point above $Q_1 - 1.5(Q_3 - Q_1)$ and the upper whisker is at the highest data point below $Q_3 + 1.5(Q_3 - Q_1)$.

Outlying points are marked using circles.

Scatter plot



Location

Mean, median, and mode are commonly used measures of location.

Let x_1, \dots, x_n be i.i.d. observations of a random variable x . Then the sample mean

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

estimates the expected value $\mathbb{E}[x] = \mu$ of the variable x .

The population median m_x of a random variable x is the value with the property

$$\mathbb{P}(x < m_x) \leq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(x \leq m_x) \geq \frac{1}{2}.$$

Let $y_1 < y_2 < \dots < y_n$ be the ordered values of the data. The sample median is the middle value of the ordered values. If the number of observations is even, then the sample median is the average of the two middle elements. The sample median estimates the population median.

The sample mode is the value x_1, \dots, x_n that has the highest frequency. Mode estimates a value of a qualitative variable or discrete quantitative variable that has the highest probability.

Percentiles

Let x_1, \dots, x_n be i.i.d. observations of a random variable x . Let $y_1 < y_2 < \dots < y_n$ be the *ordered* values of the data. Then the **sample β percentile**, $0 < \beta < 100$, is the data point y_k , where k is the closest integer that is larger than or equal to $\beta \cdot (n/100)$. The **population β percentile** of a random variable x is the value β_x with the property

$$\mathbb{P}(x < \beta_x) \leq \frac{\beta}{100} \quad \text{and} \quad \mathbb{P}(x \leq \beta_x) \geq \frac{\beta}{100}.$$

Numerical example

Consider the sample

$$\{3, 1, 2, 3, 7, 8, 3, 4, 4, 6\}.$$

The sample mean is

$$\bar{x} = \frac{1}{10} \cdot (3 + 1 + 2 + 3 + 7 + 8 + 3 + 4 + 4 + 6) = \frac{41}{10} = 4.1.$$

The sample median is

$$\hat{m}_x = \frac{3+4}{2} = \frac{7}{2} = 3.5.$$

The sample mode is 3.

Deviation/scatter

Variance, standard deviation, median absolute deviation (MAD), and range are commonly used measures of deviation/scatter.

Let x_1, \dots, x_n be i.i.d. observations of a random variable x . The **sample variance**

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

estimates the population variance $\mathbb{E}[(x - \mathbb{E}[x])^2] = \sigma^2$.

The **sample standard deviation** is the square root of the sample variance:

$$s = s_n = \sqrt{s_n^2}.$$

Chebyshev's inequality

Let x be a random variable with finite expected value $\mathbb{E}[x] = \mu$ and finite variance $\mathbb{E}[(x - \mathbb{E}[x])^2] = \sigma^2$. Let $k > 1$. Then

$$\mathbb{P}(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

If $k = 2$, then $1 - \frac{1}{k^2} = 75\%$.

If $k = 3$, then $1 - \frac{1}{k^3} \approx 88.9\%$.

In practice, the expected value and variance must be estimated.

Chebyshev's inequality can be used to evaluate the outlyingness/rareness of a single observation:

- If an observation lies further away than two times the standard deviation of the sample mean, it is considered **rare**.
- If an observation lies further away than three times the standard deviation of the sample mean, it is considered **very rare**.

These definitions are based on Chebyshev's inequality.

Rare observation under normality

If it is known that observations follow a Gaussian distribution, then the probability for a data point lying within one standard deviation of the sample mean is $\approx 68\%$. The probability for a data point lying within two standard deviations of the sample mean is $\approx 95\%$ and for three standard deviations it is $\approx 99.7\%$.

Median absolute deviation and range

Let x_1, \dots, x_n be i.i.d. observations of a random variable x and let m_x be the sample median. Then the **median absolute deviation (MAD)** is the median of the sample $|x_1 - m_x|, |x_2 - m_x|, \dots, |x_n - m_x|$.[†]

Let Max_x be the largest data point and Min_x the smallest data point. Then the sample range is the interval $[\text{Min}_x, \text{Max}_x]$ and the length of the range is $\text{Max}_x - \text{Min}_x$.

[†]To make the MAD comparable with the standard deviation, one often multiplies the MAD with a scale factor k depending on the distribution. For example, for normally distributed data, $k = \frac{1}{\phi^{-1}(3/4)} \approx 1.4826$. (In fact, this is the default scaling used in R, but for example the `scipy.stats.median_abs_deviation` function uses $k = 1$ by default.)

Numerical example

Consider the sample

$$\{3, 1, 2, 3, 7, 8, 3, 4, 4, 6\}.$$

The sample mean was calculated above and it was 4.1. The sample variance is

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - 4.1)^2 = 4.9888\dots$$

and the sample standard deviation is $s_n = \sqrt{s_n^2} = \sqrt{4.9888\dots} = 2.233\dots$

The sample median was calculated above and it was 3.5. Mean absolute deviation:

$$\begin{aligned}MAD &= \text{median}\{|3 - 3.5|, |1 - 3.5|, |2 - 3.5|, |3 - 3.5|, |7 - 3.5|, \\&\quad |8 - 3.5|, |3 - 3.5|, |4 - 3.5|, |4 - 3.5|, |6 - 3.5|\}\\&= 1.\end{aligned}$$

The range can be calculated from the minimum and maximum values of the sample:

$$[\min(x), \max(x)] = [1, 8].$$

The length of the range is $8 - 1 = 7$.

Skewness

Let x_1, \dots, x_n be i.i.d. observations of a random variable x . Then the **sample skewness coefficient** is

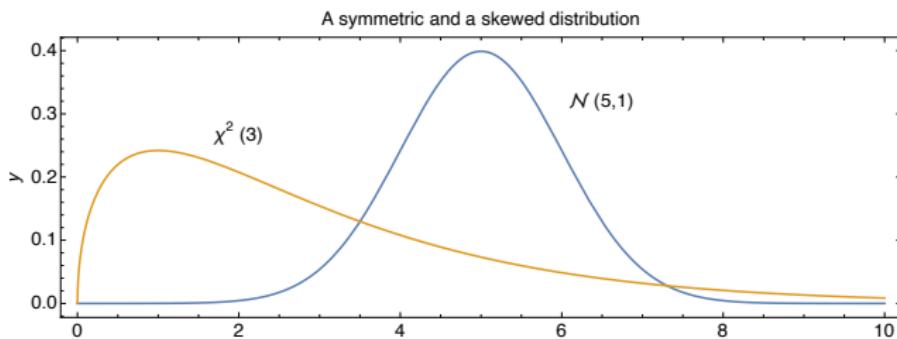
$$v = \frac{m_3}{s_n^3},$$

where

$$m_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3.$$

Sample skewness coefficient estimates the population value

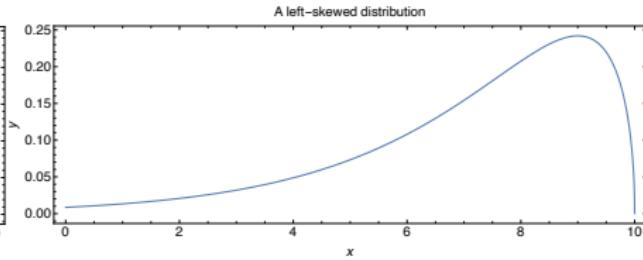
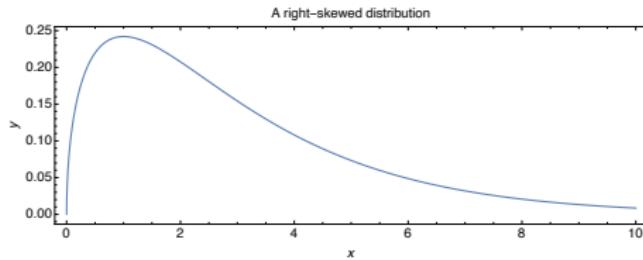
$$\mathbb{E}\left[\left(\frac{x - \mu}{\sigma}\right)^3\right]$$



Skewness

- If the skewness coefficient $\nu > 0$, then the distribution is skewed to the right (**positively skewed distribution**).
- If the skewness coefficient $\nu < 0$, then the distribution is skewed to the left (**negatively skewed distribution**).

Usually[†], a positively (right) skewed distribution has a long right tail and the mass of the distribution is concentrated on the left. A negatively (left) skewed distribution has a long left tail and the mass of the distribution is concentrated on the right.



[†]Multimodal distributions or asymmetric distributions which have one long tail but the other tail is fat can break this rule of thumb.

Skewness

Alternative skewness coefficient v_2 : Let x_1, \dots, x_n be i.i.d. observations of a random variable x . Then also

$$v_2 = \frac{\bar{x} - m_x}{s_n}$$

is a measure of skewness. (Here, m_x denotes the sample median.)

For symmetric distributions, the sample mean and sample median estimate the same population value.

Kurtosis

Let x_1, \dots, x_n be i.i.d. observations of a random variable x . Then the sample kurtosis coefficient is

$$k = \frac{m_4}{s_n^4} - 3,$$

where

$$m_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

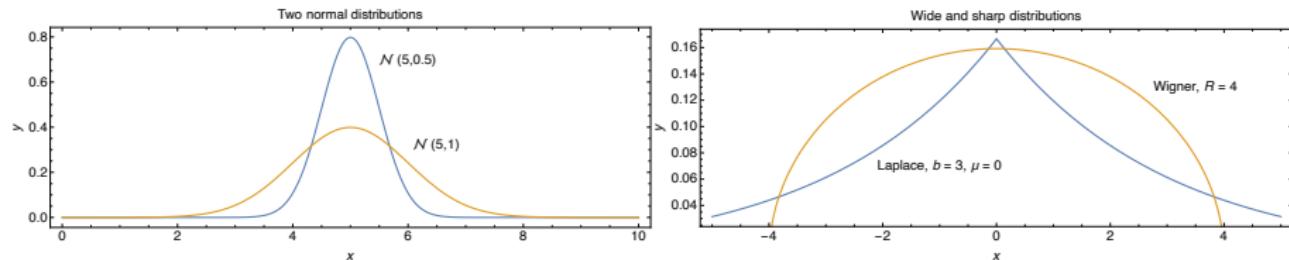
The sample kurtosis coefficient estimates the population value

$$\mathbb{E}\left[\left(\frac{x - \mu}{\sigma}\right)^4 - 3\right].$$

Kurtosis

A random variable with normal distribution has kurtosis value 0. If the kurtosis value is $k > 0$, then the distribution is more peaked than normal distribution. If $k < 0$, then the distribution is less peaked than normal distribution.

A distribution with large kurtosis value (**leptokurtic**) typically has a sharp peak and thick tails, while less peaked distributions (**platykurtic**) have round peaks and thin tails.



Linear dependence and correlation

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . Then the **sample covariance**

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

estimates the population covariance $\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \sigma_{xy}$, and

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

estimates the **Pearson correlation coefficient**

$$\rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

The Pearson correlation coefficient measures numerically the **linear dependence** of two random variables. The coefficient is always in the interval $[-1, 1]$.

Confidence interval

Confidence interval

In statistics, we often have a sample and we estimate the value of some parameter using the observations. For example, we estimate the expected value by calculating the sample mean or we estimate the population skewness coefficient by calculating the corresponding sample estimate. The simple estimate, however, still gives us quite little information. We cannot directly evaluate how good our estimate is. It would be nice to know a bit more. That is why an estimate of a parameter is often presented with a corresponding confidence interval.

Confidence interval

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. A confidence level for a confidence interval determines the probability that the confidence interval produced will contain the true parameter value.

Confidence interval

Let x be a random variable from a distribution P_x . Let θ be a parameter of the distribution P_x and let $\hat{\theta}$ be an estimate of the parameter. (For example, θ could be the population mean, population standard deviation, population median, etc., and $\hat{\theta}$ would be the corresponding sample mean, sample standard deviation, sample median, etc.)

We say that an interval (l, u) is a confidence interval for the estimate $\hat{\theta}$ at confidence level $(1 - \alpha)$ if the following holds: *before* the sample is generated, the **random** range (l, u) corresponding to $\hat{\theta}$ includes the **true parameter value** θ with probability $p = 1 - \alpha$.

After the sample has been generated and the estimate $\hat{\theta}$ and the corresponding confidence interval (l, u) has been calculated, the confidence interval either includes or does not include the true parameter value θ . If 100 samples are generated, the corresponding 100 estimates $\hat{\theta}$ and the corresponding 100 confidence intervals are calculated, then $\approx (1 - \alpha) \cdot 100$ of the confidence intervals include the true parameter value and $\approx \alpha \cdot 100$ do not include it.

Bootstrap confidence intervals

Let $\{x_1, \dots, x_n\}$ denote i.i.d. observations from the distribution P_x . Let θ be a parameter of the distribution P_x . (For example, θ could be the population mean, population standard deviation, population median, etc.) Let $\hat{\theta}$ be an estimate of the parameter θ calculated from the sample $\{x_1, \dots, x_n\}$. (For example, $\hat{\theta}$ would be the sample mean, sample standard deviation, sample median, etc., corresponding to θ .)

An estimate for the confidence interval (l, u) can now be obtained by resampling as follows:

1. Select n data points randomly with replacement from the original sample x_1, \dots, x_n . Each data point can be selected once, multiple times, or not at all. (Note that the sample size of the new sample is the same as the sample size of the original sample.)
2. Calculate a new estimate for the parameter θ from the new sample formed in the previous step.

(Continued on the next slide.)

3. Repeat the steps 1–2 k times and order the obtained estimates from the smallest to the largest. Include also the original estimate $\hat{\theta}$.
4. Calculate an estimate for a $(1 - \alpha) \cdot 100\%$ confidence interval by selecting a lower bound l that is smaller than (or equal to) $(1 - \frac{\alpha}{2}) \cdot 100\%$ of the ordered estimates and an upper bound u that is larger than (or equal to) $(1 - \frac{\alpha}{2}) \cdot 100\%$ of the estimates.

Example

Assume that we compute 999 bootstrap estimates. Then, in total, there are 1000 estimates – the original one and the 999 new ones. Now, an estimated 90% confidence interval (l, u) is obtained by choosing the 50th ordered estimate as l and the 951st estimate as u .

An estimate for the 95% confidence interval (l, u) is obtained by choosing the 25th estimate as l and the 976th estimate as u .

On the accuracy of the bootstrap confidence interval:

- The larger the original sample size, the better the confidence interval.
- The larger the number k of bootstrap samples, the better the confidence interval.

Exact confidence intervals

Bootstrap confidence intervals are nowadays easy to calculate and they have the advantage of being distribution free.

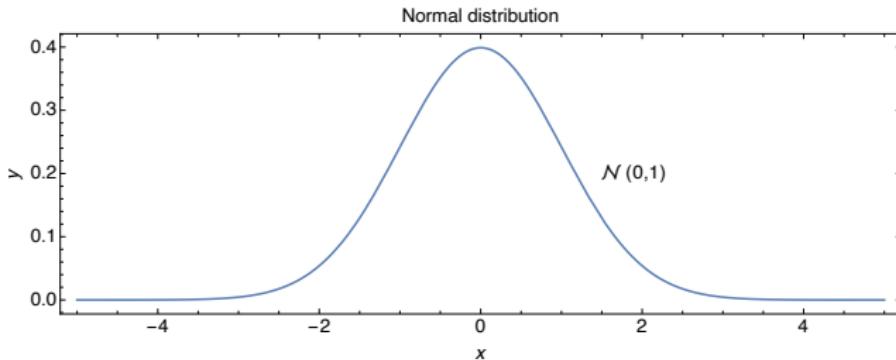
However, when the type of distribution is known, also exact confidence intervals can be calculated. It is possible to obtain exact confidence intervals for the parameters of the normal distribution or for the parameter of the Bernoulli distribution, for example.

Confidence interval, normal distribution

A random variable with normal distribution has a probability density function (PDF) of the form

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

The normal distribution has two parameters: the mean μ and the variance σ^2 .



Example (Confidence interval for population mean μ of a normal i.i.d. sample with *known variance* σ^2)

Let x_1, \dots, x_n be i.i.d. copies of $x \sim \mathcal{N}(\mu, \sigma^2)$. Suppose that we are interested in finding a level $(1 - \alpha)$ confidence interval for the population mean μ given the sample x_1, \dots, x_n . If we know the population variance σ^2 , then we can use the whitening transform

$$Z = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad (1)$$

and deduce that the $(1 - \alpha)$ confidence interval for the population mean is given by

$$\left(\bar{x}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right), \quad (2)$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ is the $(1 - \alpha/2) \cdot 100$ percentile of the standard normal distribution. E.g., if $\alpha = 0.05$, then $z_{0.025} = \Phi^{-1}(0.975) \approx 1.96$.

In practice, the population standard derivation must be *approximated* by the sample standard deviation s_n . If n is large (e.g., $n > 30$), then simply approximating $\sigma \approx s_n$ in (1)–(2) may lead to a reasonable approximation of the CI. However, simply replacing σ by s_n makes the test statistic (1) non-Gaussian in general. A better method is to note that $\frac{\bar{x}_n - \mu}{s_n/\sqrt{n}}$ follows Student's *t-distribution*.

Confidence interval, mean of normal distribution

Let x_1, \dots, x_n be i.i.d. copies of $x \sim \mathcal{N}(\mu, \sigma^2)$. We are interested in finding a level $(1 - \alpha)$ confidence interval for the population mean μ given the sample x_1, \dots, x_n . In practice, the population standard deviation σ must be approximated by the sample standard deviation s_n . Substituting the population standard deviation σ by the sample standard deviation s_n in (1) yields the t -statistic

$$t_{n-1} := \frac{\bar{x}_n - \mu}{s_n / \sqrt{n}}$$

and we say that t_{n-1} has Student's t -distribution with $n - 1$ degrees of freedom. Then the $(1 - \alpha)$ confidence interval for the population mean μ is given by

$$\left(\bar{x}_n - t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1, \alpha/2} \frac{s_n}{\sqrt{n}} \right),$$

where $t_{n-1, \alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the t_{n-1} distribution. E.g., if $n = 10$ and $\alpha = 0.05$, then $t_{9, 0.025} = F_{t_9}^{-1}(0.975) = 2.262$, where $F_{t_9}^{-1}$ is the quantile function of t_9 .

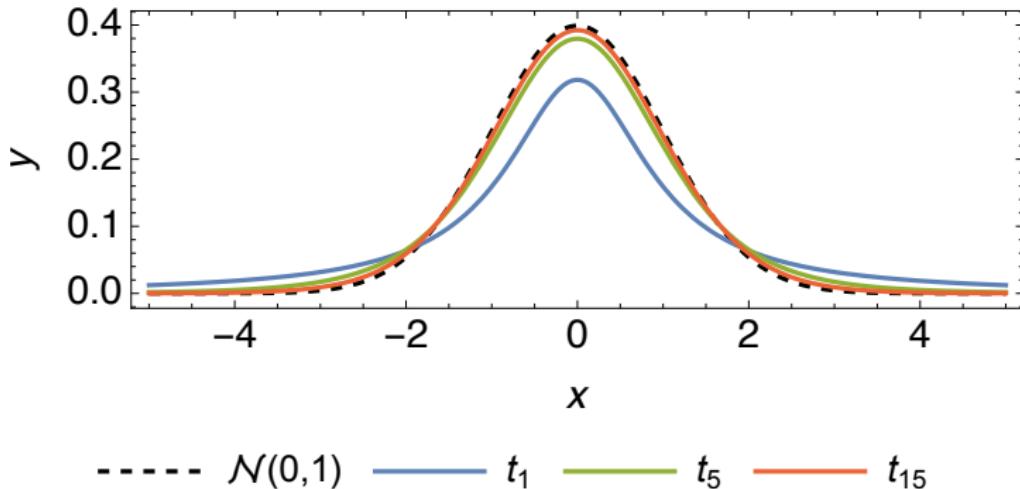


Figure: Student's t -distributions with different degrees of freedom. The t -distribution has heavier tails than the standard Gaussian distribution. As the degrees of freedom increase, the t -distributions tend to the standard Gaussian distribution.

Confidence interval, variance of normal distribution

Let x_1, \dots, x_n be i.i.d. copies of $x \sim \mathcal{N}(\mu, \sigma^2)$. We are interested in finding a level $(1 - \alpha)$ confidence interval for the **population variance** given the sample x_1, \dots, x_n . It is assumed that the population mean μ is also unknown. The statistic

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

has the χ^2 distribution with $n-1$ degrees of freedom, i.e., $Q \sim \chi^2(n-1)$. Then the level $(1 - \alpha)$ confidence interval for the variance of a normal distribution can be given as

$$\left(\frac{(n-1)s_n^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)s_n^2}{\chi_{n-1,1-\alpha/2}^2} \right),$$

where $\chi_{n-1,\alpha/2}^2$ is the $(1 - \alpha/2) \cdot 100$ percentile of the $\chi^2(n-1)$ distribution. Similarly, $\chi_{n-1,1-\alpha/2}^2$ is the $(\alpha/2) \cdot 100$ percentile of the $\chi^2(n-1)$ distribution. E.g., if $n = 10$ and $\alpha = 0.05$, then $\chi_{9,0.025}^2 = F_{\chi^2(9)}^{-1}(0.975) \approx 19.02$ and $\chi_{9,0.975}^2 = F_{\chi^2(9)}^{-1}(0.025) \approx 2.70$.

χ^2 distribution

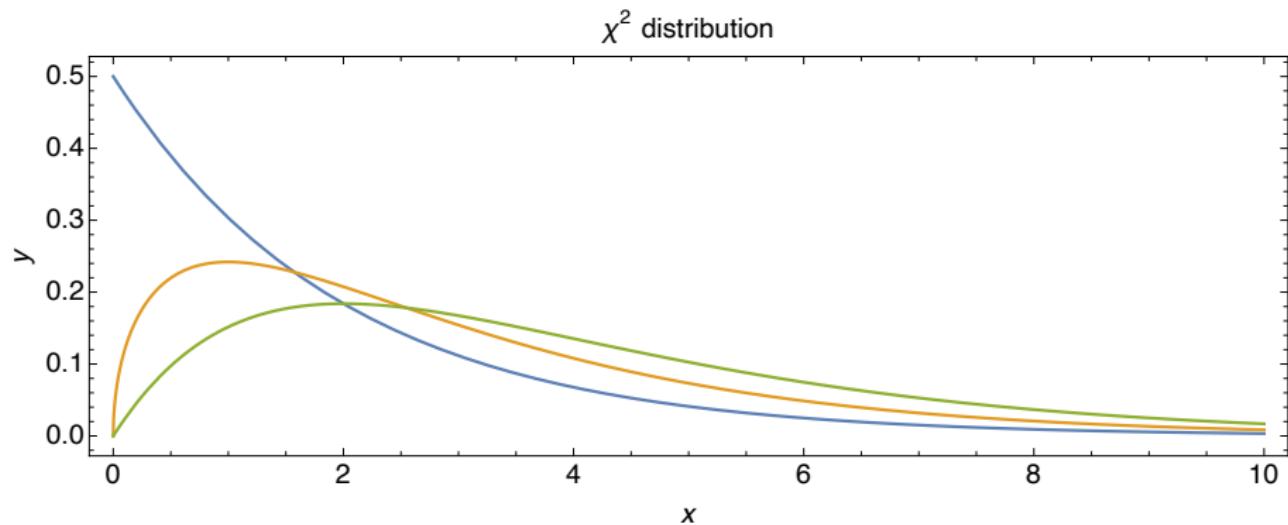


Figure: χ^2 distribution with different degrees of freedom.

Confidence interval, parameter p of Bernoulli distribution

Let $\{x_1, \dots, x_n\}$ denote i.i.d. observations of a random variable x . Assume that $\mathbb{P}(x_i = 1) = p$ and $\mathbb{P}(x_i = 0) = 1 - p$. Then $x \sim \text{Ber}(p)$, with expected value $\mathbb{E}[x] = p$ and $\mathbb{E}[(x - \mathbb{E}[x])^2] = p(1 - p)$. An unbiased estimate of the expected value p is the sample mean

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i.$$

If n is large, the level $(1 - \alpha)$ confidence interval for the **mean p** of the Bernoulli distribution can be given as

$$\left(\hat{p} - z_{\alpha/2} \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}, \hat{p} + z_{\alpha/2} \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}} \right),$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the standard normal distribution $\mathcal{N}(0, 1)$.

There exist several alternative estimates for the confidence interval for the mean of the Bernoulli distribution. If the sample size is small, one can try the Wilson score interval, for example.

Numerical example, confidence intervals

The masses of Brand X cookie packages are approximately normally distributed with expected value μ . The randomly chosen packages were weighed and the following data (measured in grams) was obtained: 397.3, 399.6, 401.0, 392.9, 396.8, 400.0, 397.6, 392.1, 400.8, 400.6.

The mean of the masses is 397.87g and the sample standard deviation is

$$s = \sqrt{\frac{1}{10 - 1} \sum_{i=1}^{10} (x_i - 397.87)^2} \approx 3.2128.$$

As we saw above, the 97.5% percentile of the Student's t -distribution with $10 - 1 = 9$ degrees of freedom is $t = 2.262$. The 95% confidence interval for the mean masses of the cookie packages is

$$(\bar{x} \pm t \frac{s}{\sqrt{n}}) = (397.87g \pm 2.262 \cdot \frac{3.2128g}{\sqrt{10}}) = (395.6g, 400.3g).$$

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Seventh lecture, November 25, 2024

Hypothesis testing

Hypothesis testing

Statistical tests are applied extensively in various fields of science. We might want to test, for example:

- If one concrete type is stronger than another (competing) concrete type.
 - If there is a difference in the average salaries of men and women across the population.
 - Whether or not a new medicine lowers systolic blood pressure.
- ⋮

Hypothesis testing

A statistical hypothesis is a hypothesis that is tested using probabilities. Statistical testing is based on setting general statistical assumptions, a null hypothesis and an alternative hypothesis, and on selecting a suitable test statistic. The value of the selected test statistic is calculated from a sample of observations.

Assumptions

- General statistical assumptions include assumptions about the population, sampling method, and about the distribution of the observations.
- Statistical assumptions hold throughout the testing process.
- Statistical assumptions may, and should, be tested separately.

Null hypothesis

- The statement about a population parameter that is being tested is called the **null hypothesis** H_0 .
- The null hypothesis is assumed to be true, unless there is strong evidence that indicates otherwise.
- If strong evidence against the null hypothesis is found, then it is rejected.
- In simple statistical tests, the null hypothesis can often be stated as

$$H_0 : \theta = \theta_0,$$

where θ is the parameter being tested and θ_0 is a fixed value of the parameter.

- The null hypothesis is often of the form “is the same” or “no difference”.

Alternative hypothesis

- If the null hypothesis H_0 is rejected, then the **alternative hypothesis** H_1 is accepted.
- If the alternative hypothesis can be stated as $H_1: \theta > \theta_0$ or $H_1: \theta < \theta_0$, then it is called a **one tailed alternative hypothesis**.
- If the alternative hypothesis can be stated as $H_1: \theta \neq \theta_0$, then it is called a **two tailed alternative hypothesis**.
- The alternative hypothesis is often of the form “not the same” or “different”.

It is not always easy to decide whether one tailed or two tailed alternative hypothesis should be used.

Do not fish for favorable results by using one tailed alternative hypothesis!

The use of one tailed alternative hypothesis must be justified by the context.

Test statistic

- A test statistic compares the observations and the null hypothesis H_0 .
- A test statistic is a random variable and its value depends on the observations.
- A test statistic is used in evaluating the probability of getting the observed value of the statistic, under the assumption that the null hypothesis H_0 is true.
- The distribution of the test statistic under the null hypothesis H_0 must be known for comparing the observations and the null hypothesis H_0 .

Critical value

- The expected value of a chosen test statistic is calculated under the null hypothesis H_0 .
- If the observed value of the test statistic is close to the expected value, no strong argument against the null hypothesis H_0 is found.
- If the observed value of the test statistic is far away from the expected value, then evidence against the null hypothesis H_0 is found.
- The set of values of the test statistic for which the null hypothesis is rejected (i.e., the set of the values that are far away from the expected value) is called the **critical region**.
- The threshold values defining the critical region are called the **critical values**.

p-value

The *p*-value of a statistical test is the probability, assuming that the null hypothesis H_0 is true, of observing at least as extreme value as the observed value of the test statistic.

Rejecting or not rejecting the null hypothesis H_0 is based on the *p*-value. Statistical software can be used to calculate the *p*-value.

The significance level α of a test statistic is the smallest *p*-value that is accepted without rejecting the null hypothesis H_0 . It is possible to use pre-selected significance levels and the corresponding critical regions. Commonly used significance levels α are 0.05, 0.1, 0.01, and 0.001.

If the significance level is $\alpha = 0.05$ and the *p*-value of the test statistic is < 0.05 , then the null hypothesis H_0 is rejected.

p-value

The *p*-value of a test statistic is calculated as follows:

- ① Calculate the value of the test statistic using the observations.
- ② Assuming that the null hypothesis H_0 is true and based on the known distribution of the test statistic, calculate the probability of the value of the test statistic being as extreme, or more extreme, as it is.

The null hypothesis H_0 can be rejected, if the *p*-value is small enough.

The smaller the *p*-value, the stronger the evidence against H_0 .

Errors

There are two types of errors related to the rejection of the null hypothesis H_0 :

- **Type 1 error:** True null hypothesis is rejected.
- **Type 2 error:** False null hypothesis is not rejected.

The **type 1 error rate** is the probability of rejecting the null hypothesis given that it is true. Thus type 1 error rate is equal to the significance level α .

The **type 2 error rate** is the probability of not rejecting the null hypothesis given that it is false. Type 2 error rate is in general a function of the possible distributions, often determined by a parameter, under the alternative hypothesis. The **power of a test statistic** is equal to $1 - (\text{type 2 error rate})$. Thus, the power of a test statistic is also a function of the possible distributions. As the power increases, the chance of a type 2 error decreases – one is more likely to detect significant differences when they truly exist.

In statistical testing, type 1 errors are generally considered worse than type 2 errors. That is why the significance level α is usually selected to be small.

p-value, one tailed and two tailed alternative hypothesis

Let z be the value of a test statistic Z calculated from the observations.

If the one tailed alternative hypothesis is given as $H_1: \theta > \theta_0$, then the *p*-value of the test is

$$p = \mathbb{P}(Z \geq z | H_0).$$

If the one tailed alternative hypothesis is given as $H_1: \theta < \theta_0$, then the *p*-value of the test is

$$p = \mathbb{P}(Z \leq z | H_0).$$

If the alternative hypothesis is two tailed, $H_1: \theta \neq \theta_0$, then the *p*-value of the test is

$$p = 2 \min(\mathbb{P}(Z \leq z | H_0), \mathbb{P}(Z \geq z | H_0)).$$

Steps of statistical hypothesis testing

- ① State the hypotheses and general assumptions.
- ② Select a test statistic.
- ③ Pick a sample such that the general assumptions hold.
- ④ Calculate the value of the test statistic using the sample.
- ⑤ Calculate the p -value corresponding to the observed value of the test statistic.
- ⑥ Draw conclusions and reject/do not reject the null hypothesis.

t-tests

One sample t -test

The **one sample t -test** compares the expected value of a random variable to a given constant.

Let x_1, \dots, x_n be i.i.d. observations of a random variable x . Assume that the observed values come from the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

The null hypothesis: $H_0: \mu = \mu_0$.

The possible alternative hypotheses:

$$H_1: \mu > \mu_0 \text{ (one tailed),}$$

$$H_1: \mu < \mu_0 \text{ (one tailed),}$$

$$H_1: \mu \neq \mu_0 \text{ (two tailed).}$$

One sample t -test

- The t -test statistic is

$$t = \frac{\bar{x} - \mu_0}{s_n / \sqrt{n}}.$$

- If the null hypothesis H_0 is true, then the test statistic follows Student's t -distribution with $n - 1$ degrees of freedom.
- The expected value of the test statistic under the null hypothesis H_0 is 0, i.e., $\mathbb{E}[t] = 0$.
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found. On the other hand, the null hypothesis H_0 is rejected if the p -value is small enough.
- Python:

```
t_stat,p_value = scipy.stats.ttest_1samp(a=x, popmean=μ₀)
```

One sample t -test, normality assumption

- When the one-sample t -test is used, it is assumed that the observations follow the normal distribution.
- If the sample size is large, then one sample t -test is not very sensitive to moderate deviations from normality.
- Even without normality, the one sample t -test is quite reliable if the sample size $n > 25$. That is, unless the distribution is very skewed.
- With sample size $n > 40$, the one sample t -test is quite reliable even for clearly skewed distributions.

One sample t -test – implementation in Python

```
import numpy as np
from scipy.stats import t as tdist

def tTest_1sample(x,mu0,alternative='two-sided'):
    n = len(x)
    xbar = np.mean(x)
    std = np.std(x,ddof=1) # Use Bessel's correction
    t_stat = (xbar-mu0)/(std/np.sqrt(n))
    q = tdist.cdf(t_stat,n-1)
    if alternative == 'less':
        return t_stat,q
    elif alternative == 'greater':
        return t_stat,1-q
    else:
        return t_stat,2*min(q,1-q)
```

Numerical example, one sample t -test

According to the package text, Brand X cookies have 12 chocolate chips in each cookie. The number of chocolate chips of ten randomly selected cookies were calculated and the following data was obtained:

$$\{12, 11, 10, 13, 14, 12, 11, 12, 12, 12\}.$$

We want to test, on significance level 5%, the hypothesis that the expected value of the number of chocolate chips in Brand X cookies is 12.

The sample mean of the chocolate chips is 11.9 and the sample standard deviation is 1.1005. One sample t -test is used and the value of the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{11.9 - 12}{1.1005/\sqrt{10}} = -0.287.$$

Assuming normality and i.i.d. observations, under the null hypothesis ($\mu = \mu_0 = 12$), the test statistic follows Student's t -distribution with 9 degrees of freedom.

With significance level 5% and 9 degrees of freedom, the critical values of the test statistic are ± 2.262 . Since the observed value of the test statistic $-0.287 > -2.262$ and $-0.287 < 2.262$, the null hypothesis is not rejected.

The p -value is often observed directly without setting any pre-selected significance level. Probabilities $\mathbb{P}(T \leq t | H_0)$ and $\mathbb{P}(T \geq t | H_0)$ are 0.6098 and 0.3902, respectively. Then the p -value is

$$p = 2 \min(\mathbb{P}(Z \leq z | H_0), \mathbb{P}(Z \geq z | H_0)) = 2 \cdot 0.3902 = 0.7804.$$

The p -value is large and no evidence against the null hypothesis is found.

What went wrong in the previous example? What are the general statistical assumptions when one sample t -test is used?

Two sample t -test

The **two sample t -test** compares the expected values of two independent variables. We first consider the case when the variances are *not assumed to be equal*.

Two sample t -test, assumptions

Let x_1, \dots, x_n be the observed values of a random variable x and let y_1, \dots, y_m be the observed values of a random variable y . Assume that the observed values x_1, \dots, x_n are i.i.d. and come from the normal distribution $\mathcal{N}(\mu_x, \sigma_x^2)$ and assume that the observed values y_1, \dots, y_m are i.i.d. and come from the normal distribution $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, assume that x_i and y_j are independent for all i, j .

The null hypothesis: $H_0: \mu_x = \mu_y$.

The possible alternative hypotheses:

$$H_1: \mu_x > \mu_y \text{ one tailed,}$$

$$H_1: \mu_x < \mu_y \text{ one tailed,}$$

$$H_1: \mu_x \neq \mu_y \text{ two tailed.}$$

Two sample t -test

- The t -test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n + s_y^2/m}}.$$

- If the null hypothesis H_0 is true, then the test statistic follows Student's t -distribution with v degrees of freedom, where

$$v = \frac{(s_x^2/n + s_y^2/m)^2}{((s_x^2/n)^2/(n-1)) + ((s_y^2/m)^2/(m-1))}.$$

- The expected value of the test statistic under the null hypothesis H_0 is 0 ($\mathbb{E}[t] = 0$).
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- Python:

```
t_stat,p_value =  
scipy.stats.ttest_ind(a=x,b=y,equal_var=False)
```

Two sample t -test – implementation Python

```
import numpy as np
from scipy.stats import t as tdist

def tTest_2sample(x,y,alternative='two-sided'):
    n = len(x); m = len(y)
    xbar = np.mean(x); ybar = np.mean(y)
    stdx = np.std(x,ddof=1); stdy = np.std(y,ddof=1)
    t_stat = (xbar-ybar)/np.sqrt(stdx**2/n+stdy**2/m)
    v = (stdx**2/n+stdy**2/m)**2 \
        /((stdx**2/n)**2/(n-1)+((stdy**2/m)**2/(m-1)))
    q = tdist.cdf(t_stat,v)
    if alternative == 'less':
        return t_stat,q
    elif alternative == 'greater':
        return t_stat,1-q
    else:
        return t_stat,2*min(q,1-q)
```

Two sample t -test, normality assumption

- When the two sample t -test is used, it is assumed that the observations follow the normal distribution.
- If the sample sizes are large, then the two sample t -test is not very sensitive to moderate deviations from normality.
- Even without normality, the two sample t -test is quite reliable, if the sample sizes $n > 25$ and $m > 25$. That is, unless the distributions are very skewed.
- If $n > 40$ and $m > 40$, then the test can be quite safely used even with clearly skewed distributions.

Two sample t -test, equal variances

The two sample t -test has a bit simpler form if the variances are assumed to be equal.

Assumptions and hypotheses are the same as in the general two sample t -test, but the variances of the distributions are assumed to be equal – that is, it is assumed that $\sigma_x^2 = \sigma_y^2$.

Two sample t -test, equal variances

- The t -test statistic

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{1/n + 1/m}},$$

where

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$

- If the null hypothesis H_0 is true, then the test statistic follows Student's t -distribution with $n+m-2$ degrees of freedom.
- The expected value of the test statistic under the null hypothesis H_0 is 0 ($\mathbb{E}[t] = 0$).
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- Normality assumption can be relaxed as in the general two sample t -test.
- Python:

```
t_stat,p_value =  
scipy.stats.ttest_ind(a=x,b=y,equal_var=True)
```

Paired *t*-test

General two sample *t*-tests can be applied when the two samples are *independent*.

The paired *t*-test can be used to **compare two measuring equipments** by using both equipments to measure the same subject in the same circumstances. (Do two pedometers give the same result?) A paired *t*-test can be used for example to **study if a treatment works** by measuring the same subjects before and after the treatment. (Does drinking have an effect on reaction time? Does malnutrition have an effect on memory?) The aim can also be to **compare two populations** by measuring the same variables of fitted pairs. (Do the voting preferences of couples living together differ from each other?)

Paired t -test

Paired t -test:

- Observations $(x_{i,1}, x_{i,2})$, $i = 1, \dots, n$, consist of measured pairs of a random variable x .
- The pairs are assumed to be independent. However, the two values inside one pair are not assumed to be independent.
- General two sample t -tests should not be used for paired observations.
- Calculate the differences $d_i = x_{i,1} - x_{i,2}$, $i = 1, \dots, n$, of the measurements $x_{i,1}$ and $x_{i,2}$.
- Measurements $x_{i,1}$ and $x_{i,2}$ have on average about the same value if the differences are on average about 0.
- It is now possible to apply the standard one sample t -test to the differences d_i .
- Python:
`t_stat,p_value = scipy.stats.ttest_rel(a=x,b=y)`

Paired t -test

- General statistical assumptions: differences d_i are i.i.d. and come from the normal distribution.
- The null hypothesis: $H_0: \mu_d = 0$.
- Possible alternative hypotheses: $H_1: \mu_d > 0$ (one tailed), $H_1: \mu_d < 0$ (one tailed) or $H_1: \mu_d \neq 0$ (two tailed).
- The t -test statistic

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}.$$

- If the null hypothesis H_0 is true, then the test statistic follows Student's t -distribution with $n - 1$ degrees of freedom.
- The expected value of the test statistic under the null hypothesis H_0 is 0 ($\mathbb{E}[t] = 0$).
- If the value of the test statistic is large/small, evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- The normality assumption can be relaxed as in the general one sample t -test.

Paired t -test – implementation in Python

```
def tTest_paired(x,y,alternative='two-sided'):
    return tTest_1sample(x-y,0,alternative)
```

Variance tests

Variance test, assumptions

Let x_1, \dots, x_n be observed values of a random variable x . Assume that the observed values are i.i.d. and come from the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

The null hypothesis: $H_0: \sigma^2 = \sigma_0^2$.

The possible alternative hypotheses:

$$H_1: \sigma^2 > \sigma_0^2 \text{ (one tailed),}$$

$$H_1: \sigma^2 < \sigma_0^2 \text{ (one tailed),}$$

$$H_1: \sigma^2 \neq \sigma_0^2 \text{ (two tailed).}$$

Variance test

- The χ^2 test statistic

$$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2}.$$

- If the null hypothesis is true, then the test statistic follows the χ^2 distribution with $n - 1$ degrees of freedom.
- The expected value of the test statistic is $n - 1$.
- Large and small values of the test statistic (compared to the expected value $n - 1$) suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p -value is small enough.
- **This test is sensitive to deviations from normality!** Variance test does not work, not even with large sample sizes, if the distribution of the observations is skewed.

Variance test – implementation in Python

```
import numpy as np
from scipy.stats import chi2

def varTest(x,sigma_squared,alternative='two-sided'):
    n = len(x)
    Q_stat = (n-1) * np.var(x,ddof=1)/sigma_squared
    q = chi2.cdf(Q_stat,n-1)
    if alternative == 'less':
        return Q_stat,q # one-sided variance test
    elif alternative == 'greater':
        return Q_stat,1-q # one-sided variance test
    else:
        return Q_stat,2*min(q,1-q) # two-sided variance test
```

Variance comparison test, assumptions

Let x_1, \dots, x_n be observed values of a random variable x and let y_1, \dots, y_m be observed values of a random variable y . Assume that the observations x_1, \dots, x_n are i.i.d. and follow the normal distribution $\mathcal{N}(\mu_x, \sigma_x^2)$ and assume that y_1, \dots, y_m are i.i.d. and follow the normal distribution $\mathcal{N}(\mu_y, \sigma_y^2)$. Furthermore, assume also that x_i and y_j are independent for all i, j .

The null hypothesis: $H_0: \sigma_x^2 = \sigma_y^2$.

The possible alternative hypotheses:

$$H_1: \sigma_x^2 > \sigma_y^2 \text{ (one tailed),}$$

$$H_1: \sigma_x^2 < \sigma_y^2 \text{ (one tailed),}$$

$$H_1: \sigma_x^2 \neq \sigma_y^2 \text{ (two tailed).}$$

Variance comparison test

- The F -test statistic

$$F = \frac{s_x^2}{s_y^2}.$$

- If the null hypothesis is true, then the test statistic follows F -distribution with $n - 1$ and $m - 1$ degrees of freedom.
- The expected value of the test statistic is ≈ 1 .
- Large and small values of the test statistic (compared to the expected value ≈ 1) suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- This test is also sensitive to deviations from normality and does not work, not even with large sample sizes, if the distribution of the observations is skewed.

Variance comparison test – implementation in Python

```
import numpy as np
from scipy.stats import f as Fdist

def Ftest(x,y,alternative='two-sided'):
    dfx = len(x)-1
    dfy = len(y)-1
    F_stat = np.var(x,ddof=1)/np.var(y,ddof=1)
    q = Fdist.cdf(F_stat,dfx,dfy)
    if alternative == 'less':
        return F_stat,q
    elif alternative == 'greater':
        return F_stat,1-q
    else:
        return F_stat,2*min(q,1-q)
```

Hypothesis testing

Nonparametric (distribution free) statistical tests

Sign tests and rank tests

Parametric tests are usually preferred over non-parametric tests since they usually have more statistical power (= lower type 2 error rate) than non-parametric tests, given that the statistical assumptions are satisfied.

The advantage of sign tests and rank tests is that they do not require strong distributional assumptions. Sign tests and rank tests are suitable for continuous quantitative variables, but can also be used for any ordinal data.

Sign test

One sample sign test

The one sample sign test is applied in similar testing problems as the one sample t -test. However, the sign test requires milder distributional assumptions.

Let x_1, \dots, x_n be observed values of a continuous random variable x with population median m . Assume that the observed values are i.i.d.

The null hypothesis: $H_0: m = m_0$.

Possible alternative hypotheses:

$$H_1: m > m_0 \text{ (one tailed)},$$

$$H_1: m < m_0 \text{ (one tailed)},$$

$$H_1: m \neq m_0 \text{ (two tailed)}.$$

One sample sign test

- Calculate the differences $d_i = x_i - m_0$, $i = 1, \dots, n$.
- The test statistic S is the number of cases where $d_i > 0$.
(Alternatively, the number of cases where $d_i < 0$.)
- If the null hypothesis H_0 is true, then the test statistic follows the binomial distribution with parameters n and $1/2$.
- Under H_0 , the expected value of the test statistic is $\frac{1}{2}n$ and the variance is $\frac{1}{4}n$.
- Large and small values of the test statistic (compared to the expected value $\frac{1}{2}n$) suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p -value is small enough.

One sample sign test, p -value

The distribution of the test statistic S is tabulated and many softwares give exact p -values of the test.

Let s denote the observed value of the test statistic S . Then the p -value of the test is given as follows:

- If the alternative hypothesis is $H_1: m > m_0$, then the p -value is $p = \mathbb{P}(S \geq s)$.
- If the alternative hypothesis is $H_1: m < m_0$, then the p -value is $p = \mathbb{P}(S \leq s)$.
- If the alternative hypothesis is $H_1: m \neq m_0$, then the p -value is $p = 2 \min(\mathbb{P}(S \geq s), \mathbb{P}(S \leq s))$.

Naturally, the probabilities $\mathbb{P}(S \leq s)$ and $\mathbb{P}(S \geq s)$ are calculated under H_0 .

Remark. The sign test can also be used for discrete variables as well. Then it is possible that for some of the observations $d_i = x_i - m_0 = 0$. If the number of zeros is small compared to the sample size, these observations can be deleted and the sample size can be modified accordingly. If the number of zeros is large, then the zeros should be dealt with such that they are against rejecting the null hypothesis. For example: consider the two-tailed null hypothesis, 3 negative signs, 15 positive signs and 6 zeros. Now the test should be conducted as if there were 9 negative signs and 15 positive ones.

Python:

```
S_stat = sum(x-m0>0)
n = sum(i!=0 for i in x - m0)
#n = len(x) # if the number of zeros in x - m0 is large
p_value = scipy.stats.binomtest(S_stat,n,p=0.5).pvalue
```

One sample sign test – implementation in Python

```
import numpy as np
from scipy.stats import binom

def signTest_1sample(x,m0,alternative='two-sided'):
    diff = x-m0
    S_stat = sum(diff>0)
    n = sum(i!=0 for i in diff)
    #n = len(x) # if the number of zeros in  $x - m_0$  is large
    q = binom.cdf(S_stat,n,0.5)
    q2 = binom.pmf(S_stat,n,0.5)+1-q # (*)
    if alternative == 'less':
        return S_stat,q
    elif alternative == 'greater':
        return S_stat,q2
    else:
        return S_stat,2*min(q,q2)
# Note that in (*), we used  $P(S \geq s) = P(S=s) + 1 - P(S \leq s)$ 
```

Asymptotic one sample sign test

If the sample size is large, then under the null hypothesis H_0 , the standardized test statistic $Z = \frac{S-n/2}{\sqrt{n/4}}$ approximately follows the standard normal distribution.

The approximation is usually good enough if $n > 20$. For smaller samples, the test relies on the exact distribution of the test statistic S .

Paired sign test

The paired sign test is applied in similar testing problems as the paired t -test.

- The observations $(x_{i,1}, x_{i,2})$, $i = 1, \dots, n$, consist of measured pairs of a random variable x .
- The pairs are assumed to be independent. However, the two values inside one pair are not assumed to be independent.
- Calculate the differences $d_i = x_{i,1} - x_{i,2}$, $i = 1, \dots, n$, of the measurements $x_{i,1}$ and $x_{i,2}$.

Paired sign test

- General statistical assumptions: the differences d_i are i.i.d. and follow a distribution with median m .
- The null hypothesis $H_0: m = 0$.
- Possible alternative hypotheses: $H_1: m > 0$ (one tailed), $H_1: m < 0$ (one tailed) or $H_1: m \neq 0$ (two tailed).
- Now it is possible to apply the one sample sign test for the differences d_i .

Paired sign test – implementation in Python

```
def signTest_paired(x,y,alternative='two-sided'):
    return signTest_1sample(x-y,0,alternative)
```

Numerical example

An **imaginary** medical study was conducted to examine the effect of medicine a in lowering plasmatoxin levels in plasma. High plasmatoxin levels in plasma are related to several diseases. Plasmatoxin levels were measured at the beginning of the study and again 8 weeks after the treatment. We wish to study, whether the medicine had the desired effect on 5% significance level.

Data

Patient	Level		Difference
	Before	After	
1	1384	1332	-52
2	1640	1564	-76
3	1122	1100	-22
4	1272	1260	-12
5	1380	1360	-20
6	624	1624	1000
7	360	1821	1461
8	456	450	-6
9	1726	1712	-14
10	332	821	489
11	1342	1338	-4
12	1630	1626	-4
13	1170	1160	-10

Table: Plasmatoxin levels ($\mu\text{g}/1000\text{ml}$) before and after treatment.

t-test

One sample *t*-test

Data: differences

$t = 1.5646$, $df = 12$, $p\text{-value}=0.9282$

Alternative hypothesis: true mean is less than 0

Sample estimates: mean of $x = 210$.

Sign test

One sample sign test

Data: differences

$s = 3$, $p\text{-value} = 0.04614$

Alternative hypothesis: true median is less than 0

Sample estimates: median of $x = -10$.

Compare the results given by the two tests. Neither one of the tests alone gives a clear view on how the medicine a affects the toxin levels. Why? Based on this sample, how does the medicine seem to affect the toxin levels? Is there anything suspicious in the testing set up? Is it OK to use one sided alternative hypothesis here?

Wilcoxon signed rank test

The one sample Wilcoxon signed rank test is applied in similar testing problems as the one sample t -test. However, the one sample Wilcoxon signed rank test requires milder distributional assumptions.

Let x_1, \dots, x_n be observed values of a continuous symmetric random variable x with population median m . Assume that the observed values are i.i.d.

The null hypothesis $H_0: m = m_0$.

Possible alternative hypotheses:

$$H_1: m > m_0 \text{ (one tailed)},$$

$$H_1: m < m_0 \text{ (one tailed)},$$

$$H_1: m \neq m_0 \text{ (two tailed)}.$$

One sample Wilcoxon signed rank test

- Calculate the absolute values of the differences $|d_i| = |x_i - m_0|$ for $i = 1, \dots, n$. Order the absolute values from the smallest to the largest. Define the signed ranks $R_*(x_i)$ such that $R_*(x_i)$ is the rank of the absolute value $|d_i| = |x_i - m_0|$ multiplied with the sign of the difference $x_i - m_0$.
- The test statistic $W_* = \sum_{R_*(x_i) > 0} R_*(x_i)$ is the sum of the positive ranks. (Alternatively, the sum of the negative ranks.)
- Under H_0 , the expected value of the test statistic is $\frac{n(n+1)}{4}$ and the variance is $\frac{n(n+1)(2n+1)}{24}$.
- Large and small values (compared to the expected value $\frac{n(n+1)}{4}$) if the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p -value is small enough.
- Python:
`__, p_value = scipy.stats.wilcoxon(x-m0)`

One sample Wilcoxon signed rank test, p -value

The distribution of the test statistic W_* is tabulated and many softwares give exact p -values of the test.

The p -value of the Wilcoxon signed rank test, where w_* is the observed value of the test statistic W_* , is given as follows:

- If the alternative hypothesis is $H_1: m > m_0$, then the p -value is $p = \mathbb{P}(W_* \geq w_*)$.
- If the alternative hypothesis is $H_1: m < m_0$, then the p -value is $p = \mathbb{P}(W_* \leq w_*)$.
- If the alternative hypothesis is $H_1: m \neq m_0$, then the p -value is $p = 2 \min(\mathbb{P}(W_* \geq w_*), \mathbb{P}(W_* \leq w_*))$.

The probabilities $\mathbb{P}(W_* \geq w_*)$ and $\mathbb{P}(W_* \leq w_*)$ are calculated under the null H_0 .

Asymptotic one sample Wilcoxon signed rank test

Under H_0 , when the sample size is large, the standardized test statistic $Z = \frac{W_* - \mathbb{E}[W_*]}{\sqrt{\text{Var}(W_*)}}$, where $\mathbb{E}[W_*] = \frac{n(n+1)}{4}$ and $\text{Var}(W_*) = \frac{n(n+1)(2n+1)}{24}$, approximately follows the standard normal distribution.

The approximation is usually good enough if $n > 20$. For smaller samples, the exact distribution of W_* is needed.

One sample Wilcoxon signed rank test

We assumed above that the observations come from a continuous distribution. The Wilcoxon signed rank test can be applied for discrete observations as well. However, it is then possible that some points share the same rank of absolute values $|x_i - m_0|$. In that case, all these points are assigned to have the median of the corresponding ranks. For example, if two sample points have the same rank, corresponding to ranks 7 and 8, then both points are assigned to have rank 7.5. If three sample points have the same rank corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

Paired Wilcoxon signed rank test

The paired Wilcoxon signed rank test is applied in similar testing problems as the paired t -test.

- The observations $(x_{i,1}, x_{i,2})$, $i = 1, \dots, n$, consist of measured pairs of a random variable x .
- The pairs are assumed to be independent. However, the two values inside one pair are not assumed to be independent.
- Calculate the differences $d_i = x_{i,1} - x_{i,2}$, $i = 1, \dots, n$, of the measurements $x_{i,1}$ and $x_{i,2}$.

Paired Wilcoxon signed rank test

- General statistical assumptions: the differences d_i are i.i.d. and follow a symmetric distribution with median m .
- The null hypothesis is $H_0: m = 0$.
- Possible alternative hypotheses: $H_1: m > 0$ (one tailed), $H_1: m < 0$ (one tailed) or $H_1: m \neq 0$ (two tailed).
- Now it is possible to apply the one sample Wilcoxon signed rank test for the differences d_i .
- Python:
`__, p_value = scipy.stats.wilcoxon(x,y)`

Numerical example

We want to compare the prices of Brand X and Brand Y cookies in different stores. The distribution of the prices is not known, but it can be assumed to be symmetrical. 10 different stores were selected randomly for this study. The cookie prices have been tabulated below.

Brand X	4.56	4.67	4.28	4.57	4.78	4.54	4.56	4.48	4.47	4.50
Brand Y	4.52	4.48	4.51	4.30	4.59	4.67	4.53	4.54	4.71	4.49
Difference	0.04	0.19	-0.23	0.27	0.19	-0.13	0.03	-0.06	-0.24	0.01

Table: Prices of Brand X and Brand Y cookie packages in different stores.

Numerical example

The price differences are assumed to be symmetrically distributed. The null hypothesis is that the theoretical medians of the prices of Brand X and Brand Y cookies do not differ, i.e., the difference of the population medians is zero. The ordered absolute values of the differences and the corresponding signed ranks are as follows.

Difference	0.01	0.03	0.04	0.06	0.13	0.19	0.19	0.23	0.24	0.27
Signed rank	1	2	3	-4	-5	6.5	6.5	-8	-9	10

Table: The ordered absolute values of the differences and the corresponding signed ranks.

The test statistic

$$W_* = \sum_{R_*(d_i) > 0} R_*(x_i) = 1 + 2 + 3 + 6.5 + 6.5 + 10 = 29.$$

The p -value (obtained using statistical software) is 0.9219. We do not reject the null hypothesis.

Signed test vs. Wilcoxon signed rank test

- Both tests are suitable for similar problems: one sample – comparison of the median to a constant, paired samples – comparison of the medians.
- The tests are non-parametric counterparts of the one sample t -test.
- The values of the test statistic do not depend on the numerical values of the observations – only the order of the observations matters.
- No assumption of the type of the population distribution is needed for the sign test. Symmetry assumption is required for the Wilcoxon signed rank test.
- The Wilcoxon signed rank test uses more information of the order of the observations.
- If the distribution can be assumed to be symmetric, use the Wilcoxon signed rank test. Otherwise, apply the sign test.

Two sample Wilcoxon rank test

The two sample Wilcoxon rank test is used in similar settings as the two sample t -test, but Wilcoxon rank test requires milder assumptions.

In practice, the two sample Wilcoxon rank test is exactly the same test statistic as the Mann-Whitney test – both names are used in the literature.

Let x_1, \dots, x_n be the observed values of a continuous random variable x and let y_1, \dots, y_m be the observed values of a continuous random variable y . Assume that the observations x_1, \dots, x_n are i.i.d. and assume that y_1, \dots, y_m are i.i.d. as well. Assume also that x_i and y_j are independent for all i, j . Assume that x is distributed as y up to a location shift (i.e., x and y follow otherwise the same distribution, but possibly with different medians) and assume that the variables have population medians m_x and m_y , respectively.

The null hypothesis $H_0: m_x = m_y$.

Possible alternative hypotheses: $H_1: m_x > m_y$ (one tailed), $H_1: m_x < m_y$ (one tailed) or $H_1: m_x \neq m_y$.

Two sample Wilcoxon rank test

Consider the samples x_1, \dots, x_n and y_1, \dots, y_m . Assume (without loss of generality) that $n \leq m$.

The two sample Wilcoxon rank test is based in analyzing the order of all the observations. Combine the samples x_1, \dots, x_n and y_1, \dots, y_m to one sample z_1, \dots, z_{n+m} . Order the observations z_i from the smallest to the largest. Let $R(z_i)$ be the rank of z_i in the combined sample z_1, \dots, z_{n+m} .

- The test statistic $W = \sum_{i=1}^n R(x_i)$ is the sum of the ranks of the smaller sample.
- Under H_0 , the expected value of the test statistic is $n(n + m + 1)/2$ and the variance is $nm(n + m + 1)/12$.
- Large and small values of the test statistic (compared to the expected value $n(n + m + 1)/2$) suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- Python:
`__, p_value = scipy.stats.mannwhitneyu(x,y)`

Two sample Wilcoxon rank test, p -value

The distribution of the test statistic W is tabulated and many softwares give the exact p -values.

The p -value of the two sample Wilcoxon rank test, where w is the observed value of the test statistic W , is defined as follows:

- If the alternative hypothesis is $H_1: m_x > m_y$, then the p -value is $p = \mathbb{P}(W \geq w)$.
- If the alternative hypothesis is $H_1: m_x < m_y$, then the p -value is $p = \mathbb{P}(W \leq w)$.
- If the alternative hypothesis is $H_1: m_x \neq m_y$, then the p -value is $p = 2 \min(\mathbb{P}(W \geq w), \mathbb{P}(W \leq w))$.

Naturally, $\mathbb{P}(W \geq w)$ and $\mathbb{P}(W \leq w)$ are calculated under H_0 .

Asymptotic two sample Wilcoxon rank test

Assuming that the null hypothesis is true, if the sample size is large, the standardized test statistic $z = \frac{W - \mathbb{E}[W]}{\sqrt{\text{Var}(W)}}$, where $\mathbb{E}[W] = n(n + m + 1)/2$ and $\text{Var}(W) = nm(n + m + 1)/12$, approximately follows the standard normal distribution.

The approximation is usually good enough if $n, m > 10$. For smaller samples, the exact distribution of the test statistic W is needed.

Two sample Wilcoxon rank test

The Wilcoxon rank test can be used also when the observations are discrete. Then it is possible that some of the sample points have the same rank. In that case, all those points are assigned to have the median of the corresponding ranks. For example, if two observations have the same rank, corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same rank, corresponding to ranks 3,5, and 5, then each is assigned to have rank 4.

Note that ranks can be used even when the variables cannot be measured numerically, but they can be ordered. (For example, one could order/rank singers, or qualities of apartments, without measuring them numerically.)

Two sample Wilcoxon rank test

- The two sample Wilcoxon rank test is the non-parametric counterpart of the two sample t -test.
- The value of the test statistic depends on the order/rank of the observed value, not on the exact numerical values of the observations.
- The test is an excellent alternative to two sample t -test, when the populations are not normally distributed.

Numerical example

The height of 10 randomly chosen students was measured in the corridor of the Department of Mathematics. The students were put to stand in line from the shortest to the tallest. There were both, male and female students, in the sample. We wish to know if there is a difference in the distribution of male and female students. The null hypothesis is that the population median of the heights of the female students is equal to the population median of the heights of the male students.

The following table displays the gender and rank of the height of the students.

Student	F	F	M	M	M	F	M	M	F	M
Rank	1	2	3	4	5	6	7	8	9	10

Table: Female and male students ordered according to the rank of their height.

The test statistic

$$W = \sum_{i=1}^4 R(x_i) = 1 + 2 + 6 + 9 = 18$$

is the sum of the ranks of the smaller, female, sample. We decide to use the two-tailed alternative hypothesis (why?) and significance level 0.05. Since the samples are small, we take the critical values of the test statistic from tabulated values. The critical values are 12 and 32. Since $12 < 18 < 32$, we do not reject the null hypothesis.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eighth lecture, December 2, 2024

Proportion test

Proportion test

Proportion tests can be used for example when testing proportions of faulty products in a production process.

Let x_1, \dots, x_n be the observed values of a random variable x . Assume that the observed values are i.i.d. and come from the Bernoulli distribution with parameter p .[†]

The null hypothesis: $H_0: p = p_0$.

Possible alternative hypotheses:

$$H_1: p > p_0 \text{ (one tailed),}$$

$$H_1: p < p_0 \text{ (one tailed),}$$

$$H_1: p \neq p_0 \text{ (two tailed).}$$

[†]Now $\mathbb{P}(x = 1) = p$, $\mathbb{P}(x = 0) = 1 - p$, $\mathbb{E}[p] = p$, and $\text{Var}(x) = p(1 - p)$.

Proportion test

- The test statistic $C = \sum_{i=1}^n x_i$.
- If the null hypothesis H_0 is true, then the test statistic follows the binomial distribution with parameters n and $p = p_0$.
- Under the null hypothesis H_0 , the expected value of the test statistic is np_0 ($\mathbb{E}[C] = np_0$) and the variance of the test statistic is $np_0(1 - p_0)$.
- If the value of the test statistic is large or small compared to the expected value np_0 , evidence against the null hypothesis is found.
- The null hypothesis is rejected if the p -value is small enough.
- Python:

```
C_stat = sum(x) # x is a (0,1) vector of length n  
# containing the outcomes of  
# Bernoulli trials  
  
p_value = scipy.stats.binomtest(C_stat,n,p=p0).pvalue
```

Proportion test, p -value

The distribution of the test statistic C is tabulated and statistical software can be used to calculate p -values of the test.

Let c denote the observed value of the test statistic C . Then the p -value of the test is given as follows:

- If the alternative hypothesis is $H_1: p > p_0$, then the p -value is $p = \mathbb{P}(C \geq c)$.
- If the alternative hypothesis is $H_1: p < p_0$, then the p -value is $p = \mathbb{P}(C \leq c)$.
- If the alternative hypothesis is $H_1: p \neq p_0$, then the p -value is usually[†] defined as $p = \sum_{k: p_C(k) \leq p_C(c)} p_C(k)$, where p_C denotes the PMF of $\text{Bin}(n, p_0)$.

The probabilities $\mathbb{P}(C \geq c)$ and $\mathbb{P}(C \leq c)$ are calculated under H_0 .

[†]Especially, statistical software such as R or the Scipy library use this formula.

Asymptotic proportion test

If the sample size is large, then under the null hypothesis H_0 , the standardized test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}},$$

where $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$ is the unbiased estimator of the parameter p , approximately follows the standard normal distribution.

The approximation is usually good enough if $n\hat{p} > 10$ and $n(1 - \hat{p}) > 10$. For smaller samples, the test relies on the exact distribution of the test statistic.

Numerical example

In anticipation of an upcoming election, an opinion poll was conducted. In the poll, the sample size was 1000 and 420 out of the 1000 eligible voters reported that they support the mayor. We want to test on significance level 5% whether the true support is less than 50% of the population.

Null hypothesis: $H_0: p = 0.5$.

Alternative hypothesis: $H_1: p < 0.5$.

Since $n = 1000$ and $\hat{p} = \frac{420}{1000} = 0.42$ satisfy $n\hat{p} > 10$ and $n(1 - \hat{p}) > 10$, we can use normal approximation. The observed value of the Z-statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/1000}} = \frac{0.42 - 0.50}{\sqrt{0.5^2/1000}} \approx -5.06.$$

The p -value is $p = \mathbb{P}(Z \leq z) = \Phi(-5.06) \approx 2.10 \cdot 10^{-7}$. H_0 is rejected.

Python:

```
>>>scipy.stats.binomtest(420,1000,p=0.5,alternative='less')
2.348554631632085e-07 # exact binomial test
>>>scipy.stats.norm.cdf((0.42-0.50)/numpy.sqrt(0.5*0.5/1000))
2.1001969880109918e-07 # normal approximation
```

Two sample proportion test

In the two sample proportion test, parameters of two different Bernoulli distributed samples are compared.

Let x_1, \dots, x_n be the observed values of a random variable x and let y_1, \dots, y_m be the observed values of a random variable y . Assume that the observed values x_1, \dots, x_n are i.i.d. and come from the Bernoulli distribution with parameter p_x , and assume that the observed values y_1, \dots, y_m are i.i.d. and come from the Bernoulli distribution with parameter p_y . Furthermore, assume that x_i and y_j are independent for all i, j .

The null hypothesis: $H_0: p_x = p_y$.

Possible alternative hypotheses:

$$H_1: p_x > p_y \text{ (one tailed)},$$

$$H_1: p_x < p_y \text{ (one tailed)},$$

$$H_1: p_x \neq p_y \text{ (two tailed)},$$

Two sample proportion test

- Calculate the sample proportions $\hat{p}_x = \frac{1}{n} \sum_{i=1}^n x_i$ and $\hat{p}_y = \frac{1}{m} \sum_{i=1}^m y_i$, and $\hat{p} = \frac{n\hat{p}_x + m\hat{p}_y}{n+m}$.
- Calculate the test statistic

$$Z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}.$$

- If the sample size is large, then under the null hypothesis H_0 , the test statistic Z approximately follows the standard normal distribution. The approximation is usually good enough if $n\hat{p}_x > 5$, $n(1 - \hat{p}_x) > 5$, $m\hat{p}_y > 5$, and $m(1 - \hat{p}_y) > 5$.
- If the value of the test statistic has large absolute value, then evidence against the null hypothesis H_0 is found.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Testing general statistical assumptions

In statistics, we very often make assumptions about the underlying distribution. Most statistical methods become ineffective or give false results if these assumptions do not hold. Hence **it is very important to test the distributional assumptions separately.**

Testing normality

Normality assumption

The normal distribution has a central role in statistics. Multiple methods for testing the normality of observations have been developed. Here, we take a look at a couple of them.

In what follows, let x_1, \dots, x_n be i.i.d. observations of a random variable x .

The null hypothesis is H_0 : “the random variable x is normally distributed.”

The alternative hypothesis is H_1 : “the random variable x is not normally distributed.”

The Bowman and Shenton normality test

The Bowman and Shenton normality test is a function of skewness and kurtosis:

$$BS = n \left(\frac{v^2}{6} + \frac{k^2}{24} \right),$$

where v is the sample skewness coefficient and k is the sample kurtosis coefficient.

If the skewness or kurtosis differ a lot from the skewness and/or kurtosis of the normal distribution, the test statistic gets large values.

Bowman and Shenton normality test

- If n is large, then under the null hypothesis H_0 , the test statistic BS follows approximately the $\chi^2(2)$ distribution.
- The expected value of the test statistic under the null hypothesis H_0 is $\mathbb{E}[BS] = 2$.
- Large values of the test statistic compared to the expected value suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- If one uses the formulae $\hat{v} = \frac{m_3}{\hat{s}^3}$ and $\hat{k} = \frac{m_4}{\hat{s}^4} - 3$, where

$\hat{s} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ is the *biased* sample standard deviation, then one obtains the closely related **Jarque–Bera** test statistic

$$JB = n \left(\frac{\hat{v}^2}{6} + \frac{\hat{k}^2}{24} \right),$$

also used to assess normality. This test statistic is implemented in the Python Scipy library as `scipy.stats.jarque_bera`

- Note that the Bowman and Shenton (resp. Jarque–Bera) normality test is suitable for large samples only!

Implementation using Python

```
import numpy as np
from scipy.stats import chi2

def BowmanShentonTest(x):
    n = len(x)
    xbar = np.mean(x)
    std = np.std(x,ddof=1)
    v = (1/n)*sum((x-xbar)**3)/std**3
    k = (1/n)*sum((x-xbar)**4)/std**4-3
    BS = n*(v**2/6+k**2/24)
    q = chi2.cdf(BS,2)
    return BS,1-q

# Note: if the distribution has 0 skewness and 0 kurtosis
# (ideal case for the normal distribution), then the test
# statistic BS == 0. Thus we choose a one sided alternative
# hypothesis of type 'greater' since only large values of BS
# would be evidence of non-normality.
```

Rank plot

Let x_1, \dots, x_n be i.i.d. observations from some distribution F_x . Let $z_1 \leq \dots \leq z_n$ be the observations x_1, \dots, x_n ordered from the smallest to the largest one. Let $y_1 \leq \dots \leq y_n$ be the ordered values of n i.i.d. observations from the standard normal distribution $\mathcal{N}(0, 1)$ and let $\mathbb{E}[y_i]$ be the expected value of y_i .

Plot the pairs $(\mathbb{E}[y_i], z_i)$, $i = 1, \dots, n$. If the x_i come from a normal distribution, then the points $(\mathbb{E}[y_i], z_i)$ should approximately lie on a line. If the points do not lie on a line, the sample differs from the normal distribution. The plot can be used in detecting skewness of the distribution and in finding outliers.

Rank plots are useful for quick visual assessment of the distribution of the data: cf., e.g., the excellent StackExchange post

<https://stats.stackexchange.com/a/101290>

Shapiro–Wilk normality test

- The Shapiro–Wilk normality test statistic is the squared value of the Pearson sample correlation coefficient calculated from the rank plot points $(\mathbb{E}[y_i], z_i)$, $i = 1, \dots, n$.
- Small values of the test statistic suggest that the assumption of normality does not hold. Large values of the test statistic are in line with the null hypothesis.
- The null hypothesis is rejected if the p -value is small enough. The test requires a large sample.
- Statistical software can be used to calculate the p -value of the test.
Python: `scipy.stats.shapiro(x)`

Numerical example

During the previous lecture, we considered an example where we compared Brand X and Brand Y cookies. In the example, the price differences were assumed to be symmetrically distributed. The data consisted of the cookie prices in 10 randomly selected stores. We now wish to test the normality of the price differences. The price differences are given below.

Difference: 0.04 | 0.19 | -0.23 | 0.27 | 0.19 | -0.13 | 0.03 | -0.06 | -0.24 | 0.01

Table: The differences of Brand X and Brand Y cookie prices.

The Bowman and Shenton test: In order to calculate the test statistic, the sample skewness and kurtosis coefficients v and k are needed. The sample standard deviation is $s \approx 0.176$ and the sample mean is $\bar{x} \approx 0.07$. Now

$$v = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3} \approx -0.0139$$
$$k = \frac{m_4}{s^4} - 3 = \left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} \right) - 3 \approx -1.506.$$

The value of the test statistic is

$$BS = n \left(\frac{v^2}{6} + \frac{k^2}{24} \right) \approx 0.945.$$

Under the null hypothesis, the test statistic follows the $\chi^2(2)$ distribution. We decide to use the significance level 0.05. The critical values are then 0.051 and 7.378. Since $0.051 < 0.945 < 7.378$, evidence of non-normality was not found.

Rank plot (Q-Q plot):

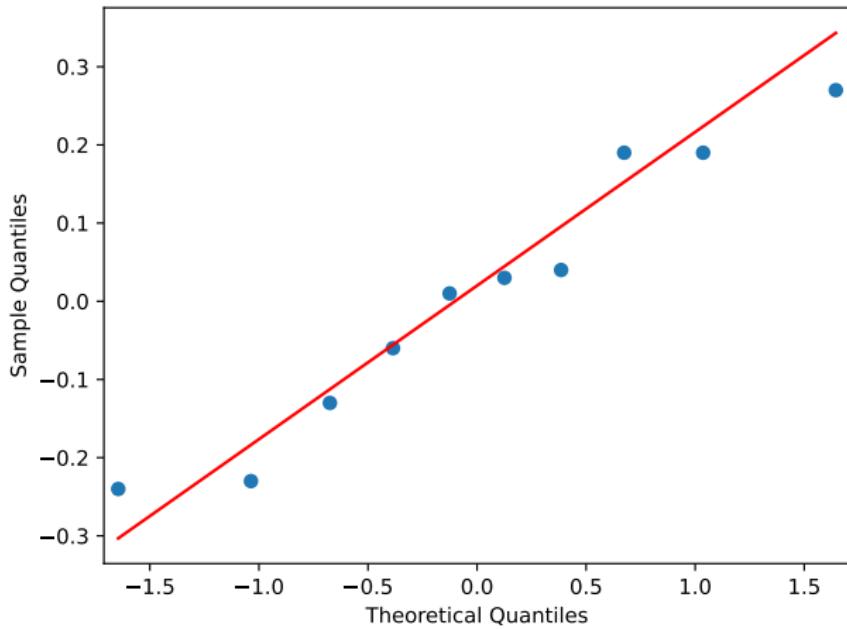


Figure: Rank plot of the price differences.

Shapiro–Wilk test: calculated in Python using the function
`scipy.stats.shapiro`

data: differences

$$W = 0.9439, \text{ } p\text{-value} = 0.5966$$

The p -value is large and thus evidence of non-normality was not found.

Can these results be trusted? Were all the required assumptions fulfilled?
What was the type 2 error?

χ^2 tests

Multinomial distribution

Consider a situation, where a random experiment has k mutually exclusive outcomes and consider n independent runs of that experiment. The multinomial distribution models the frequency distribution of the outcome of these n independent random experiments.

The random variables x_1, \dots, x_k follow the multinomial distribution with parameters n, p_1, \dots, p_k , if the probability mass function is

$$p(x_1, \dots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

where

$$\sum_{i=1}^k x_i = n \quad \text{and} \quad \sum_{i=1}^k p_i = 1.$$

Assume that x_1, \dots, x_k follow multinomial distribution with parameters n, p_1, \dots, p_k . If n is large, then

$$\sum_{i=1}^k \frac{(x_i - np_i)^2}{np_i}$$

approximately follows the $\chi^2(k - 1)$ distribution.

χ^2 goodness-of-fit test

The χ^2 goodness-of-fit test examines the discrepancy between observed values and the values expected under some particular distribution of a random variable x .

The null hypothesis H_0 : “The random variable x follows distribution F_x (with or without unknown parameters).”

The alternative hypothesis H_1 : “The random variable x does not follow distribution F_x (with or without unknown parameters).”

χ^2 goodness-of-fit test

Let x_1, \dots, x_n be i.i.d. observations of a random variable x .

- Categorize the n observations into k categories.
- Calculate the frequencies O_i , $i = 1, \dots, k$, where O_i is the observed frequency of the category i . Note that $\sum_{i=1}^k O_i = n$.
- Let p_i be the probability that, under the null hypothesis, the random variable x belongs to the category i . Calculate the expected frequencies $E_i = np_i$ of the observations in category i . Note that $\sum_{i=1}^k p_i = 1$.
- Now, under the null hypothesis, the random variables O_1, \dots, O_k follow the multinomial distribution with parameters n, p_1, \dots, p_k .

χ^2 goodness-of-fit test

- Calculate the test statistic

$$\chi_g^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}.$$

- If n is large, then under the null hypothesis, the test statistic χ_g^2 approximately follows $\chi^2(k - 1 - e)$ distribution, where e is the number of estimated parameters.
- The expected value of the test statistic, under the null hypothesis, is $\mathbb{E}[\chi_g^2] = k - 1 - e$.
- Large values of the test statistic (compared to the expected value) suggest that the null hypothesis H_0 does not hold.
- If the p -value is small enough, then the null hypothesis H_0 is rejected.
- If the value of the test statistic is large, the sample frequencies differ greatly from the expected value and it is clear that the null hypothesis should be rejected. However, if the value is very small, then the sample frequencies differ less than expected. This is called **overfitting** – usually, we are not concerned about this, so typically a one tailed alternative hypothesis (of type alternative='greater') is used.

Goodness-of-fit test, Example 1

Let us examine the quality of giant mugs made in a ceramics factory. The null hypothesis is that:

- an error in the **shape of the mug** occurs with probability $2/14$,
- a **color error** occurs with probability $2/14$,
- **both errors** occur simultaneously with probability $1/14$,
- the probability of an **error-free** product is $9/14$.

Consider a sample of 200 randomly selected mugs such that

- 40 mugs have an **error in the shape**,
- 44 have a **color error**,
- 26 mugs have **both errors**,
- 90 mugs are **error-free**.

Now $O_1 = 40$, $O_2 = 44$, $O_3 = 26$, $O_4 = 90$

$$E_1 = 200 \cdot \frac{2}{14}, E_2 = 200 \cdot \frac{2}{14}, E_3 = 200 \cdot \frac{1}{14}, E_4 = 200 \cdot \frac{9}{14}$$

$\therefore \chi_g^2 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i} = 34.08$. Under the null hypothesis, the test statistic approximately follows the $\chi^2(4 - 1) = \chi^2(3)$ distribution. Since $\mathbb{P}(\chi^2(3) \geq 34.08) < 0.00001$, the null hypothesis is rejected.

Goodness-of-fit test, Example 1

The χ^2 goodness-of-fit test is implemented in the Python Scipy library as `scipy.stats.chisquare`

For example, we can solve the previous example numerically as follows:

```
from scipy.stats import chisquare
O = [40,44,26,90]
E = [200*2/14,200*2/14,200*1/14,200*9/14]
chisquare(O,E)
```

The output is

```
Power_divergenceResult(statistic=34.08,
                        pvalue=1.905621048402571e-07)
```

Goodness-of-fit test, Example 2

Consider testing whether the monthly salary of Germans follows the normal distribution. Select randomly n Germans and document the salaries. The null hypothesis is that the observations come from a normal distribution with an unknown expected value and an unknown variance.

- Estimate the unknown parameters (μ and σ^2) from the sample.
- Discretize the continuous salary variable.
- Calculate the observed category frequencies O_1, \dots, O_k , i.e., calculate the number of observations in each category.
- Calculate the category probabilities from the normal distribution. For example,

$$\dots, \mathbb{P}(1900 < X \leq 2000), \mathbb{P}(2000 < X \leq 2100), \dots$$

- Calculate the expected category frequencies E_1, \dots, E_k .
- Calculate the test statistic. Under the null hypothesis, the test statistic approximately follows the $\chi^2(k - 1 - 2) = \chi^2(k - 3)$, where k is the number of the used categories and we estimated $e = 2$ parameters (μ and σ^2). Calculate the p -value and based on that, either reject or do not reject the null hypothesis.

χ^2 homogeneity test

In the χ^2 homogeneity test, several (r) samples are examined.

The null hypothesis H_0 : “The samples come from (some) same distribution.”

The alternative hypothesis H_1 : “The samples do not come from the same distribution.”

χ^2 homogeneity test

Consider several (r) independent samples. Assume that the observations of each sample are i.i.d. Assume that the sample i , $i \in \{1, \dots, r\}$, has n_i observations.

- Categorize all the observations into c categories of size C_j .
- Calculate the frequencies O_{ij} , $i = 1, \dots, r$, $j = 1, \dots, c$, where O_{ij} is the observed frequency of the observations of the sample i in category j

	1	2	\dots	c	sum
1	O_{11}	O_{12}	\dots	O_{1c}	n_1
2	O_{21}	O_{22}	\dots	O_{2c}	n_2
\dots	\dots	\dots	\dots	\dots	\dots
r	O_{r1}	O_{r2}	\dots	O_{rc}	n_r
sum	C_1	C_2	\dots	C_c	n

Table: The observed frequencies.

- Let $p_j = C_j/n$. Under the null hypothesis, for each sample i , the probability of the category j is the same p_j .
- Calculate the expected frequencies $E_{ij} = n_i p_j$.

χ^2 homogeneity test

	1	2	\dots	c	sum
1	E_{11}	E_{12}	\dots	E_{1c}	n_1
2	E_{21}	E_{22}	\dots	E_{2c}	n_2
\dots	\dots	\dots	\dots	\dots	\dots
r	E_{r1}	E_{r2}	\dots	E_{rc}	n_r
sum	C_1	C_2	\dots	C_c	n

Table: The expected frequencies.

χ^2 homogeneity test

- Calculate the value of the test statistic

$$\chi_h^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- If n is large, then under the null hypothesis, the test statistic χ_h^2 approximately follows the $\chi^2((r - 1)(c - 1))$ distribution.
- Under the null hypothesis, the expected value of the test statistic is $(r - 1)(c - 1)$. (That is, $\mathbb{E}[\chi_h^2] = (r - 1)(c - 1)$.)
- Large values of the test statistic compared to the expected value suggest that the null hypothesis H_0 is false. Small values of the test statistic compared to the expected value are indicative of **overfitting** – the data fits the model “too well”. Usually, we are not too concerned about this, so typically a one tailed alternative hypothesis (of type alternative='greater') is used.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Homogeneity test, Example

A city council is about to make decisions about building a new library. There was a preliminary plan and 250 randomly selected men and 300 randomly selected women were asked to comment the plan. 169 men and 125 women thought that the plan was good, 52 men and 144 women did not like the plan, and 29 men and 31 women did not have an opinion about the plan.

	good plan	bad plan	no opinion	Total
Men	169	52	29	250
Women	125	144	31	300
Total	294	196	60	550

Table: Observed frequencies

	good plan	bad plan	no opinion	Total
Men	133.6	89.1	27.3	250
Women	160.4	106.9	32.7	300
Total	294	196	60	550

Table: Expected frequencies

Homogeneity test, Example

The value of the test statistic:

$$\chi_h^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 45.7105.$$

Under the null hypothesis, the test statistic approximately follows the $\chi^2((2 - 1)(3 - 1)) = \chi^2(2)$ distribution. Since $\mathbb{P}(\chi^2(2) \geq 45.7105) < 0.00001$, it can be concluded that the opinions about the preliminary plan do differ between men and women.

Solution using Python:

```
import pandas as pd
from scipy.stats import chisquare
O = pd.DataFrame({'good plan': [169, 125], 'bad plan':
                  [52, 144], 'no opinion': [29, 31]}, index=['Men', 'Women'])
tmp = O.values # create expected frequency table
E = pd.DataFrame((tmp.sum(0)*tmp.sum(1)[:,None])/tmp.sum(),
                  columns=O.columns, index=O.index)
chisquare(O, E, ddof=(O.shape[0]-1)*(O.shape[1]-1), axis=None)
```

χ^2 test of independence

The χ^2 test of independence is applied to study whether two random variables (factors) are stochastically independent.

Null hypothesis H_0 : “the variables are independent.”

Alternative hypothesis: H_1 : “the variables are not independent.”

χ^2 test of independence

Consider a simple random sample of size n . Divide the observations to r classes with respect to a factor A and to c classes with respect to a factor B . Let R_i be the frequency of the observations in class i with respect to the factor A . Let C_j be the frequency of the observations in class j with respect to the factor B . Let O_{ij} be the observed frequency of the observations in class i with respect to the factor A and class j with respect to the factor B .

	1	2	...	c	sum
1	O_{11}	O_{12}	...	O_{1c}	R_1
2	O_{21}	O_{22}	...	O_{2c}	R_2
...
r	O_{r1}	O_{r2}	...	O_{rc}	R_r
sum	C_1	C_2	...	C_c	n

Table: The observed frequencies.

- Let $P_j = C_j/n$. Under the null hypothesis, for each category i of the factor A , the probability of category j of the factor B has the same probability P_j .
- Calculate the expected frequencies $E_{ij} = R_i P_j$.

	1	2	...	c	sum
1	E_{11}	E_{12}	...	E_{1c}	R_1
2	E_{21}	E_{22}	...	E_{2c}	R_2
...
r	E_{r1}	E_{r2}	...	E_{rc}	R_r
sum	C_1	C_2	...	E_c	n

Table: The expected frequencies.

χ^2 test of independence

- Calculate the value of the test statistic

$$\chi_I^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- If n is large, then under the null hypothesis, the test statistic approximately follows the $\chi^2((r - 1)(c - 1))$ distribution.
- The expected value of the test statistic is $(r - 1)(c - 1)$. That is, $\mathbb{E}[\chi_I^2] = (r - 1)(c - 1)$.
- Large values (compared to the expected values) of the test statistic suggest that the null hypothesis is false. Small values of the test statistic compared to the expected value are indicative of **overfitting** – the data fits the model “too well”. Usually, we are not too concerned about this, so typically a one tailed alternative hypothesis (of type alternative='greater') is used.
- The null hypothesis is rejected if the p -value is small enough.

Test of independence, Example

There was an interesting presidential election and we wish to examine the independence of the voting behavior of married men (M) and women (W). The sample consists of 120 married couples and the presidential candidates were A, B, C. In total, there are nine categories: AA, AB, AC, BA, BB, BC, CA, CB, CC.

	A, man	B, man	C, man	Total
A, woman	15	7	8	30
B, woman	20	25	5	50
C, woman	10	10	20	40
Total	45	42	33	120

Table: Observed frequencies

	A, man	B, man	C, man	Total
A, woman	11.25	10.50	8.25	30
B, woman	18.75	17.50	13.75	50
C, woman	15.00	14.00	11.00	40
Total	45	42	33	120

Table: Expected frequencies

The value of the test statistic

$$\chi_r^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 21.46.$$

Under the null hypothesis, the test statistic approximately follows the $\chi^2((3-1)(3-1)) = \chi^2(4)$ distribution. Since $\mathbb{P}(\chi^2(4) \geq 21.46) = 0.000257$, we conclude that the voting behavior of married men and women is not independent.

Solution using Python:

```
import pandas as pd
from scipy.stats import chisquare
O = pd.DataFrame({'A, man': [15,20,10], 'B, man': [7,25,10],
                   'C, man': [8,5,20]}, index=['A, woman', 'B, woman', 'C, woman'])
tmp = O.values # create expected frequency table
E = pd.DataFrame((tmp.sum(0)*tmp.sum(1)[:,None])/tmp.sum(),
                  columns=O.columns, index=O.index)
chisquare(O,E,ddof=(O.shape[0]-1)*(O.shape[1]-1), axis=None)
```

Remark. The χ^2 test of independence and the χ^2 homogeneity test are very similar. The test statistics and the degrees of freedom are calculated identically. However, the tests measure very different phenomena.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Ninth lecture, December 9, 2024

Analysis of variance

Analysis of variance

The two sample t -test generalizes into **analysis of variance**.

In analysis of variance – **ANOVA** – the population consists of two or more independent groups. Observations are assumed to follow a normal distribution. Each group is independently sampled.

ANOVA tests the equality of the expected values of the groups.

For example, we could test if there is a difference in the mean monthly salary in the 10 largest cities in Germany.

ANOVA

ANOVA can be generalized in several different ways:

- In multivariate analysis of variance, MANOVA, the tested expected values are vectors. We could test the equality of the mean monthly salary *and* weekly overtime in the 10 largest cities in Germany.
 - The population could also be divided into groups based on multiple factors (multifactor ANOVA), of which some can be continuous (analysis of covariance, ANCOVA). For example, we could divide people into groups based on the city they live in and based on their gender.
 - In multifactor MANOVA, the tested expected values are vectors.
- ⋮

In what follows, we only consider cases where the population is divided into groups with respect to just one factor and the expected value that is tested is univariate.

ANOVA

Let $x_{1,j}, x_{2,j}, \dots, x_{n_j,j}$ be observed values of a random variable x_j , $j \in \{1, \dots, k\}$. Assume that the observations $x_{1,j}, x_{2,j}, \dots, x_{n_j,j}$ are i.i.d. and follow the normal distribution $\mathcal{N}(\mu_j, \sigma^2)$, $j \in \{1, \dots, k\}$.

That is, we now have k random samples from univariate normal distributions, and the variance of all the k normal distributions are assumed to be equal.

Assume that the k samples are independent.

Null hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_k$.

Alternative hypothesis $H_1: \mu_i \neq \mu_j$ for some $i \neq j$.

ANOVA

In analysis of variance, the total variance is divided into two parts. The first part measures the variation between the group means, and the second part measures the variation within the groups. If the first part is much larger than the second part, there is evidence against the null hypothesis and it can be rejected.

The test of equality of the expected values is based on comparison of between-groups variance and within-groups variance. Hence the name of the method – analysis of variance.

ANOVA

Calculate the group means

$$\bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{i,j}$$

and the combined sample mean

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{i,j},$$

where $n = \sum_{j=1}^k n_j$.

ANOVA

Consider the **total sum of squares**

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{i,j} - \bar{x})^2,$$

the variance between groups (**group sum of squares**)

$$SSG = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2,$$

and the variance within groups (**error sum of squares**)

$$SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2 = \sum_{j=1}^k (n_j - 1) s_j^2,$$

where $s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2$.

Now the total sum of squares $SST = SSG + SSE$.

ANOVA

- The F -test statistic

$$F = \frac{n - k}{k - 1} \frac{SSG}{SSE}.$$

- Under the null hypothesis, the test statistic follows the F -distribution with parameters $(k - 1)$ and $(n - k)$.
- The expected value of the test statistic under H_0 is $\mathbb{E}[F] = \frac{n-k}{n-k-2}$.
- Large values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p -value is small enough.
- Python:

```
F_stat,p_value = scipy.stats.f_oneway(group1,...,group_k)
```

F -distribution

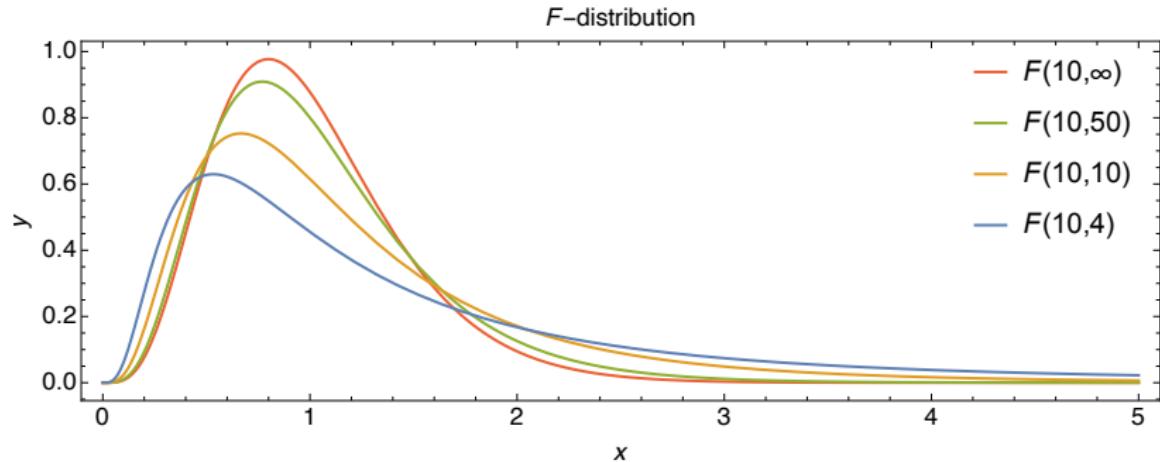


Figure: F -distributions with different parameters.

Numerical example

A research group was set to study whether the expected value of a specific laboratory test L differs between patients that are on different medications (A, B, C). 10 patients receiving medication A (group 1), 10 patients receiving medication B (group 2), and 10 receiving medication C (group 3) were picked randomly and lab test L was taken. The next table shows the accurately measured laboratory test results.

Group 1 (A)	Group 2 (B)	Group 3 (C)
0.111	0.109	0.119
0.123	0.107	0.124
0.109	0.103	0.125
0.120	0.104	0.117
0.115	0.098	0.111
0.112	0.110	0.120
0.117	0.101	0.118
0.110	0.115	0.116
0.119	0.099	0.122
0.116	0.111	0.119

Table: Laboratory test results for groups 1, 2, and 3.

The group means are $\bar{x}_1 = 0.1152$, $\bar{x}_2 = 0.1057$, and $\bar{x}_3 = 0.1191$ and the combined mean is $\bar{x} = 0.1133$. The group variances are $s_1^2 = 2.173333 \cdot 10^{-5}$, $s_2^2 = 3.134444 \cdot 10^{-5}$, and $s_3^2 = 1.654444 \cdot 10^{-5}$.

The total sum of squares is

$$SST = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{i,j} - \bar{x})^2 = \sum_{i=1}^{10} (x_{1,i} - 0.1133)^2 + \sum_{i=1}^{10} (x_{2,i} - 0.1133)^2 \\ + \sum_{i=1}^{10} (x_{3,i} - 0.1133)^2 = 0.001576667,$$

the group sum of squares is

$$SSG = \sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2 = \sum_{j=1}^3 10 \cdot (\bar{x}_j - 0.1133)^2 = 0.0009500667,$$

and the error sum of squares is

$$SSE = \sum_{j=1}^k (n_j - 1) s_j^2 \\ = 9 \cdot (2.173333 \cdot 10^{-5} + 3.134444 \cdot 10^{-5} + 1.654444 \cdot 10^{-5}) \\ = 0.0006265999.$$

The value of the test statistic is

$$F = \frac{n - k}{k - 1} \frac{SSG}{SSE} = \frac{27}{2} \cdot \frac{0.0009500667}{0.0006265999} = 20.46904.$$

Under the null hypothesis, the test statistic follows the F -distribution with parameters 2 and 27. The one-tailed critical value on 5% significance level is $3.354 < 20.46904$. The null hypothesis can be rejected.

Solution using Python:

```
import pandas as pd
from scipy.stats import f_oneway
data = pd.DataFrame({
    "A": [.111,.123,.109,.120,.115,.112,.117,.110,.119,.116],
    "B": [.109,.107,.103,.104,.098,.110,.101,.115,.099,.111],
    "C": [.119,.124,.125,.117,.111,.120,.118,.116,.122,.119]
})
F_stat,p_value = f_oneway(*data.T.values)
```

Pairwise comparison

Pairwise comparison

If the null hypothesis (equality of the expected values) is rejected based on the F -test, then we know **at least two of the groups differ** (but we do not know which ones).

The next step in the analysis is pairwise comparison. The goal in pairwise comparison is to identify the groups with statistically significant differences in expected values.

A simple way to do this is to analyze the groups in pairs of two with the t -test.

There are $c = \frac{k(k-1)}{2}$ pairs in total to compare and conducting all possible comparisons has the side effect that the **probability of type 1 error is inflated greatly above its set level**.

Bonferroni's method for pairwise comparison

Consider pairwise comparison of the expected values. There are $c = \frac{k(k-1)}{2}$ pairs in total to compare. Let us consider analyzing the pairs with the t -test.

Let β be the significance level of the c pairwise comparisons, i.e., the (upper bound for the) probability that H_0 is incorrectly rejected in a single comparison. Let α be the probability that H_0 is incorrectly rejected in at least one test when the test is repeated c times, i.e., the probability of making at least one type 1 error during the c tests.

Probability theory shows that $\alpha \leq c\beta$. For this reason, if the significance level α is chosen for the combined comparison, the individual comparisons must be done on level $\beta = \frac{\alpha}{c}$. (For pairwise tests, instead of p -value α , one looks for p -values $\leq \frac{\alpha}{c}$.) This is known as the **Bonferroni correction**.

Example. We want to investigate, on significance level $\alpha = 0.05$, the differences in expected values for $k = 5$ groups. Then there are $c = 10$ pairs to compare. The t -test should be carried out at significance level $\frac{0.05}{10} = 0.005$ for each of the 10 pairs.

Bartlett's test for equality of variances

Bartlett's test for equality of variances

ANOVA makes two key assumptions:

1. The groups are normally distributed.
2. The groups have equal variances.

As usual, the first assumption can (by CLT) be replaced with a large enough sample size n .

The second assumption is also required for large samples. However, ANOVA is robust to moderate violations from it. As a rule of thumb, the largest group variance should be at most 4 times the smallest group variance.

The variance assumption can also be tested using Bartlett's test.

Bartlett's test for equality of variances

Let $x_{1,j}, x_{2,j}, \dots, x_{n_j,j}$ be observed values of a random variable x_j , $j \in \{1, \dots, k\}$. Assume that the observations are i.i.d. and follow a normal distribution $\mathcal{N}(\mu_j, \sigma_j^2)$, $j \in \{1, \dots, k\}$. Assume that all the k samples are independent.

The null hypothesis $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$.

The alternative hypothesis $H_1: \sigma_i^2 \neq \sigma_j^2$ for some $i \neq j$.

Bartlett's test for equality of variances

Let

$$s^2 = \frac{1}{n - k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2,$$

and

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (x_{i,j} - \bar{x}_j)^2.$$

Let

$$B = \frac{Q}{h},$$

where

$$Q = (n - k) \ln s^2 - \sum_{j=1}^k (n_j - 1) \ln s_j^2$$

and

$$h = 1 + \frac{1}{3(k-1)} \left(\left(\sum_{j=1}^k \frac{1}{n_j - 1} \right) - \frac{1}{n - k} \right).$$

Bartlett's test for equality of variances

- Bartlett's test statistic

$$B = \frac{Q}{h},$$

where

$$Q = (n - k) \ln s^2 - \sum_{j=1}^k (n_j - 1) \ln s_j^2$$

and

$$h = 1 + \frac{1}{3(k-1)} \left(\left(\sum_{j=1}^k \frac{1}{n_j - 1} \right) - \frac{1}{n-k} \right).$$

- If the sample size is large, then under the null hypothesis the test statistic approximately follows the χ^2 distribution with $(k - 1)$ degrees of freedom.
- The expected value of the test statistic under H_0 is $\mathbb{E}[B] = k - 1$.
- Large values of the test statistic suggest that the null hypothesis H_0 is false. The null hypothesis H_0 is rejected if the p -value is small enough.
- Python:

```
Q_stat, p_value = scipy.stats.bartlett(group1, ..., group_k)
```

Kruskal–Wallis test

Analysis of variance

ANOVA tests the equality of the expected values of normally distributed samples. Next we consider a non-parametric alternative to ANOVA.

Kruskal–Wallis test

The Kruskal–Wallis test is similar to one way analysis of variance, but it does not require the normality assumption.

It is a generalization of the two sample Wilcoxon rank test.

Kruskal–Wallis test

Let $x_{1,j}, x_{2,j}, \dots, x_{n_j,j}$ be observed values of a continuous random variable x_j , $j \in \{1, \dots, k\}$. Assume that the observations $x_{1,j}, x_{2,j}, \dots, x_{n_j,j}$ are i.i.d. Assume also that the k samples are independent and that the variables x_j , $j \in \{1, \dots, k\}$, follow the same distribution up to location shifts (i.e., x_j follow otherwise the same distribution, but possibly with different medians) and assume that the variables x_j have population medians m_j , $j \in \{1, \dots, k\}$.

The null hypothesis $H_0: m_1 = m_2 = \dots = m_k$.

The alternative hypothesis $H_1: m_i \neq m_j$ for some $i \neq j$.

Kruskal–Wallis test

The Kruskal–Wallis test is based on examining the ranks of the observations.

Kruskal–Wallis test

Combine the groups $x_{1,j}, x_{2,j}, \dots, x_{n_j,j}$, $j \in \{1, \dots, k\}$, into one big sample z_1, z_2, \dots, z_n , where $n = \sum_{j=1}^k n_j$. Order the observations z_s from the smallest value to the largest value. Let $R(z_s)$ be the rank of the observation z_s in the combined sample z_1, z_2, \dots, z_n .

Calculate the group means of the ranks

$$\bar{r}_j = \frac{1}{n_j} \sum_{\substack{s=1 \\ z_s=x_{i,j}}}^{n_j} R(z_s),$$

and the mean of the combined sample

$$\bar{r} = \frac{1}{n} \sum_{s=1}^n R(z_s).$$

Kruskal–Wallis test

Consider the group sum of squares, which describes the variation of the ranks between the groups

$$\sum_{j=1}^k n_j(\bar{r}_j - \bar{r})^2$$

and the total sum of squares, which describes the variation of the ranks in the combined sample

$$\sum_{s=1}^n (R(z_s) - \bar{r})^2.$$

Kruskal–Wallis test

- Kruskal–Wallis test statistic

$$K = (n - 1) \frac{\sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2}{\sum_{s=1}^n (R(z_s) - \bar{r})^2}.$$

- Under the null hypothesis H_0 , if the sample size is large, the test statistic approximately follows the χ^2 distribution with $k - 1$ degrees of freedom.
- Under H_0 , the (asymptotic) expected value of the test statistic is $k - 1$.
- Large values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis is rejected if the p -value is small enough.
- Python:

```
K_stat, p_value = scipy.stats.kruskal(group1, ..., groupk)
```

Kruskal–Wallis test

Statistical software often calculate exact p -values for the Kruskal–Wallis test when the sample size is small. With large sample sizes, the calculation of exact p -values requires intense computations and in these cases asymptotic p -values (based on the χ^2 distribution) are used.

Discrete distributions

We assumed above that the observations follow some continuous distribution. However, the Kruskal–Wallis test can be used for discrete observations as well. Then it is possible that some of the observations have the same rank. In this case, all those observations are assigned to have the median of the corresponding ranks. For example, if two observations have the same rank corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same ranks corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

ANOVA vs. Kruskal–Wallis

ANOVA is explicitly a test for the equality of the expected values. The Kruskal–Wallis test can, technically, be seen as a comparison of the expected ranks. Hence, the Kruskal–Wallis test is in fact more general than a test for the equality of the medians. It tests whether the probability that a random observation from each group is equally likely to be above or below a random observation from another group. The test is sensitive to differences in medians and that is why it is usually considered a test for the equality of medians.

Example

Consider three student groups (1, 2, 3) and their statistics exam scores. The table below displays the scores and the corresponding ranks (in parenthesis).

Group 1	Group 2	Group 3
18.0 (14)	16.5 (11)	23 (22)
11.0 (4.5)	10.0 (3)	22 (20)
17.0 (12)	15.0 (8.5)	23 (22)
14.0 (7)	15.0 (8.5)	24 (24)
11.0 (4.5)	20.5 (17)	21 (18)
9.5 (2)	8.0 (1)	21.5 (19)
16.0 (10)	12.0 (6)	23 (22)
		20.0 (16)
		17.5 (13)
		19.0 (15)

Example

Calculate the rank means within groups:

$$\bar{r}_1 = \frac{1}{7}(14 + 4.5 + 12 + 7 + 4.5 + 2 + 10) = \frac{54}{7} = 7.714286,$$

$$\bar{r}_2 = \frac{1}{7}(11 + 3 + 8.5 + 8.5 + 17 + 1 + 6) = \frac{55}{7} = 7.857143,$$

$$\bar{r}_3 = \frac{1}{10}(22 + 20 + 22 + 24 + 18 + 19 + 22 + 16 + 13 + 15) = \frac{191}{10} = 19.1,$$

and the mean rank of the combined sample

$$\bar{r} = \frac{1}{24}(54 + 55 + 191) = \frac{300}{24} = 12.5.$$

Calculate the group sum of squares:

$$\sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2 = 7 \cdot (7.714286 - 12.5)^2 + 7 \cdot (7.857143 - 12.5)^2 \\ + 10 \cdot (19.1 - 12.5)^2 = 746.8143$$

and the total sum of squares:

$$\sum_{s=1}^n (R(z_s) - \bar{r})^2 \\ = (14 - 12.5)^2 + (4.5 - 12.5)^2 + (12 - 12.5)^2 + (7 - 12.5)^2 \\ + (4.5 - 12.5)^2 + (2 - 12.5)^2 + (10 - 12.5)^2 + (11 - 12.5)^2 \\ + (3 - 12.5)^2 + (8.5 - 12.5)^2 + (8.5 - 12.5)^2 \\ + (17 - 12.5)^2 + (1 - 12.5)^2 + (6 - 12.5)^2 \\ + (22 - 12.5)^2 + (20 - 12.5)^2 + (22 - 12.5)^2 + (24 - 12.5)^2 \\ + (18 - 12.5)^2 + (19 - 12.5)^2 + (22 - 12.5)^2 \\ + (16 - 12.5)^2 + (13 - 12.5)^2 + (15 - 12.5)^2 \\ = 1147.$$

Example

Now

$$K = (n - 1) \frac{\sum_{j=1}^k n_j (\bar{r}_j - \bar{r})^2}{\sum_{s=1}^n (R(z_s) - \bar{r})^2} = (24 - 1) \frac{746.8143}{1147} = 14.97535.$$

The p -value of the test is clearly less than 0.05 – the value 5.79 corresponds approximately to p -value 0.05. The null hypothesis can be rejected. There is a statistically significant difference in the exam success between the three groups.

Solution using Python:

```
import pandas as pd
from scipy.stats import kruskal
nan = float('nan')
data = pd.DataFrame({
    "1": [18, 11, 17, 14, 11, 9.5, 16, nan, nan, nan],
    "2": [16.5, 10, 15, 15, 20.5, 8, 12, nan, nan, nan],
    "3": [23, 22, 23, 24, 21, 21.5, 23, 20, 17.5, 19]})
K_stat, p_value = kruskal(*data.T.values, nan_policy='omit')
```

Bonferroni's method pairwise comparison

If the null hypothesis of the Kruskal–Wallis test is rejected, then the next step is pairwise comparison.

Bonferroni's method pairwise comparison

Compare the pairwise equality/difference of the medians. There are $c = \frac{k(k-1)}{2}$ pairs to compare.

The first idea would be to use the Wilcoxon two sample rank test for pairwise comparison. It should be noted that if the comparison is done on significance level α , then the pairwise comparisons should be done on significance level $\beta = \frac{\alpha}{c}$. For example, if the significance level 0.05 is used for the combined comparison, then the pairwise comparisons can be used to reject the null hypothesis if the p -value is smaller than or equal to $\frac{0.05}{c}$.

Numerical example

Previously we applied ANOVA to examine whether the expected value of a specific laboratory test L differs between patients that are on different medications (A, B, C). The null hypothesis was rejected. We are now a bit worried about the normality assumption and we decide to conduct pairwise comparisons using the Wilcoxon two sample rank test. We decide to use significance level 0.05.

Group 1 (A)	Group 2 (B)	Group 3 (C)
0.111	0.109	0.119
0.123	0.107	0.124
0.109	0.103	0.125
0.120	0.104	0.117
0.115	0.098	0.111
0.112	0.110	0.120
0.117	0.101	0.118
0.110	0.115	0.116
0.119	0.099	0.122
0.116	0.111	0.119

Table: Laboratory test results for groups 1, 2, and 3.

There are $c = \frac{k(k-1)}{2} = 3$ pairs to compare. Thus, in pairwise comparison, we reject the null hypothesis (equality of the medians) if the p -value is smaller than or equal to $\frac{0.05}{3} = 0.0166\dots$

Wilcoxon rank sum test with continuity correction

data: A and B

W=91, p-value = 0.002169

alternative hypothesis: true location shift is not equal to 0

Wilcoxon rank sum test with continuity correction

data: A and C

W=26, p-value = 0.07478

alternative hypothesis: true location shift is not equal to 0

Wilcoxon rank sum test with continuity correction

data: B and C

W=1.5, p-value = 0.0002821

alternative hypothesis: true location shift is not equal to 0

Two (A vs. B and B vs. C) of the p -values are smaller than $\frac{0.05}{3} = 0.0166\dots$. We conclude that the median of the laboratory test L of the patients that are on medication B differs from the median of the laboratory test L of the patients that are on medication A or on medication C.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Tenth lecture, December 16, 2024

Stochastic independence

Independence

Two random variables are independent if the result of one does not in any way help us predict the result of the other.

Formally, if $\mathbb{P}(x \in A, y \in B) = \mathbb{P}(x \in A)\mathbb{P}(y \in B)$ for all events A and B , then the random variables x and y are **independent**.

If the above does not hold, then the random variables x and y are **dependent**.

In statistics, the dependence of random variables is of great interest:

- The dependence between the unemployment rate and the GDP growth rate.
- The dependence between alcohol consumption and the price of alcohol.
- The dependence between lung cancer incidences and smoking.

Linear dependence

Let x and y be random variables. Let

$$y = ax + b, \quad a, b \in \mathbb{R}, \quad a \neq 0.$$

Then the random variable y is a linear combination of the variable x and thus the variables x and y are (completely) linearly dependent. Linear dependence between two variables can be measured, for example, using the Pearson correlation coefficient.

Linear dependence

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . Then the **sample covariance**

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

estimates the population covariance

$$\sigma_{xy} = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

and

$$\hat{\rho}(x, y) = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

estimates the **Pearson correlation coefficient**

$$\rho(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Linear dependence

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) .

- If the variables x and y are independent, then

$$\mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}[x - \mathbb{E}[x]]\mathbb{E}[y - \mathbb{E}[y]] = 0$$

and the Pearson correlation coefficient $\rho(x, y) = 0$.

- If $y = ax + b$, $a > 0$ and $b \in \mathbb{R}$, then $\rho(x, y) = 1$.
- If $y = ax + b$, $a < 0$ and $b \in \mathbb{R}$, then $\rho(x, y) = -1$.

In general, the Pearson correlation coefficient is a measure of the strength of linear dependence between two random variables. The coefficient $\rho \in [-1, 1]$.

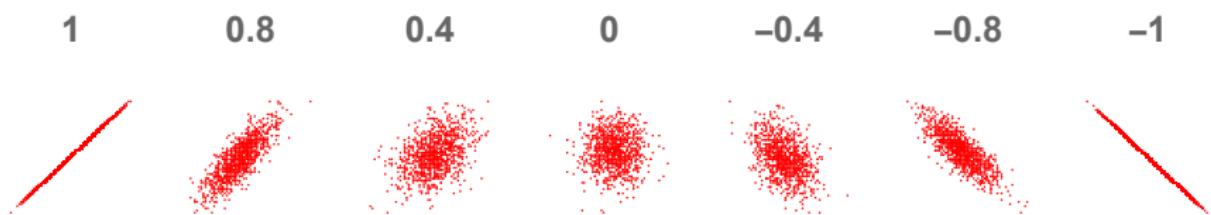
Pearson correlation coefficient

Note that linear independence does not guarantee independence.

For example, if $x \sim \mathcal{U}([-1, 1])$ and $y = x^2$, then the (linear) correlation between the variables x and y is 0, even though they do depend on each other.

Recall that normally distributed random variables are uncorrelated if and only if they are independent.

Example 1



Example 2



Example 3

Correlation coefficients



Probability density function of a bivariate normal distribution:

$$f(x, y) = \frac{1}{2\pi\sqrt{1 - \rho^2(x, y)}\sigma_x\sigma_y} \\ \times \exp\left(-\frac{1}{2(1 - \rho^2(x, y))}\left(\frac{(x - \mu_x)^2}{\sigma_x^2} - 2\rho(x, y)\frac{(x - \mu_x)}{\sigma_x}\frac{(y - \mu_y)}{\sigma_y} + \frac{(y - \mu_y)^2}{\sigma_y^2}\right)\right).$$

Parametric confidence interval

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . Assume that (x, y) follows a bivariate normal distribution. Let

$$l = \frac{(1 + \hat{\rho}(x, y)) - (1 - \hat{\rho}(x, y)) \exp(2z_{\alpha/2}/\sqrt{n-3})}{(1 + \hat{\rho}(x, y)) + (1 - \hat{\rho}(x, y)) \exp(2z_{\alpha/2}/\sqrt{n-3})}$$

and let

$$u = \frac{(1 + \hat{\rho}(x, y)) - (1 - \hat{\rho}(x, y)) \exp(-2z_{\alpha/2}/\sqrt{n-3})}{(1 + \hat{\rho}(x, y)) + (1 - \hat{\rho}(x, y)) \exp(-2z_{\alpha/2}/\sqrt{n-3})},$$

where $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ is the $(1 - \alpha/2) \cdot 100$ percentile of the standard normal distribution.

If the sample size n is large, then (l, u) estimates a level $(1 - \alpha)$ confidence interval for the Pearson correlation coefficient. Note that this confidence interval can only be used under the assumption of bivariate normal distribution. Note also that the confidence intervals for the Pearson correlation coefficient provided by different statistical softwares are almost always based on normality assumption.

Nonparametric confidence interval

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . One can use bootstrapping to obtain nonparametric confidence intervals for the Pearson correlation coefficient:

1. Pick a new random sample of size n from the observed values $(x_1, y_1), \dots, (x_n, y_n)$ with replacement, such that the new values are selected one-by-one and the selected observation is returned back to the original sample. (Note that this means that the same observation can be selected multiple times.)
2. Calculate the Pearson correlation coefficient for the new sample formed in the previous step.

Continued on the next slide!

3. Repeat the previous steps several times and order the obtained estimates from the smallest to the largest. Include also the original estimate of the Pearson correlation coefficient.
4. Calculate an estimate for a $(1 - \alpha)\%$ confidence interval by selecting a lower bound l that is smaller than (or equal to) $1 - \frac{\alpha}{2}$ of the ordered estimates and an upper bound u that is larger than (or equal to) $1 - \frac{\alpha}{2}$ if the estimates. (Assume, for example, that there are 999 bootstrap estimates. Then, in total, there are 1000 estimates – the original one and the 999 new ones. Now, an estimated 90% confidence interval (l, u) is obtained by choosing the 50th ordered estimate as l and the 951st estimate as u . An estimate for the 95% confidence interval (l, u) is obtained by choosing the 25th estimate as l and the 976th estimate as u .)

One sample test for the Pearson correlation coefficient

One sample test for the Pearson correlation coefficient

The one sample test for the Pearson correlation coefficient compares the Pearson correlation coefficient to a given constant.

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . Assume that (x, y) follows a bivariate normal distribution.

The null hypothesis: $H_0: \rho(x, y) = \rho_0$.

The possible alternative hypotheses:

$$H_1: \rho(x, y) > \rho_0 \text{ (one tailed)},$$

$$H_1: \rho(x, y) < \rho_0 \text{ (one tailed)},$$

$$H_1: \rho(x, y) \neq \rho_0 \text{ (two tailed)}.$$

One sample test for the Pearson correlation coefficient

- The test statistic

$$z = \frac{\text{ar tanh}(\hat{\rho}(x, y)) - \text{ar tanh}(\rho_0)}{\sqrt{\frac{1}{n-3}}} = \frac{\frac{1}{2} \log\left(\frac{1+\hat{\rho}(x, y)}{1-\hat{\rho}(x, y)}\right) - \frac{1}{2} \log\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}}.$$

- If the sample size n is large, then under the null hypothesis, the test statistic z approximately follows the standard normal distribution.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Two sample test for Pearson correlation coefficients

Two sample test for Pearson correlation coefficients

The two sample test (correlation comparison test) compares the Pearson correlation coefficients of two independent samples.

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) and let $(z_1, w_1), \dots, (z_m, w_m)$ be i.i.d. observations of a bivariate random variable (z, w) . Assume that (x, y) follows a bivariate normal distribution with Pearson correlation coefficient $\rho(x, y)$ and that (z, w) follows a bivariate normal distribution with Pearson correlation coefficient $\rho(z, w)$. Assume that (x_i, y_i) and (z_j, w_j) are independent for all i, j .

The null hypothesis $H_0: \rho(x, y) = \rho(z, w)$.

The possible alternative hypotheses:

$$H_1: \rho(x, y) > \rho(z, w) \text{ (one tailed)},$$

$$H_1: \rho(x, y) < \rho(z, w) \text{ (one tailed)},$$

$$H_1: \rho(x, y) \neq \rho(z, w) \text{ (two tailed)}.$$

Two sample test for Pearson correlation coefficients

- The test statistic

$$z = \frac{\frac{1}{2} \log \left(\frac{1+\hat{\rho}(x,y)}{1-\hat{\rho}(x,y)} \right) - \frac{1}{2} \log \left(\frac{1+\hat{\rho}(z,w)}{1-\hat{\rho}(z,w)} \right)}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}}.$$

- If n and m are large, then under the null hypothesis, the test statistic z approximately follows the standard normal distribution.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 is false.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Parametric significance test

Parametric significance test

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . Assume that (x, y) follows a bivariate normal distribution.

The null hypothesis $H_0: \rho(x, y) = 0$.

The possible alternative hypotheses:

$H_1: \rho(x, y) > 0$ (one tailed),

$H_1: \rho(x, y) < 0$ (one tailed),

$H_1: \rho(x, y) \neq 0$ (two tailed).

Parametric significance test

- The test statistic

$$t = \hat{\rho}(x, y) \sqrt{\frac{n - 2}{1 - \hat{\rho}(x, y)^2}}.$$

- Under the null hypothesis, the test statistic follows Student's t -distribution with $n - 2$ degrees of freedom.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 does not hold.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Nonparametric significance test

Nonparametric significance test

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) .

The null hypothesis $H_0: \rho(x, y) = 0$.

The possible alternative hypotheses:

$H_1: \rho(x, y) > 0$ (one tailed),

$H_1: \rho(x, y) < 0$ (one tailed),

$H_1: \rho(x, y) \neq 0$ (two tailed).

Nonparametric significance test

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . The significance of the observed Pearson sample correlation coefficient under the null hypothesis can be assessed using a Monte Carlo permutation test:

1. Form n new pairs $(x_1, y_1^*), \dots, (x_n, y_n^*)$ from the original observed values $(x_1, y_1), \dots, (x_n, y_n)$ such that each original y_j is used exactly once in the new sample.
2. Calculate the Pearson correlation coefficient $\hat{\rho}(x, y^*)$ for the sample $(x_1, y_1^*), \dots, (x_n, y_n^*)$.
3. Repeat steps 1 and 2 several times and estimate the probability of the estimate $\hat{\rho}(x, y)$ under the null hypothesis using the values from step 2.
 - That is, calculate the percentage of the estimates in step 2 that
 - satisfy $\hat{\rho}(x, y^*) \geq \hat{\rho}(x, y)$ (one tailed $H_1: \rho(x, y) > 0$);
 - satisfy $\hat{\rho}(x, y^*) \leq \hat{\rho}(x, y)$ (one tailed $H_1: \rho(x, y) < 0$);
 - satisfy $|\hat{\rho}(x, y^*)| \geq |\hat{\rho}(x, y)|$ (two tailed $H_1: \rho(x, y) \neq 0$).

Remark. A more accurate procedure can be obtained by using the permutation test without simulations: instead of simulating new pairs, all the $n!$ possible combinations are used. The probability of $\hat{\rho}(x, y)$ under the null hypothesis is estimated using all $n!$ correlation coefficients.

Spearman (rank) correlation coefficient

Monotonic dependence

Let x and y be random variables. Let $y = g(x)$, where g is a monotonic (increasing or decreasing) function. Then the variable y is a monotonic function of the variable x and the variables x and y are (completely) monotonically dependent.

The monotonic dependence between two random variables can be measured using the Spearman rank correlation coefficient.

Spearman correlation coefficient

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) . Let $R(x_i)$, $i \in \{1, \dots, n\}$, be the rank of the observation x_i in the sample x_1, \dots, x_n and let $R(y_i)$, $i \in \{1, \dots, n\}$, be the rank of the observation y_i in the sample y_1, \dots, y_n .

The **Spearman rank correlation coefficient** $\rho_S(x, y)$ is the Pearson correlation coefficient calculated for the rank sample

$$(R(x_1), R(y_1)), \dots, (R(x_n), R(y_n)).$$

The Spearman correlation coefficient is a measure of the strength of **monotonic dependence** between the two random variables. The coefficient $\rho_S \in [-1, 1]$.

Confidence intervals for the Spearman correlation coefficient can be estimated using bootstrap.

Significance test

Let $(x_1, y_1), \dots, (x_n, y_n)$ be i.i.d. observations of a bivariate random variable (x, y) .

The null hypothesis $H_0: \rho_S(x, y) = 0$.

The possible alternative hypotheses:

$$H_1: \rho_S(x, y) > 0 \text{ (one tailed)},$$

$$H_1: \rho_S(x, y) < 0 \text{ (one tailed)},$$

$$H_1: \rho_S(x, y) \neq 0 \text{ (two tailed)}.$$

Significance test

- The test statistic

$$t = \hat{\rho}_S(x, y) \sqrt{\frac{n - 2}{1 - \hat{\rho}_S(x, y)^2}}.$$

- If n is large, then under the null hypothesis the test statistic t approximately follows Student's t -distribution with $n - 2$ degrees of freedom. If the sample size is small, statistical software can be used to calculate exact p -values for the test statistic.
- The expected value of the test statistic is 0.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 is not true.
- The null hypothesis H_0 is rejected if the p -value is small enough.

The significance of the Spearman rank correlation coefficient can alternatively be tested using the permutation test.

Spearman rank correlation coefficient

It is possible that some of the sample points have the same rank. In that case, all those points are assigned to have the median of the corresponding ranks. For example, if two observations have the same rank, corresponding to ranks 7 and 8, then both are assigned to have rank 7.5. If three observations have the same rank, corresponding to ranks 3, 4, and 5, then each is assigned to have rank 4.

Numerical example

Twin sisters were asked to rank different cookie brands according to the taste. The goal was to test, on significance level 5%, whether the cookie preferences were monotonically dependent. The null hypothesis is $\rho(x, y) = 0$.

rank	10	9	8	7	6	5	4	3	2	1
X (twin 1)	J	G	D	H	A	C	B	I	E	F
Y (twin 2)	G	H	D	C	A	B	J	E	I	F

Table: Cookie preferences of the twins.

The tabulated values can be converted to rank pairs:

(6, 6), (4, 5), (5, 7), (8, 8), (2, 3), (1, 1), (9, 10), (7, 9), (3, 2), (10, 4).

The sample standard deviations are $s_X = 3.02765$ and $s_Y = 3.02765$ and the sample covariance is $s_{XY} = 6.5$. The Spearman rank correlation coefficient is $\hat{\rho}_S(X, Y) = 0.7090909$. The test statistic has the value

$$t = \hat{\rho}_S(X, Y) \sqrt{\frac{n - 2}{1 - \hat{\rho}_S(X, Y)^2}} = \frac{0.7090909 \cdot \sqrt{8}}{1 - (0.7090909)^2} = 2.844367.$$

Under the null hypothesis, the test statistic approximately follows Student's t -distribution with $10 - 2 = 8$ degrees of freedom. The critical values on significance level 5% are -2.306 and 2.306 . Since $2.844 > 2.306$, the null hypothesis is rejected and the alternative hypothesis is accepted. The cookie preferences of the twins are monotonically dependent.

Q: What went wrong with the previous example?

A: The sample size in this example is quite small, so using asymptotic p -values is questionable. It would be better to use the exact p -value computed using statistical software or to use the permutation test.

Words of warning

Dependence \neq linear dependence!

Dependence does not imply causation! See *Spurious Correlations*:
<https://www.tylervigen.com/spurious-correlations>

Regression analysis

Regression analysis

The aim in regression analysis is to study how a dependent variable changes when one or more explanatory variables are varied. It can be used to study, e.g., if the number of violent crimes depends on alcohol consumption and if it does, how strong is this dependence.

- Does salary depend on the education level and if it does, how strong is this dependence?
- Does a parent's smoking have an effect on the height of a child and if it does, how strong is this dependence?
- Do crime rates depend on the income inequality level and if yes, how strong is this dependence?
- :

Possible goals in regression analysis:

- Description of the dependence between the explanatory and dependent variables. What is the type of the relationship? How strong is the dependence?
- Predicting the values of the dependent variable.
- Controlling the values of the dependent variable.

Linear model

There are several different models that can be used in regression analysis. Today, we focus on linear regression.

Consider n observations (pairs) $(x_1, y_1), \dots, (x_n, y_n)$ of (x, y) . Assume that the values y_i are observed values of a random variable y and assume that the values x_i are observed non-random values of x . Assume that the values y_i depend linearly on the value x_i . A simple (one explanatory variable) **linear model** can be represented in the following way:

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where the **regression coefficients** b_0 and b_1 are unknown constants and the expected value of the **residuals** ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

Simple linear model

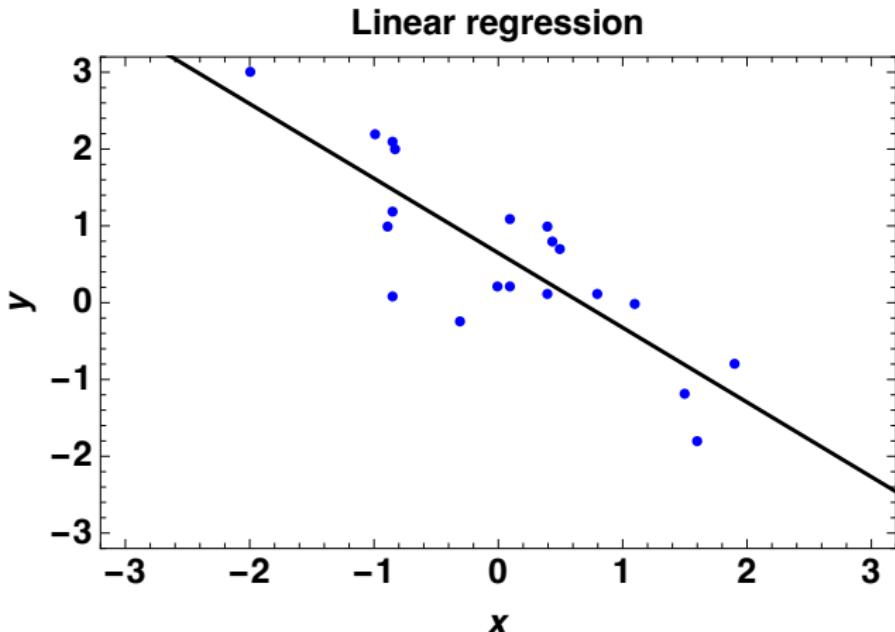


Figure: As the values of the variable x increase, the values of the variable y decrease.

Linear model, assumptions

The following assumptions are usually made when simple linear models are considered.

- The measurement of the values x_i is error-free.
- The residuals are independent of the values x_i .
- The residuals are independently and identically distributed (i.i.d.).
- The expected value of the residuals is $\mathbb{E}[\varepsilon_i] = 0$, $i \in \{1, \dots, n\}$.
- The residuals have the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$, $i \in \{1, \dots, n\}$.
- The residuals are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

Under these assumptions, the variable y has the following properties:

- The expected value $\mathbb{E}[y_i] = b_0 + b_1 x_i$, $i \in \{1, \dots, n\}$.
- The variance $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2$, $i \in \{1, \dots, n\}$.
- The correlation coefficient $\rho(y_i, y_j) = 0$, $i \neq j$.

Linear regression

Linear regression

The linear model

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

has the following parameters: the regression coefficients b_0 and b_1 , and the variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. These parameters are usually unknown and must be estimated from the observations.

Under the assumption $\mathbb{E}[\varepsilon_i] = 0$ for all $i \in \{1, \dots, n\}$, the linear model can be given as

$$y_i = \mathbb{E}[y_i] + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where $\mathbb{E}[y_i] = b_0 + b_1 x_i$ is the so-called **systematic part** and ε_i is the **random part of the model**.

Regression line

The systematic part

$$\mathbb{E}[y_i] = b_0 + b_1 x_i$$

of the linear model defines the regression line

$$y = b_0 + b_1 x,$$

where

- b_0 is the intersection of the regression line and the y -axis;
- the slope b_1 tells us how much the independent variable y changes when the explanatory variable x grows by one unit;
- the variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ describes the deviation of the observed values from the regression line.

The aim in linear regression analysis is to find estimates for the regression coefficients b_0 and b_1 . The estimates should be such that the estimated regression line would explain the variation of the values of the dependent variable with great accuracy.

Least squares method

In the so-called l_2 regression (least squares method), the least squares estimates are

$$\hat{b}_1 = \frac{s_{xy}}{s_x^2} = \hat{\rho}(x, y) \frac{s_y}{s_x}$$

and

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

These estimates minimize the sum of squared differences

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

The least squares estimates now give an estimated regression line

$$\begin{aligned}\hat{y} &= \hat{b}_0 + \hat{b}_1 x = \bar{y} - \hat{b}_1 \bar{x} + \hat{\rho}(x, y) \frac{s_y}{s_x} x \\ &= \bar{y} + \hat{\rho}(x, y) \frac{s_y}{s_x} (x - \bar{x}).\end{aligned}$$

Properties of the estimated regression line:

- If $\hat{\rho}(x, y) > 0$, then the line is increasing.
- If $\hat{\rho}(x, y) < 0$, then the line is decreasing.
- If $\hat{\rho}(x, y) = 0$, then the line is horizontal.

Fitted values and residuals

The fitted value of the variable y_i , i.e., the value given to the variable y by the regression line at points x_i , is

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i, \quad i \in \{1, \dots, n\}.$$

The residual $\hat{\varepsilon}_i$ of the estimated model is the difference

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i \in \{1, \dots, n\},$$

between the observed value y_i (of the variable y) and the fitted value \hat{y}_i .

Note that $y_i = \hat{y}_i + \hat{\varepsilon}_i$, $i \in \{1, \dots, n\}$.

The regression model explains the observed values of the dependent variables the better, the closer the fitted values are to the observed ones.
In other words, the regression model explains the observed values of the dependent variable the better, the smaller the residuals of the estimated model are.

Example

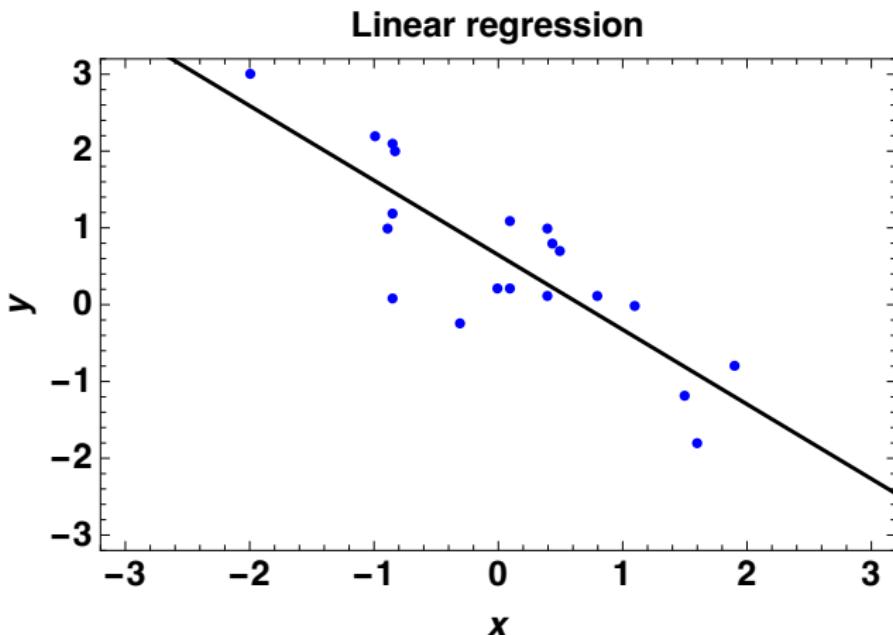


Figure: The estimated regression line minimizes the squared sum of the residuals.

Numerical example

We wish to model the dependence of the sales of Brand X cookies and Brand Y cookies. We assume that the sales are linearly dependent, and try to apply linear regression.

Brand X	Brand Y
5673	5489
4892	5987
5735	5362
5382	5738
5982	4988
5487	5576
5764	5481
5933	4999
5298	5832
5561	5591
5721	5298
5386	5632

Table: Monthly sales of Brand X and Brand Y cookies.

The sample standard errors $s_X = 302.95$ and $s_Y = 302.85$, the sample covariance $s_{XY} = -86145.95$, and the sample means $\bar{X} = 5567.833$ and $\bar{Y} = 5497.75$. The estimated regression parameters

$$\hat{b}_1 = \frac{s_{XY}}{s_X^2} = \frac{-86145.95}{302.95^2} = -0.938\dots$$

and

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 5497.75 - (-0.938\dots) \cdot 5567.833 = 10723.87.$$

An estimated regression model can now be given as

$$\hat{Y}_i = 10723.87 - 0.938X_i.$$

Fit	Actual	Residual
5399.040	5489	89.96
6132.108	5987	-145.11
5340.845	5362	21.15
5672.181	5738	65.82
5109.004	4988	-121.00
5573.625	5576	2.38
5313.625	5481	167.37
5154.997	4999	-156.00
5751.025	5832	80.97
5504.166	5591	86.83
5353.986	5298	-55.99
5668.426	5632	-36.43

Table: Fitted values and actual sales of Brand X cookies. The residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$ have been tabulated as well.

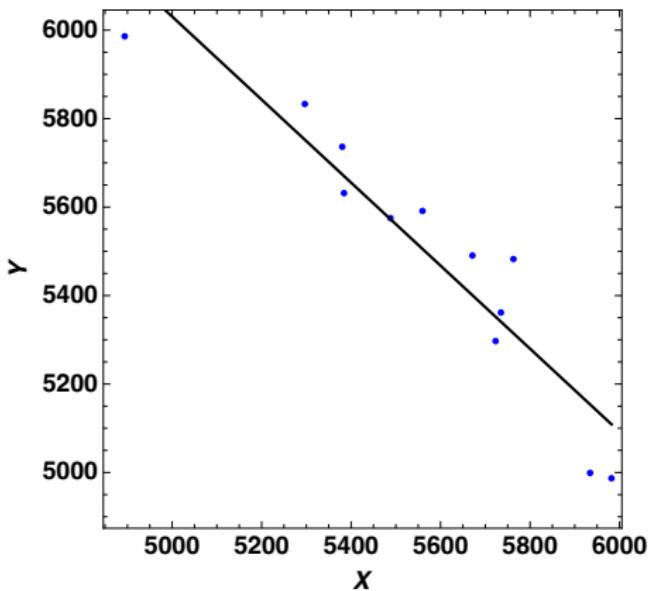


Figure: Brand Y cookies, sales and fit.

Residual mean square estimation

If the assumptions of the linear model hold, then an unbiased estimate of $\text{Var}(\varepsilon_i) = \sigma^2$ is

$$\text{Var}(\hat{\varepsilon}) = \frac{1}{n-2} \sum_{i=1}^n (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-2} \sum_{i=1}^n (\hat{\varepsilon}_i)^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

In the formula above, the number of the estimated parameters (b_0 and b_1) is subtracted from the sample size n .

Error sum of squares

Consider the total sum of squares (SST)

$$\sum_{i=1}^n (y_i - \bar{y})^2,$$

and the error sum of squares (SSE)

$$\sum_{i=1}^n (\hat{\varepsilon}_i)^2.$$

It can be shown that

$$SSE = \sum_{i=1}^n (\hat{\varepsilon}_i)^2 = (1 - \hat{\rho}(x, y)^2) \sum_{i=1}^n (y_i - \bar{y})^2 = (1 - \hat{\rho}(x, y)^2) SST.$$

Since $\hat{\rho}(x, y) \in [-1, 1]$, we have that $SSE \leq SST$.

Error sum of squares

The error sum of squares SSE is 0 if and only if all the observed values lie on the same line. In this case, the linear regression model explains the values of the dependent variable perfectly.

The error sum of squares SSE equals the total sum of squares if and only if the sample correlation coefficient $\hat{\rho}(x, y) = 0$. In this case, the linear regression model fails to explain any part of the values of y .

Model sum of squares

The model sum of squares SSM is defined as

$$SSM = SST - SSE.$$

The model sum of squares SSM describes the part of variation of the observed values of y that is explained by the regression model.

There holds

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

and since $\bar{y} = \bar{\hat{y}}$, the equation can be given as

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2.$$

Coefficient of determination

The coefficient of determination is defined as

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}.$$

The coefficient of determination R^2 measures the proportion of SST explained by the model.

There holds $0 \leq R^2 \leq 1$, and the coefficient of determination is usually given as a percentage $100R^2\%$.

The coefficient of determination $R^2 = (\hat{\rho}(y, \hat{y}))^2$, where $\hat{\rho}(y, \hat{y})$ is the sample correlation coefficient of the observed values of the dependent variable and the corresponding fitted values. In a simple linear regression model with one explaining variable, $R^2 = (\hat{\rho}(y, x))^2$.

Properties of the coefficient of determination

The following conditions are equivalent:

- The coefficient of determination $R^2 = 1$.
- All the residuals vanish: $\hat{\varepsilon}_i = 0, i \in \{1, \dots, n\}$.
- All the observations (x_i, y_i) lie on the same line.
- The sample correlation coefficient $\hat{\rho}(x, y) = \pm 1$.
- The regression model completely explains the variation of the observed values of the dependent variable y .

Properties of the coefficient of determination

The following conditions are equivalent:

- The coefficient of determination $R^2 = 0$.
- The regression coefficient $\hat{b}_1 = 0$.
- The sample correlation coefficient $\hat{\rho}(x, y) = 0$.
- The regression model fails completely in explaining the variation of the observed values of the dependent variable y .

Numerical example

The numerical example from above continues...

Calculate the total sum of squares

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^{12} (y_i - 5497.75)^2 = 1008932.25,$$

the error sum of squares

$$SSE = \sum_{i=1}^n (\hat{\varepsilon}_i)^2 = 119482.3$$

and the model sum of squares

$$SSM = SST - SSE = 1008932.25 - 119482.3 = 889449.95.$$

Now, the coefficient is determination

$$R^2 = \frac{SSM}{SST} = \frac{889449.95}{1008932.25} \approx 0.8816.$$

Is this a good model?

About the assumptions: We assumed above that the values x_i of the explanatory variable x are non-random. In linear regression, the values x_i can very well also be assumed to be random.

Words of warning:

- The regression model should not be used to predict any values of the range of x . Tail behavior can differ from majority of the data.
- If there is nonlinear dependence between x and y , then linear regression is not a suitable approach.
- The least squares method (l_2 regression) is very sensitive to outliers (i.e., it is non-robust).

Example, linear regression

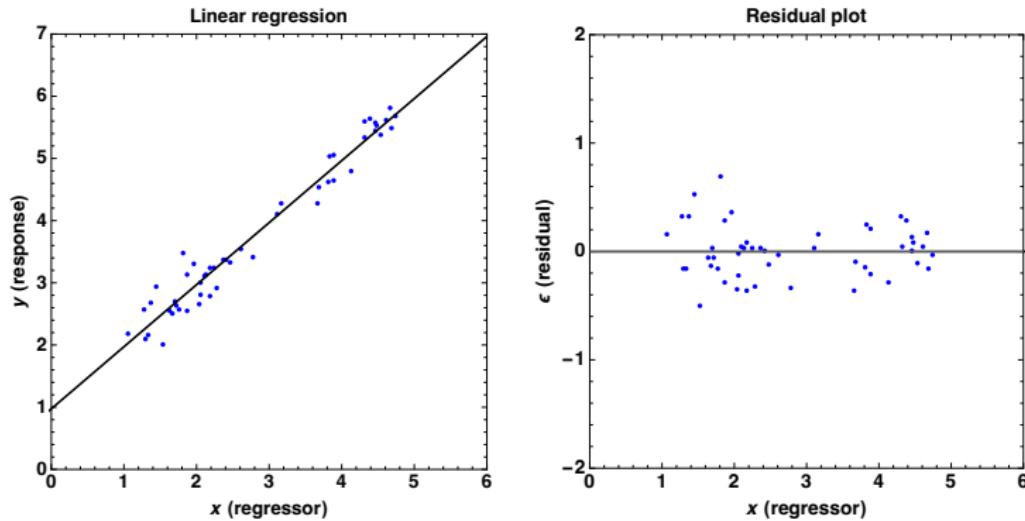


Figure: Estimated regression line and residuals.

Example, outlier

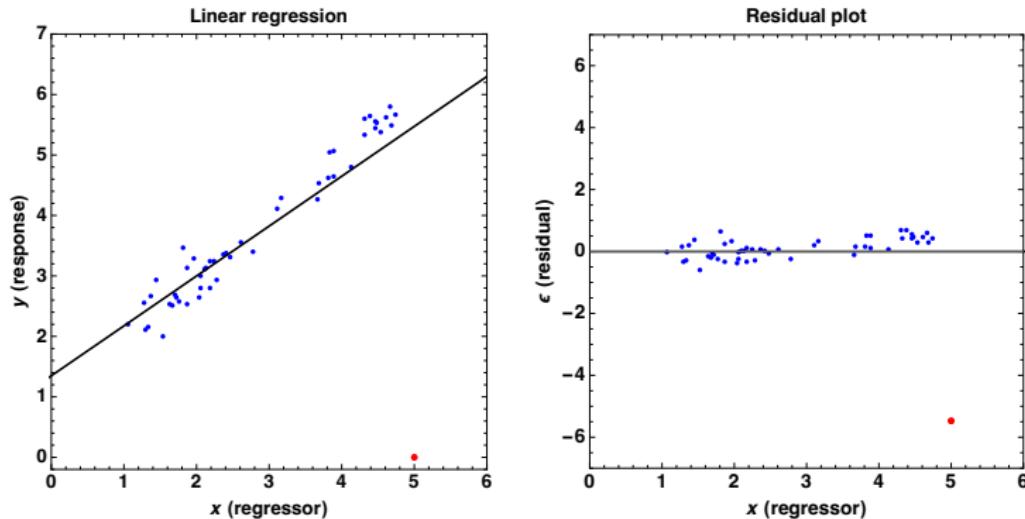


Figure: Estimated regression line and residuals. Note the effect of an outlier.

Example, heteroscedasticity

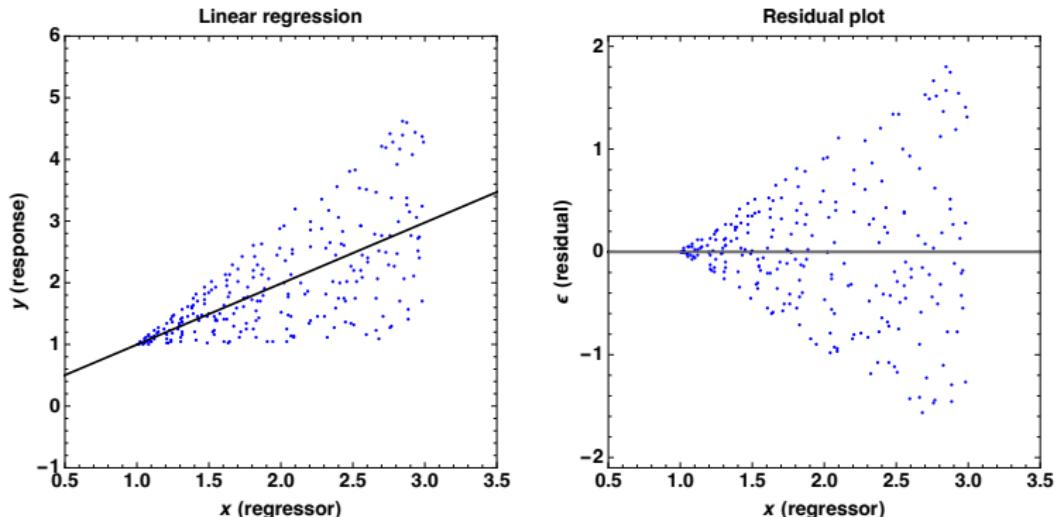


Figure: Estimated regression line and residuals. Note that the variance of the residuals increases.

Example, non-linear dependence

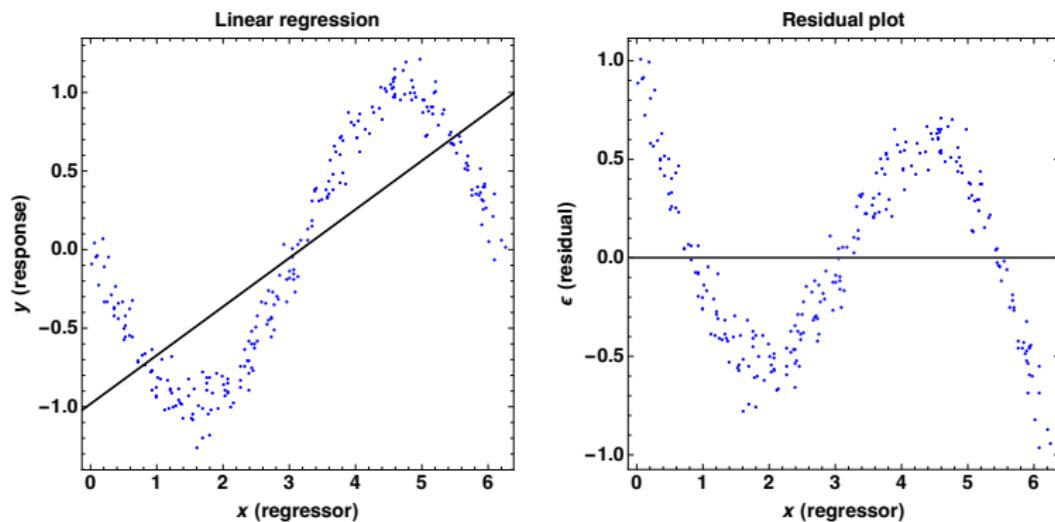


Figure: Estimated regression line and residuals. Note the clear non-linear dependence.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eleventh lecture, January 6, 2025

Tests and confidence intervals for linear regression

Consider n observations (pairs) $(x_1, y_1), \dots, (x_n, y_n)$ of (x, y) . Assume that the values y_i are observed values of a random variable y and assume that the values x_i are observed non-random values of x . Assume that the values y_i depend linearly on the value x_i . A simple (one explanatory variable) **linear model** can be presented in the following way:

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where the **regression coefficients** b_0 and b_1 are unknown constants and the expected value of the residuals ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

Linear model, assumptions for parametric tests and confidence intervals

We now consider testing the parameters of a linear regression model and calculating confidence intervals for the estimated parameters under classical assumptions.

- Measurement of the values x_i is error-free.
- The residuals are independent of the values x_i .
- The residuals are independently and identically distributed (i.i.d.).
- The expected value of the residuals is $\mathbb{E}[\varepsilon_i] = 0$.
- The residuals have the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.
- The residuals are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.
- The residuals are normally distributed.

Slope of the regression line

Testing the slope of the regression line

The null hypothesis:

$$H_0: b_1 = b_1^0$$

(typically null hypothesis $b_1 = 0$ is tested).

Possible alternative hypotheses:

$$H_1: b_1 > b_1^0 \text{ (one tailed),}$$

$$H_1: b_1 < b_1^0 \text{ (one tailed),}$$

$$H_1: b_1 \neq b_1^0 \text{ (two tailed).}$$

Testing the slope of the regression line

- t -test statistic

$$t = \frac{\hat{b}_1 - b_1^0}{s/(s_x \sqrt{n-1})},$$

where $s^2 = \text{Var}(\hat{\varepsilon}) = \frac{1}{n-2} \sum_{i=1}^n (\hat{\varepsilon}_i)^2$ (see previous lecture) and s_x^2 is the sample variance of the variable x .

- Under the null hypothesis H_0 , the test statistic follows Student's t -distribution with $n - 2$ degrees of freedom.
- Under the null hypothesis H_0 , the expected value of the test statistic is $\mathbb{E}[t] = 0$.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 does not hold.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Slope of the regression line, confidence interval

Under the normality assumption on the residuals, the $(1 - \alpha) \cdot 100\%$ confidence interval for the slope of the regression line can be given as

$$\left(\hat{b}_1 - t_{n-2,\alpha/2} \frac{s}{s_x \sqrt{n-1}}, \hat{b}_1 + t_{n-2,\alpha/2} \frac{s}{s_x \sqrt{n-1}} \right),$$

where $s^2 = \text{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x , t_{n-2} is Student's t distribution with $n - 2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the $t(n - 2)$ distribution.

Intercept/constant term

Testing the constant term of the regression line

The null hypothesis:

$$H_0: b_0 = b_0^0.$$

Possible alternative hypotheses:

$$H_1: b_0 > b_0^0 \text{ (one tailed),}$$

$$H_1: b_0 < b_0^0 \text{ (one tailed),}$$

$$H_1: b_0 \neq b_0^0 \text{ (two tailed).}$$

Testing the constant term of the regression line

- t -test statistic

$$t = \frac{\hat{b}_0 - b_0^0}{\frac{s\sqrt{\sum_{i=1}^n x_i^2}}{s_x \sqrt{n(n-1)}}},$$

where $s^2 = \text{Var}(\hat{\varepsilon}) = \frac{1}{n-2} \sum_{i=1}^n (\hat{\varepsilon}_i)^2$ and s_x^2 is the sample variance of the variable x .

- Under the null hypothesis H_0 , the test statistic follows Student's t -distribution with $n - 2$ degrees of freedom.
- Under the null hypothesis H_0 , the expected value of the test statistic is $\mathbb{E}[t] = 0$.
- Large absolute values of the test statistic suggest that the null hypothesis H_0 does not hold.
- The null hypothesis H_0 is rejected if the p -value is small enough.

Intercept, confidence interval

Under normality assumption, $(1 - \alpha) \cdot 100\%$ confidence interval for the constant term of the regression line can be given as

$$\left(\hat{b}_0 - t_{n-2,\alpha/2} \frac{s \sqrt{\sum_{i=1}^n x_i^2}}{s_x \sqrt{n(n-1)}}, \hat{b}_0 + t_{n-2,\alpha/2} \frac{s \sqrt{\sum_{i=1}^n x_i^2}}{s_x \sqrt{n(n-1)}} \right),$$

where $s^2 = \text{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x , t_{n-2} is Student's t -distribution with $n - 2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the $t(n - 2)$ distribution.

Predicting

Predicting the values of variable y

A prediction \tilde{y} for the value of the variable y , when x has value \tilde{x} , can be given as

$$\tilde{y}|\tilde{x} = \hat{b}_0 + \hat{b}_1\tilde{x}.$$

The more there are observations, the smaller the variance σ^2 is, and the closer \tilde{x} is to the sample mean of x , then the better (more accurate) the prediction is. Note that \tilde{x} should be on the range of the observed values of the variable x .

Predicting the values of variable y

Under normality assumption, a $(1 - \alpha) \cdot 100\%$ confidence interval for the value of y , when x has value \tilde{x} , can be given as

$$\hat{b}_0 + \hat{b}_1 \tilde{x} \pm t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{(n-1)s_x^2}},$$

where $s^2 = \text{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x , t_{n-2} is Student's t -distribution with $n - 2$ degrees of freedom, and $t_{n-2,\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the $t(n - 2)$ distribution.

Predicting the expected value of variable y

A prediction \hat{y}_y for the expected value $\mathbb{E}[y]$, when x has value \tilde{x} , can be given as

$$\hat{y}_y|\tilde{x} = \hat{b}_0 + \hat{b}_1\tilde{x}.$$

Remarks:

- Note that $\tilde{y}|\tilde{x}$ estimates the value of a random variable while $\hat{y}_y|\tilde{x}$ estimates the expected value (constant). The estimate $\tilde{y}|\tilde{x}$ estimates the values of the variable on an individual level when x has value \tilde{x} , while the estimate $\hat{y}_y|\tilde{x}$ estimates the mean value of the variable y when x has value \tilde{x} .
- Even though the estimates are the same, the corresponding confidence intervals are not! The confidence interval for the value of y is wider. It is easier to predict average behavior than to predict individual values.

Predicting the expected value of variable y

Under normality assumption, a $(1 - \alpha) \cdot 100\%$ confidence interval for $\mathbb{E}[y]$, when x has value \tilde{x} , can be given as

$$\hat{b}_0 + \hat{b}_1 \tilde{x} \pm t_{n-2, \alpha/2} s \sqrt{\frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{(n-1)s_x^2}},$$

where $s^2 = \text{Var}(\hat{\varepsilon})$, s_x^2 is the sample variance of the variable x , t_{n-2} is Student's t -distribution with $n - 2$ degrees of freedom, and $t_{n-2, \alpha/2}$ is the $(1 - \alpha/2) \cdot 100$ percentile of the $t(n - 2)$ distribution.

Numerical example

Last lecture, we obtained the regression model

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x, \quad \hat{b}_0 = 10723.87 \text{ and } \hat{b}_1 = -0.9386$$

for the cookie sales of Brand Y (dependent variable) with respect to the cookie sales of Brand X (explanatory variable). We wish to derive the 95% confidence interval for the sales when 5500 units of Brand X cookies are sold.

On the condition that $\tilde{c} = 5500$ units of Brand X cookies are sold, the prediction of the sales of Brand Y cookies is

$$\tilde{y}|\tilde{c} = \hat{b}_0 + \hat{b}_1 \tilde{c} = 10723.87 - 0.9386 \cdot 5500 = 5561.57.$$

The corresponding confidence interval can be given as

$$\hat{b}_0 + \hat{b}_1 \tilde{c} \pm t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(\tilde{c} - \bar{c})^2}{(n-1)s_c^2}} = 5561.57 \pm 257.974,$$

where we plugged in the values $t_{n-2,\alpha/2} = t_{10,0.025} = 2.228$, $\bar{c} = 5567.833$, $s_c = 302.95$, and $s^2 = 11948.42$.

\therefore If 5500 units of Brand X cookies are sold, then the prediction for the sales of Brand Y cookies is 5562 units. A 95% confidence interval for the prediction is (5308,5816).

Bootstrap confidence intervals

Bootstrap confidence intervals for the regression coefficients

Consider the estimated residuals $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ and the fitted values $\hat{y}_1, \dots, \hat{y}_n$ of the regression model. Collect a new sample $\check{\varepsilon}_1, \dots, \check{\varepsilon}_n$ by picking n data points randomly with replacement from $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$. Form a bootstrap sample

$$(x_1, \check{y}_1), \dots, (x_n, \check{y}_n),$$

where

$$\check{y}_i = \hat{y}_i + \check{\varepsilon}_i.$$

Calculate estimates for the regression coefficients b_0 and b_1 from the bootstrap sample. Repeat this several times, for example 999 times. Order now all the estimates (the original ones and the 999 bootstrap estimates) from the smallest to the largest. Now an estimate for the 90% confidence interval (l, u) is obtained by choosing the 50th ordered estimate as l and the 951st estimate as u . An estimate for the 95% confidence interval (l, u) is obtained by choosing the 25th estimate as l and the 976th estimate as u .

Prediction, bootstrap confidence intervals

A prediction $\hat{\mu}_y$ for the expected value $\mathbb{E}[y]$, when x has value \tilde{x} , was given as

$$\hat{\mu}_y|\tilde{x} = \hat{b}_0 + \hat{b}_1\tilde{x}.$$

Consider bootstrap estimates for the regression coefficients b_0 and b_1 . One can calculate bootstrap confidence intervals for $\hat{\mu}_y|\tilde{x}$ by replacing \hat{b}_0 and \hat{b}_1 by bootstrap estimates in the formula above. That is then repeated, for example, 999 times. After that, all the 1000 predictions are ordered and bootstrap confidence intervals are obtained.

Coefficient of determination, bootstrap confidence intervals

Bootstrap samples

$$(x_1, \check{y}_1), \dots, (x_n, \check{y}_n)$$

can be used also for calculating bootstrap confidence intervals for the coefficient of determination of the model. Coefficient of determination is estimated (separately) from every bootstrap sample. One can use, for example, 999 bootstrap samples. After that, all the 1000 estimates are ordered and bootstrap confidence intervals are obtained.

Bootstrap confidence intervals, alternative approach

Instead of bootstrapping from the estimated residuals, one may take bootstrap samples directly from the original observations $(x_1, y_1), \dots, (x_n, y_n)$. Parameter estimates are then calculated from the bootstrap samples, the estimates are ordered and bootstrap confidence intervals are obtained.

Multivariate linear regression

Multiple linear model

Consider n observations (pairs) $(x_1, y_1), \dots, (x_n, y_n)$ of (x, y) . Assume that the values y_i are observed values of a random variable y and assume that the values x_i are observed non-random values of a p -dimensional x . (Here, x_i is a p -vector.) Assume that $p < n$ and that the values of the variable y depend linearly on the values of the variable x .

A **multiple linear model** can be presented in the following way

$$y_i = b_0 + b_1(x_i)_1 + b_2(x_i)_2 + \dots + b_p(x_i)_p + \varepsilon_i, \quad i \in \{1, \dots, n\}, \quad (1)$$

where the **regression coefficients** b_0, \dots, b_p are unknown constants and the expected value of the residuals ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

The model (1) can also be expressed in vectorized form as

$$y_i = b_0 + b^T x_i + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where $b = [b_1, \dots, b_p]^T$ and $x_i = [(x_i)_1, \dots, (x_i)_p]^T$.

Linear model, general assumptions

The following assumptions are usually made when multiple linear models are considered.

- The measurement of the values x_i is error-free.
- The values $(x_i)_s, (x_i)_k, s \neq k$, are mutually independent.
- The residuals are independent of the values x_i .
- The residuals are independently and identically distributed (i.i.d.).
- The expected value of the residuals is $\mathbb{E}[\varepsilon_i] = 0$.
- The residuals have the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2, i = 1, \dots, n$.
- The residuals are uncorrelated, i.e., $\rho(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.

Under the assumptions above, the variable y has the following properties:

- The expected value $\mathbb{E}[y_i] = b_0 + b^T x_i, i = 1, \dots, n$.
- The variance $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2, i = 1, \dots, n$.
- The correlation coefficient $\rho(y_i, y_j) = 0, i \neq j$.

Multiple linear regression

Multiple linear regression

The multiple linear model

$$y_i = b_0 + b^T x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

has the following parameters: regression coefficients b_0 and $b = (b_1, \dots, b_p)^T$ and the variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. These parameters are usually unknown and must be estimated from the observations.

Under the assumption $\mathbb{E}[\varepsilon_i] = 0$ for all $i = 1, \dots, n$, the linear model can be given as

$$y_i = \mathbb{E}[y_i] + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\mathbb{E}[y_i] = b_0 + b^T x_i$ is the systematic part and ε_i is the random part of the model.

Regression plane

The systematic part of the linear model

$$\mathbb{E}[y_i] = b_0 + b^T x_i$$

defines the regression plane

$$y = b_0 + b^T x.$$

The variance of the residuals $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ describes the deviation of the observed points from the regression plane.

The aim in multiple linear regression analysis is to find estimates for the regression coefficients b_0 and $b = (b_1, \dots, b_p)^T$. The estimates should be such that the estimated regression plane would explain the variation of the values of the dependent variable with great accuracy.

Least squares method

Let $\beta = (b_0, b_1, \dots, b_p)^T$. Let X be an $n \times (p + 1)$ data matrix, where the elements of the first column are all equal to 1 and where the columns $2, \dots, p + 1$ are the observations x_i . Let $Y = (y_1, \dots, y_n)^T$ be an $n \times 1$ data vector.

The least squares estimates for b_0 and $b = (b_1, \dots, b_p)^T$ are given by

$$\hat{\beta} = (\hat{b}_0, \hat{b}_1, \dots, \hat{b}_p)^T = (X^T X)^{-1} X^T Y.$$

These estimates minimize the sum of the squared differences

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - b_0 - b^T x_i)^2.$$

Remark. We assumed above that the matrix $X^T X$ is non-singular. If $X^T X$ is singular, then some of the explanatory variables must be fully linearly dependent. In that case, some of the variables can be excluded from the analysis without losing any information.

Fits and residuals

The least squares estimates now give an estimated regression plane

$$\hat{y} = \hat{b}_0 + \hat{b}^T x.$$

The fitted values of the variable y_i , i.e., the values given to the variable y by the regression plane at point x_i , are

$$\hat{y}_i = \hat{b}_0 + \hat{b}^T x_i, \quad i = 1, \dots, n.$$

The residuals $\hat{\varepsilon}_i$ of the estimated model are the differences

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

of the observed values y_i (of the variable y) and the fitted values \hat{y}_i .

The regression model explains the observed values of the dependent variable the better, the closer the fitted values are to the observed ones. In other words, the regression model explains the observed values of the dependent variable the better, the smaller the residuals of the estimated model are.

Residual mean square estimation

If the assumptions of the linear model hold, then an unbiased estimate of the $\text{Var}(\varepsilon_i) = \sigma^2$ is

$$\text{Var}(\hat{\varepsilon}) = \frac{1}{n - p - 1} \sum_{i=1}^n \hat{\varepsilon}_i^2.$$

(In the formula above, the number of the estimated parameters (b_0, b_1, \dots, b_p) is subtracted from the sample size n .)

Sums of squares

The total sum of squares (SST)

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

measures total variation of the observed values y_i . The error sum of squares (SSE)

$$\sum_{i=1}^n (\hat{\varepsilon}_i)^2$$

measures the variation of the residuals $\hat{\varepsilon}_i$. The model sum of squares (SSM)

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

measures the part of the variation of the dependent variable y that is explained by the regression model.

Coefficient of determination

The coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$$

measures the proportion of SST explained by the model.

There holds $0 \leq R^2 \leq 1$ and the coefficient of determination is usually given as a percentage $100R^2\%$.

Numerical example

The effect of nonpareils and chocolate chops on the mass of cookies is examined in a lab.

Nonpareil	Chocolate chip	Mass
15	5	24
13	7	28
12	9	26
11	7	27
10	10	29
9	12	31
17	2	19
16	4	21
12	8	25
3	15	36

Table: The number of nonpareils and chocolate chips, as well as the measured masses of a sample of cookies.

The least squares estimates for the regression coefficients $(b_0, b_1, b_2)^T$ can be calculated using

$$X = \begin{bmatrix} 1 & 15 & 5 \\ 1 & 13 & 7 \\ 1 & 12 & 9 \\ 1 & 11 & 7 \\ 1 & 10 & 10 \\ 1 & 9 & 12 \\ 1 & 17 & 2 \\ 1 & 16 & 4 \\ 1 & 12 & 8 \\ 1 & 3 & 15 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} 24 \\ 28 \\ 26 \\ 27 \\ 29 \\ 31 \\ 19 \\ 21 \\ 25 \\ 36 \end{bmatrix}.$$

The estimates are

$$(\hat{b}_0, \hat{b}_1, \hat{b}_2)^T = (X^T X)^{-1} X^T Y = (29.9718, -0.6562, 0.5533)^T.$$

Now one obtains the fits $\hat{y}_i = \hat{b}_0 + \hat{b}^T x_i$ for the mass and can calculate the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Nonpareil	Chocolate chip	Mass	Fit	Residual
15	5	24	22.8953	1.1047
13	7	28	25.3143	2.6857
12	9	26	27.0771	-1.0771
11	7	27	26.6267	0.3733
10	10	29	28.9428	0.0572
9	12	31	30.7056	0.2944
17	2	19	19.9230	-0.9230
16	4	21	21.6858	-0.6858
12	8	25	26.5238	-1.5238
3	15	36	36.3027	-0.3027

Table: The effect of nonpareils and chocolate chips on the mass. Also the fitted values and residuals are tabulated.

The sample mean of the mass $\bar{y} = 26.6$ and the total sum of squares

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^{10} (y_i - 26.6)^2 = 214.4.$$

The error sum of squares

$$SSE = \sum_{i=1}^{10} (\hat{\varepsilon}_i)^2 = 13.5586$$

and the model sum of squares

$$SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{10} (\hat{y}_i - 26.6)^2 = 200.8307.$$

Thus, the coefficient of determination is

$$R^2 = \frac{SSM}{SST} = \frac{200.8307}{214.4} = 0.9367 = 93.67\%.$$

Multivariate linear regression

Multivariate linear model

Consider n observations (pairs) $(x_1, y_1), \dots, (x_n, y_n)$ of (x, y) . Assume that the values y_i are the observed values of a q -variate random vector y and assume that the values x_i are observed non-random values of a p -variate x . Assume that $p < n$ and that the values of the variable y depend linearly on the variable x .

A **multivariate linear model** can be given as

$$y_i = b_0 + B^T x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where the elements of a $q \times 1$ vector b_0 and $p \times q$ regression matrix B are unknown constants and the expected value of the **residuals** ε_i is $\mathbb{E}[\varepsilon_i] = 0$.

Linear model, general assumptions

The following assumptions are usually made when multivariate linear models are considered.

- The measurement of the values x_i is error-free.
- The values $(x_i)_s, (x_i)_k$, $s \neq k$, are mutually independent.
- The residuals are independent of the values x_i .
- The residuals are independently and identically distributed (i.i.d.).
- The expected value of the residuals $\mathbb{E}[\varepsilon_i] = 0$, $i = 1, \dots, n$.
- The residuals have the same covariance matrix $\mathbb{E}[\varepsilon_i \varepsilon_i^T] = \Sigma$, $i = 1, \dots, n$.
- The residuals are uncorrelated, i.e., $\rho((\varepsilon_i)_k, (\varepsilon_j)_k) = 0$ for all k and for all $i \neq j$.

Generalized least squares

Let $\beta = (b_0, b_1, \dots, b_p)^T$. Let X be an $n \times (p + 1)$ data matrix, where the elements of the first column are all equal to 1 and where the columns $2, \dots, p + 1$ are the observations x_i . Let Y be an $n \times q$ data matrix, where the columns are the observations y_i .

Now the regression parameters b_0 and B can be estimated using

$$\hat{\beta} = [\hat{b}_0, \hat{B}^T]^T = (X^T X)^{-1} X^T Y.$$

Fits and residuals

The fitted values of the variable y_i , i.e., the values given to the variable y by the regression model at points x_i , are

$$\hat{y}_i = \hat{b}_0 + \hat{B}^T x_i, \quad i = 1, \dots, n.$$

The fits can also be expressed as a matrix

$$\hat{Y} = X\hat{\beta}.$$

The residuals $\hat{\varepsilon}_i$ of the estimated model are the differences

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, \dots, n,$$

of the observed values y_i (of the variable y) and the fitted values \hat{y}_i .

Trace correlation and determinant correlation

Assume that the matrix Y is centered so that the columns of Y have zero mean. (That is, the sample mean is subtracted from the original observations.) Let X be as above, and let $\hat{\beta}$ be calculated for the centered data. Let

$$\hat{Y} = X\hat{\beta},$$

$$\hat{E} = Y - \hat{Y}$$

and let

$$D = (Y^T Y)^{-1} \hat{E}^T \hat{E}.$$

It is straightforward to see that the matrix $\hat{E}^T \hat{E}$ ranges between zero, when all the variation of Y is explained by the regression model, and $Y^T Y$, when no part of the variation in Y is explained by X . Therefore $I - D$ varies between the identity matrix and the zero matrix. It can be shown that all the eigenvalues of $I - D$ lie between 1 and 0.

Trace correlation and determinant correlation

It would be desirable that a multivariate coefficient of determination would range between zero and one. This is obtained by either using trace correlation r_T or determinant correlation r_D :

$$r_T^2 = \frac{1}{p} \text{tr}(I - D),$$

$$r_D^2 = \det(I - D).$$

Note that the coefficient r_D is zero if and only if at least one of the eigenvalues of $I - D$ is zero, while r_T is zero if and only if all the eigenvalues of $I - D$ are zero.

It is possible to construct parametric tests and confidence intervals for the parameters in multiple and multivariate regression analysis. Alternatively, one can consider bootstrapping.

Selecting variables

Selecting variables

In multiple and multivariate regression analysis, the explanatory variables are usually assumed to be independent. Perfect independence is rarely achieved if more than one explanatory variables are used. Still, the explanatory variables may not be highly correlated. **Multicollinearity makes the model unstable and complicates assessing the effects of different explanatory variables separately.**

Variance inflation factor

The variance inflation factor (VIF) can be used to measure the multicollinearity of the explanatory variables. The VIF for the explanatory variable $(x_i)_k$ is defined as

$$VIF_k = \frac{1}{1 - R_k^2},$$

where R_k^2 is the coefficient of determination for a model where $(x_i)_k$ is the dependent variable and the rest of $(x_i)_s$ are explanatory variables. VIF is calculated separately for each explanatory variable $(x_i)_k$. If the variable $(x_i)_k$ is independent from the other explanatory variables, then $VIF = 1$. On the other hand, $VIF \geq 10$ suggests that multicollinearity is present.

In multiple and multivariate regression models the aim is to **select variables such that the coefficient of determination is as high as possible and the explanatory variables are as independent as possible**. VIF (or some other measure of dependence) can be used in selecting the variables. Variables can be added and removed one by one and the changes in VIF and coefficient of determination can be tracked.

Cookie example continues

In this example VIF is used to assess multicollinearity of nonpareils and chocolate chips.

Nonpareil	Chocolate chip
15	5
13	7
12	9
11	7
10	10
9	12
17	2
16	4
12	8
3	15

Table: Cookie data, number of nonpareils and chocolate chips.

The sample standard deviation for nonpareil $s_x = 4.022161$ and chocolate chips $s_y = 3.842742$, the sample means $\bar{x} = 11.8$ and $\bar{y} = 7.9$, and the sample correlation coefficient $\hat{\rho}(x, y) = -0.9647379$ are needed. Fit

$$\hat{y}_i = \bar{y} + \hat{\rho}(x, y) \frac{s_y}{s_x} (x_i - \bar{x}) = 7.8 + (-0.9647379) \frac{3.842742}{4.022161} (x_i - 11.8).$$

Total sum of squares $SST = 113$, error sum of squares $SSE = 9.307418$, and model sum of squares $SSM = 123.6926$. Coefficient of determination

$$R^2 = \frac{SSM}{SST} = \frac{123.6926}{133} = 0.9300195$$

and

$$VIF = \frac{1}{1 - R^2} = \frac{1}{1 - 0.930\ldots} = 14.28969.$$

Words of warning

- Regression models should not be used to predict any values outside of the range of x . Tail behavior can differ from majority of the data.
- If there is nonlinear dependence between x and y , then linear regression is not a suitable approach.
- The least squares method (l_2 regression) is very sensitive to outliers (i.e., it is non-robust).

Parameter identification for non-linear models and the maximum likelihood estimator

Linear regression is a prototypical example of a parameter identification problem. In addition to linear models, one may also be interested in parameter identification for other types of models.

Example

Given i.i.d. normally distributed data $y_1, \dots, y_n \sim \mathcal{N}(\mu, \sigma^2)$, estimate μ and σ^2 .

Example

Given data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$, find parameters $a, b, c \in \mathbb{R}$ such that

$$y_i = ax_i^2 + bx_i + c + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where the residuals ε_i satisfy $\mathbb{E}[\varepsilon_i] = 0$.

Example

Given data $y \in \mathbb{R}^k$, find the unknown parameter $x \in \mathbb{R}^d$ such that $y = Ax + \varepsilon$, where the residual ε satisfies $\mathbb{E}[\varepsilon] = 0$.

Let y_1, \dots, y_n be the data, which are i.i.d. random variables. We assume that these follow a parameter-dependent probability distribution with PDF (resp. PMF) $f(x, y)$ for some realization of the parameter $x \in X$. (With a slight abuse of notation, one might write $y_1, \dots, y_n \stackrel{\text{i.i.d.}}{\sim} f(x, \cdot)$ for some unknown $x \in X$.)

Thus we are interested in identifying the value of the parameter $x \in X$, which (in some sense) best approximates the data out of the set

$$\mathcal{F} = \{f(x, y) \mid x \in X\}$$

containing all the possible candidates for the PDFs $f(x, y)$ which could have generated the data.

A common method to estimate parameters in a parametric models is the **maximum likelihood method**.

Maximum likelihood

Let y_1, \dots, y_n be i.i.d. with the PDF $f(x, y)$.

Definition

The likelihood function is defined by

$$\mathcal{L}_n(x) = \prod_{i=1}^n f(x, y_i).$$

The log-likelihood function is defined by $\ell_n(x) = \log \mathcal{L}_n(x)$.

The likelihood function is simply the joint density of the data, except that we treat it as a function of the parameter x . The likelihood function is not a density function: in general, the function $\mathcal{L}_n(x)$ does not integrate to 1 with respect to x .

Definition

The **maximum likelihood (ML) estimator** is defined as a maximizer of the likelihood function

$$\hat{x}_{\text{ML}} = \arg \max_{x \in X} \mathcal{L}_n(x).$$

Remarks:

- The ML estimator satisfies

$$\mathcal{L}_n(\hat{x}_{\text{ML}}) \geq \mathcal{L}_n(x) \quad \text{for all } x \in X.$$

It answers the question: *Which value of the unknown x is the most likely to produce the measured data?*

- The ML estimator may not be unique.
- The maximum of the log-likelihood $\ell_n(x)$ occurs at the same point as the maximum of $\mathcal{L}_n(x)$. It is often more convenient to work with

$$\hat{x}_{\text{ML}} = \arg \max_{x \in X} \ell_n(x).$$

- Multiplying $\mathcal{L}_n(x)$ by any positive constant c (not depending on x) will not change the ML estimator, so the constants in the likelihood function are often dropped.

Example

Assume that $y_1, \dots, y_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ for some unknown mean parameter $\mu \in \mathbb{R}$. The likelihood function for $\mu \in \mathbb{R}$ is given by

$$\mathcal{L}_n(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i - \mu)^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2}$$

and the log-likelihood is given by

$$\ell_n(\mu) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2.$$

Differentiating this with respect to μ yields

$$\ell'_n(\mu) = \sum_{i=1}^n (y_i - \mu) = n(\bar{y}_n - \mu).$$

Setting this to 0 yields the ML estimator $\hat{\mu}_{\text{ML}} = \bar{y}_n$. Thus the ML estimator coincides with the empirical mean of y_1, \dots, y_n .

Example

Consider n observations (pairs) $(x_1, y_1), \dots, (x_n, y_n)$ of (x, y) . Assume that the values y_i are observed values of a random variable y and assume that the values x_i are observed non-random values of x . Assume that the values y_i depend linearly on the values x_i through a simple linear model

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \quad i \in \{1, \dots, n\},$$

where the residuals are assumed to be Gaussian $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $\sigma > 0$. Writing $b = (b_0, b_1)$, the ML estimator $\hat{b} = \hat{b}_{\text{ML}}$ is given by

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\rho}(x, y) \frac{s_y}{s_x},$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

Example

Assume that $y_1, \dots, y_n \in \mathbb{R}^k$ are i.i.d. realizations of random variable y , which come from some mathematical model

$$y_i = F(x) + \varepsilon_i,$$

where $x \in \mathbb{R}^d$ is the unknown parameter, $F: \mathbb{R}^d \rightarrow \mathbb{R}^k$ is a function, and $\varepsilon_1, \dots, \varepsilon_k$ are i.i.d. realizations of measurement noise ε with PDF ρ_n .

Now

$$\begin{aligned}\mathbb{P}(y \in B) &= \mathbb{P}(F(x) + \varepsilon \in B) = \mathbb{P}(\varepsilon \in B - F(x)) = \int_{B-F(x)} \rho_n(t) dt \\ &= \int_B \rho_n(t - F(x)) dt \quad \text{for all events } B.\end{aligned}$$

This means that $f(x, y_i) = \rho_n(y_i - F(x))$, and the likelihood function is

$$\mathcal{L}_n(x) = \prod_{i=1}^n \rho_n(y_i - F(x)).$$

Example

Assume that $y \in \mathbb{R}^k$ is an observation of the mathematical model

$$y = F(x) + \varepsilon,$$

where $x \in \mathbb{R}^d$ is the unknown parameter and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian measurement noise, with $\sigma > 0$ and I is the $k \times k$ identity matrix. In this case, the noise has the PDF

$$\rho_n(\varepsilon) = \frac{1}{(2\pi\sigma^2)^{k/2}} e^{-\frac{1}{2\sigma^2}\|\varepsilon\|^2}$$

and the likelihood function is given by

$$\mathcal{L}_n(x) = \frac{1}{(2\pi\sigma^2)^{k/2}} e^{-\frac{1}{2\sigma^2}\|y - F(x)\|^2}.$$

The ML estimator can therefore be found as the *minimizer*(!)

$$\hat{x}_{\text{ML}} = \arg \min_{x \in \mathbb{R}^d} \|y - F(x)\|^2.$$

Example

Assume that $y \in \mathbb{R}^k$ is an observation of the linear mathematical model

$$y = Ax + \varepsilon,$$

where $x \in \mathbb{R}^d$ is the unknown parameter, $A \in \mathbb{R}^{k \times d}$ is a matrix, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian measurement noise, with $\sigma > 0$ and I is the $k \times k$ identity matrix. This corresponds to $F(x) = Ax$ in the previous example, with the likelihood

$$\mathcal{L}_n(x) = \frac{1}{(2\pi\sigma^2)^{k/2}} e^{-\frac{1}{2\sigma^2} \|y - Ax\|^2}$$

and ML estimator

$$\hat{x}_{\text{ML}} = \arg \min_{x \in \mathbb{R}^d} \|y - Ax\|^2.$$

If $A^T A$ is invertible, then the ML estimator is precisely the least squares solution

$$A^T A \hat{x}_{\text{ML}} = A^T y.$$

(If $A^T A$ is not invertible, then the ML estimator is not unique.)

Computing ML estimates

In special cases the ML estimator \hat{x}_{ML} can be solved analytically, but more often, the optimization problem needs to be solved numerically. If the log-likelihood ℓ_n is twice continuously differentiable, one can use, e.g., the Newton–Raphson algorithm. Suppose that the parameter $x \in \mathbb{R}$ is one-dimensional. If x is a good guess for \hat{x}_{ML} (in the sense that $x \approx \hat{x}_{\text{ML}}$, then Taylor's theorem implies that

$$0 = \ell'_n(\hat{x}_{\text{ML}}) \approx \ell'_n(x) + (\hat{x}_{\text{ML}} - x)\ell''_n(x).$$

Solving for \hat{x}_{ML} yields $\hat{x}_{\text{ML}} = x - \frac{\ell'_n(x)}{\ell''_n(x)}$.

Repeating this process iteratively yields the following algorithm.

Let $x_0 \in \mathbb{R}$ be an initial guess for \hat{x}_{ML} .

for $j = 1, 2, \dots$, **do**

Set $x_j = x_{j-1} - \frac{\ell'_n(x_{j-1})}{\ell''_n(x_{j-1})}$

until $|\ell'_n(x_j)| < TOL$

In the multiparameter case $x \in \mathbb{R}^d$, the ML estimator \hat{x}_{ML} is a vector and the method is the following:

Let $x_0 \in \mathbb{R}^d$ be an initial guess for \hat{x}_{ML} .

for $j = 1, 2, \dots$, **do**

Set $x_j = x_{j-1} - H(x_{j-1})^{-1} \nabla \ell(x_{j-1})$

until $\|\nabla \ell_n(x_j)\| < TOL$

Here, $H(x)$ is the $d \times d$ Hessian matrix defined by

$$H(x) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} \ell_n(x) & \frac{\partial^2}{\partial x_1 \partial x_2} \ell_n(x) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_d} \ell_n(x) \\ \frac{\partial^2}{\partial x_2 \partial x_1} \ell_n(x) & \frac{\partial^2}{\partial x_2^2} \ell_n(x) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_d} \ell_n(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_d \partial x_1} \ell_n(x) & \frac{\partial^2}{\partial x_d \partial x_2} \ell_n(x) & \cdots & \frac{\partial^2}{\partial x_d^2} \ell_n(x) \end{bmatrix}.$$

Remark. Depending on the application, any other reasonable or natural optimization procedure might also work: e.g., Gauss–Newton method, Levenberg–Marquardt method, conjugate gradient or Krylov subspace methods, (stochastic) gradient descent...

Properties of the ML estimator

The ML estimator has many desirable qualities under somewhat relaxed assumptions:

- The ML estimator is *consistent*: $\hat{x}_{\text{ML}} \xrightarrow{P} x_*$ as $n \rightarrow \infty$, where x_* denotes the true value of the parameter x .
- The ML estimator is *asymptotically normal*: $\frac{\hat{x}_{\text{ML}} - x_*}{\widehat{\text{se}}} \xrightarrow{d} \mathcal{N}(0, 1)$.
- The ML estimator is *asymptotically optimal*: roughly, this means that among all well-behaved estimators, the ML estimator has the smallest variance, at least for large samples.

⋮

As the sample size $n \rightarrow \infty$, the ML estimator is an ideal estimator from a *frequentist* point of view.

However, in some applications one might have a limited amount of data and/or the data generation is not repeatable, so the asymptotic properties of the ML estimator may not be of much use. Next week, we will start discussing the *Bayesian paradigm*, where the fundamental conceit is that only a finite amount of data is available: probability is not defined as the limit of relative frequencies, but as a (subjective) degree of belief.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Twelfth lecture, January 13, 2025

Frequentist methods

The statistical methods that we have discussed so far are known as **frequentist (or classical)** methods. The frequentist point of view is based on the following postulates:

- F1 Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.
- F2 Parameters are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameters.
- F3 Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should contain the true value of the parameter with limiting frequency at least 95 percent.

Bayesian methods

There is another approach to inference called **Bayesian inference**. The Bayesian approach is based on the following postulates:

- B1 Probability describes degree of belief, not limiting frequency. As such, we can make probability statements about lots of things, not just data which are subject to random variation. For example, “the probability that Albert Einstein drank a cup of tea on August 1, 1948” is 0.35. This does not refer to any limiting frequency; it reflects a subjective strength of belief that the proposition is true.
- B2 The parameters are modeled as random variables, not as fixed, unknown constants. We can make probability statements about the parameters.
- B3 We can make inferences about a parameter x by producing a probability distribution for x . Inferences, such as point estimates and interval estimates, may then be extracted from this distribution.

Bayesian inference embraces a subjective notion of probability. In general, Bayesian methods provide no guarantees on long run performance.

Notation / recap on conditional and marginal PDFs

Let x and y be random variables with values in \mathbb{R}^d and \mathbb{R}^k , respectively. If the random variable (x, y) has a probability density $f_{x,y}$, i.e., if

$$\mathbb{P}(x \in A, y \in B) = \mathbb{P}((x, y) \in A \times B) = \int_{A \times B} f_{x,y}(u, v) \, du \, dv$$

for all events $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^k$, then $f_{x,y}$ is called the *joint probability density* of x and y . Here, $\mathbb{P}(x \in A, y \in B) := \mathbb{P}(x \in A \text{ and } y \in B)$. To simplify notation, we also write $f(x, y) = f_{x,y}(x, y)$.

Now, the *marginal probability density* f_x of x is defined by

$$f_x(u) = \int_{\mathbb{R}^k} f_{x,y}(u, v) \, dv \quad \text{for all } u \in \mathbb{R}^d.$$

Analogously, the marginal density of y is

$$f_y(v) = \int_{\mathbb{R}^d} f_{x,y}(u, v) \, du \quad \text{for all } v \in \mathbb{R}^k.$$

The marginal density of x is indeed the probability density for x in the situation where we have no information about the random variable y , because

$$\begin{aligned}\mathbb{P}(x \in A) &= \mathbb{P}(x \in A, y \in \mathbb{R}^k) = \int_{A \times \mathbb{R}^k} f_{x,y}(u, v) \, du \, dv \\ &= \int_A \left(\int_{\mathbb{R}^k} f_{x,y}(u, v) \, dv \right) \, du = \int_A f_x(u) \, du\end{aligned}$$

for every event $A \subset \mathbb{R}^d$.

The random variables x and y are independent (denoted by $x \perp y$) if

$$\mathbb{P}(x \in A, y \in B) = \mathbb{P}(x \in A)\mathbb{P}(y \in B)$$

for all events $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^k$ or, equivalently, if

$$f_{x,y}(u, v) = f_x(u)f_y(v) \quad \text{for all } u \in \mathbb{R}^d, v \in \mathbb{R}^k.$$

To simplify notation, we will also write $f(x) := f_x(x)$ and $f(y) := f_y(y)$.

Next, we consider the random variable x in the opposite situation where we know everything about the random variable y : we have observed it and know what value it has taken.

We say we consider the random variable x , *given* that we know the value y_0 taken by y , and denote this by $x|y = y_0$. For $y_0 \in \mathbb{R}^k$ with $f_y(y_0) > 0$, the *conditional probability density* of $x|y = y_0$, $f_{x|y=y_0}$, is then defined by

$$f_{x|y=y_0}(u) = \frac{f_{x,y}(u, y_0)}{f_y(y_0)}.$$

If x and y are independent and $f_y(y_0) > 0$, then

$$f_{x|y=y_0}(u) = f_x(u).$$

To simplify notation, we will also write $f(x|y) := f_{x|y}(x) := f_{x|y=y}(x)$.

Bayesian inference

Bayesian inference is usually carried out in the following way.

- ① We choose a probability density $f(x)$ – called the **prior distribution** – that expresses our beliefs about a parameter x before we see any data.
- ② We choose a statistical model $f(y|x)$ that reflects our beliefs about y given x .
- ③ After observing data y_1, \dots, y_n , we update our beliefs and calculate the **posterior distribution** $f(x|y_1, \dots, y_n)$.

In what follows, we will consider continuous \mathbb{R}^d -valued random variables.

Bayes' formula

Let (x, y) be a random variable with joint density $f(x, y)$ on $\mathbb{R}^d \times \mathbb{R}^k$. If $f(y) > 0$, then the conditional probability density of x , given y , equals

$$f(x|y) = \frac{f(x, y)}{\int_{\mathbb{R}^d} f(x, y) dx}.$$

On the other hand, the conditional probability density of y in case we know the value of the unknown x , is the **likelihood function**

$$f(y|x) = \frac{f(x, y)}{f(x)}, \quad \text{if } f(x) > 0.$$

Since $f(x, y) = f(y|x)f(x)$, this leads to **Bayes' formula**

$$f(x|y) = \frac{f(y|x)f(x)}{Z(y)}, \quad Z(y) := \int_{\mathbb{R}^d} f(y|x)f(x) dx.$$

If we have n i.i.d. observations y_1, \dots, y_n , then we replace $f(y|x)$ with

$$f(y_1, \dots, y_n|x) = \prod_{i=1}^n f(y_i|x) = \mathcal{L}_n(x).$$

Bayes' formula presents a way to express the conditional probability density of x , given y , assuming that the conditional density of y , given x , and the marginal density of x are known.

Example

Consider the problem of estimating an unknown parameter $x \in \mathbb{R}^d$ from data $y \in \mathbb{R}^k$ that is connected to x via the model

$$y = F(x) + \varepsilon. \quad (1)$$

If

- A1 The noise ε has the probability density ν on \mathbb{R}^k ;
- A2 The parameter x has the probability density f on \mathbb{R}^d ;
- A3 The random variables x and ε are independent;

then the likelihood is

$$f(y|x) = \nu(y - F(x)).$$

This is because

$$\begin{aligned} f(y|x) &= f_{y|x}(y) = f_{F(x)+\varepsilon|x}(y) = f_{\varepsilon|x}(y - F(x)) = f_\varepsilon(y - F(x)) \\ &= \nu(y - F(x)) \end{aligned}$$

due to the assumptions $\varepsilon \perp x$ and $\varepsilon \sim \nu$.

Example

If assumptions A1–A3 hold and

$$Z(y) = \int_{\mathbb{R}^d} \nu(y - F(x))f(x) dx > 0,$$

then the posterior density corresponding to (1) is

$$f(x|y) = \frac{\nu(y - F(x))f(x)}{Z(y)}.$$

Remarks.

- The condition that the marginal density $f(y)$ of the observed data y is positive means that the observed data is assumed to be consistent with the probabilistic assumptions A1–A3.
- An event cannot have positive probability under the posterior distribution if it does not have positive probability under the prior distribution.

Case study: source localization

Suppose that a particle with unit charge is located at some (unknown) point $x^* \in (0, 1)$ and our goal is to locate it based on measurements of voltage at the interval end points $x = 0$ and $x = 1$. The mathematical model for the voltage at any point $x \in [0, 1]$ is given by

$$y(x) = \frac{1}{|x^* - x|}.$$

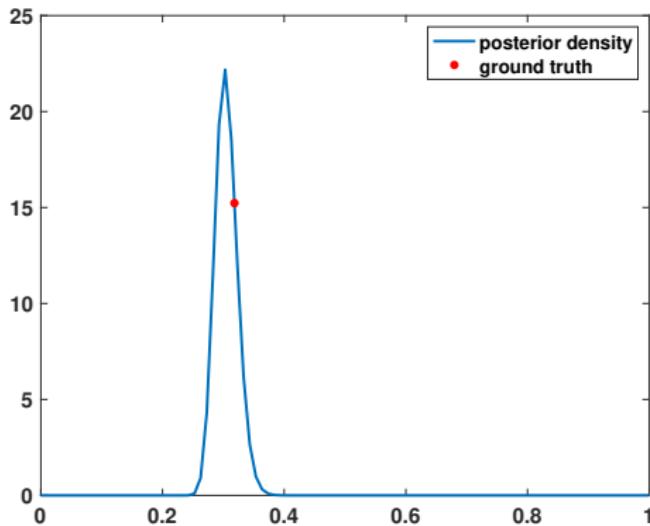
Our noisy measurements are modeled by $y_1 = \frac{1}{|x^*-0|} + \varepsilon_1$ and $y_2 = \frac{1}{|x^*-1|} + \varepsilon_2$, where ε_1 and ε_2 are i.i.d. realizations of $\mathcal{N}(0, \sigma^2)$. We take $x^* = 1/\pi$ (ground truth) and $\sigma = 0.2$ in the numerical experiments.

- The likelihood is given by $f(y|x) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{j=0}^1 \left(y_{j+1} - \frac{1}{|x-j|}\right)^2\right)$.
- We consider the prior $f(x) = \mathbf{1}_{(0,1)}(x) = \begin{cases} 1 & \text{if } x \in (0, 1), \\ 0 & \text{otherwise.} \end{cases}$

Then the posterior density is given by Bayes' formula

$$f(x|y) \propto \mathbf{1}_{(0,1)}(x) \exp\left(-\frac{1}{2\sigma^2} \sum_{j=0}^1 \left(y_{j+1} - \frac{1}{|x-j|}\right)^2\right).$$

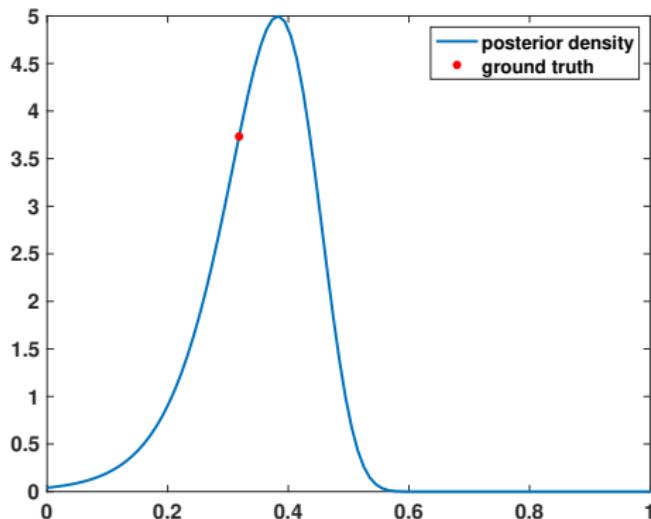
Let us visualize the posterior density against the ground truth solution.
(See also the file `source.py` on the course website!)



We see that the posterior is localized around the true parameter value (“ground truth”). **Note that in this case, the prior hardly plays any role.**

We could take, e.g., the mean or mode of the posterior density as a point estimate for the unknown location of the point charge.

What if we modify the problem so that we have access to only one boundary measurement at $x = 1$?



The resulting posterior distribution carries substantially more uncertainty since we now have less measurement data!

Note that the posterior will generally be high-dimensional, meaning that it is usually not possible to visually inspect the posterior density.

Let $x \in \mathbb{R}^d$, $y \in \mathbb{R}^k$ be random variables (the unknown parameter and the measurement, respectively). Bayes' formula:

$$f(x|y) = \frac{f(y|x)f(x)}{Z(y)}, \quad Z(y) := \int_{\mathbb{R}^d} f(y|x)f(x) dx > 0.$$

- The *prior model* $f(x)$ describes *a priori* information. It should assign high probability to objects x which are typical in light of *a priori* information, and low probability to unexpected x .
- The *likelihood model* $f(y|x)$ processes measurement information. It gives low probability to objects that produce simulated data which is very different from the measured data.
- The number $Z(y)$ can be treated as a normalization constant. It is often not of significant interest. If needed, we can recover it by computing the value of the integral $Z(y) = \int_{\mathbb{R}^d} f(y|x)f(x) dx$.
- The *posterior distribution* $f(x|y)$ represents the updated knowledge about the parameter of interest x , given the evidence y .

Since the normalization constant $Z(y)$ is often not of interest, we write

$$f(x|y) \propto f(y|x)f(x),$$

where \propto means equality up to a constant factor (not depending on x).

Bayesian estimators

The posterior distribution can be used to define estimators for the conditional random variable $x|y \sim f(x,y)$, where $y = (y_1, \dots, y_n)$. In general, an estimator \hat{x} is any function of the data y . The estimate $\hat{x} = \hat{x}(y)$ is itself an \mathbb{R}^d -valued random variable whose properties give information about the usefulness and quality of the estimator.

Bayesian estimators are those defined via the posterior distribution $f(x|y)$. We present the two most prominent ones. The **conditional mean (CM) estimator** is defined as the mean of the posterior distribution

$$\hat{x}_{\text{CM}} = \mathbb{E}[x|y] = \int_{\mathbb{R}^d} x f(x|y) dx$$

This is a high-dimensional integration problem.

The **maximum a posteriori (MAP) estimator** is defined as the mode

$$\hat{x}_{\text{MAP}} = \arg \max_{x \in \mathbb{R}^d} f(x|y)$$

of the posterior distribution (if a unique mode exists). *This is a high-dimensional optimization problem.*

One way to estimate spread are Bayesian **credible sets**. A level $1 - \alpha$ credible set \mathcal{C}_α with $\alpha \in (0, 1)$ satisfies

$$\mathbb{P}(x \in \mathcal{C}_\alpha | y) = \int_{\mathcal{C}_\alpha} f(x|y) dx = 1 - \alpha.$$

For small α , it is a region that contains a large fraction of the posterior mass.

Example. Assume that $x \in \mathbb{R}$ and that the posterior density is given by

$$f(x|y) = \frac{c}{\sigma_1} \phi\left(\frac{x}{\sigma_1}\right) + \frac{1-c}{\sigma_2} \phi\left(\frac{x-1}{\sigma_2}\right),$$

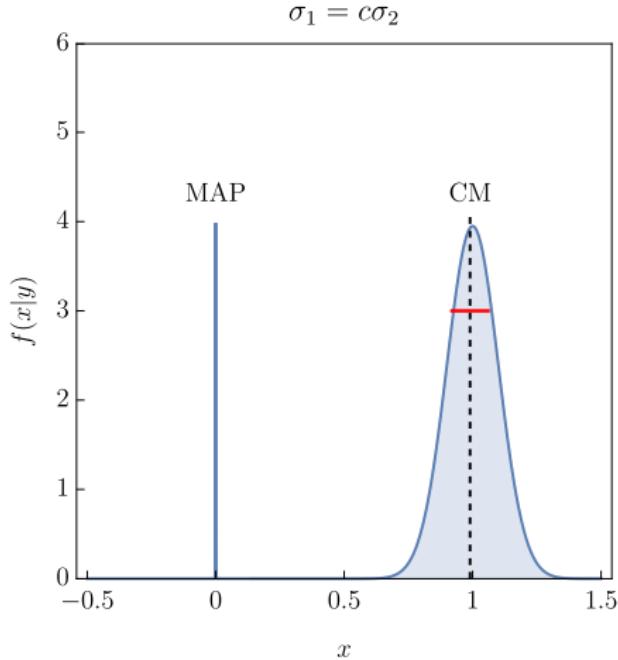
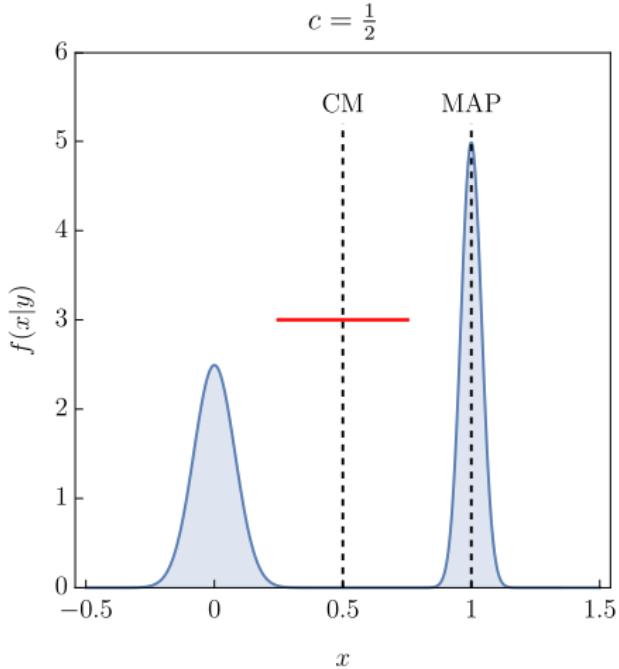
where $c \in (0, 1)$, $\sigma_1, \sigma_2 > 0$, and ϕ is the density of the standard normal distribution, $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. In this case,

$$\hat{x}_{CM} = 1 - c \quad \text{and} \quad \hat{x}_{MAP} = \begin{cases} 0 & \text{if } c/\sigma_1 > (1-c)/\sigma_2, \\ 1 & \text{if } c/\sigma_1 < (1-c)/\sigma_2. \end{cases}$$

If $c = \frac{1}{2}$ and σ_1, σ_2 are small, the probability that x takes values near \hat{x}_{CM} is small. On the other hand, if $\sigma_1 = c\sigma_2$, then $c/\sigma_1 = 1/\sigma_2 > (1-c)/\sigma_2$, so that $\hat{x}_{MAP} = 0$. If c is small, this is, however, a bad estimate for x , since the probability for x to take values near 0 is small. Last of all, we notice that when the conditional mean gives a poor estimate, this is reflected in a larger posterior variance

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \hat{x}_{CM})^2 f(x|y) dx.$$

We cannot say that one estimator is better than the other in all applications.



Left: the density with $\sigma_1 = 0.08$, $\sigma = 0.04$, and $c = \frac{1}{2}$. The CM estimate represents the distribution poorly. Notice that when the CM gives a poor estimate, this is reflected in wider variance (1 standard deviation is depicted as a red line). Right: the density with $\sigma_1 = 0.001$, $\sigma_2 = 0.1$, and $c = 0.01$. The MAP gives a poor estimate since it is in an unlikely part of the computational domain.

The maximum likelihood estimate

$$\hat{x}_{\text{ML}} = \arg \max_{x \in \mathbb{R}^d} f(y|x)$$

answers the question: “which value of the unknown is the most likely to produce the measured data?”

The ML estimate is a non-Bayesian estimate, and if the sample size is not large, it is not considered very useful by Bayesian statisticians.

Prior modeling

The prior density should reflect our beliefs on the unknown variable of interest before taking the measurements into account.

Often, the prior knowledge is qualitative in nature, and transferring the information into quantitative form expressed through a prior density can be challenging.

The prior probability distribution should be concentrated on those values of x we expect to see and assign a clearly higher probability to them than to the unexpected ones.

Gaussian priors

Gaussian densities

$$f(x) = \frac{1}{(2\pi)^{d/2} \sqrt{\det C}} \exp\left(-\frac{1}{2}(x - m)C^{-1}(x - m)\right)$$

are easy to construct and form a versatile class of distributions. They also often lead to explicit estimators.

Random samples from a standard normal distribution $\mathcal{N}(0, I)$ can be generated directly, for example via `numpy.random.normal` in Python. Samples from a general normal distribution $\mathcal{N}(m, C)$ and from a wide class of other distributions can then be derived from those, so that it is often not necessary to employ the inverse transform method.

Case study: signal recovery

Suppose that we want to estimate a one-dimensional signal $g: [0, 1] \rightarrow \mathbb{R}$ from indirect observations. We discretize the interval $[0, 1]$ by points $t_j = j/d$, $j \in \{1, \dots, d\}$, and write $x_j = g(t_j)$. In what follows, we consider some priors we could place for the unknown signal $x \in \mathbb{R}^d$.

Gaussian priors with covariance $C = \alpha^2 I$, $\alpha > 0$, are often called **(Gaussian) white noise priors**. The variance α^2 controls the magnitude of the realizations, but the values at t_1, \dots, t_d are independent.

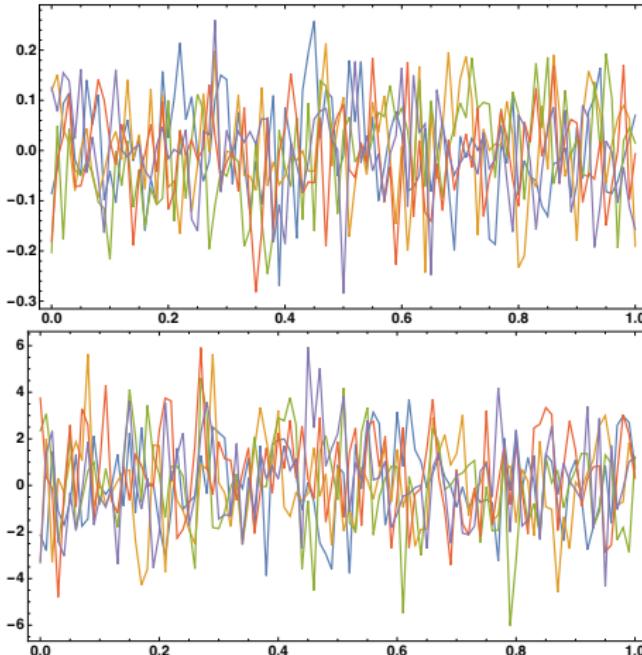


Figure: Top: 5 realizations of the Gaussian white noise prior with $\alpha = 0.1$.
Bottom: 5 realizations of the Gaussian white noise prior with $\alpha = 2$.

Gaussian priors with covariance $C = \alpha^2(L^T L)^{-1}$, where $\alpha > 0$ and $L = \text{tridiag}(-1, 2, -1)$ (we will discuss the construction of this prior next week), are often called **(Gaussian) smoothness priors**. The parameter α^2 controls the variability of the realizations. Note that this prior enforces the Dirichlet boundary condition $g(0) = g(1) = 0$.

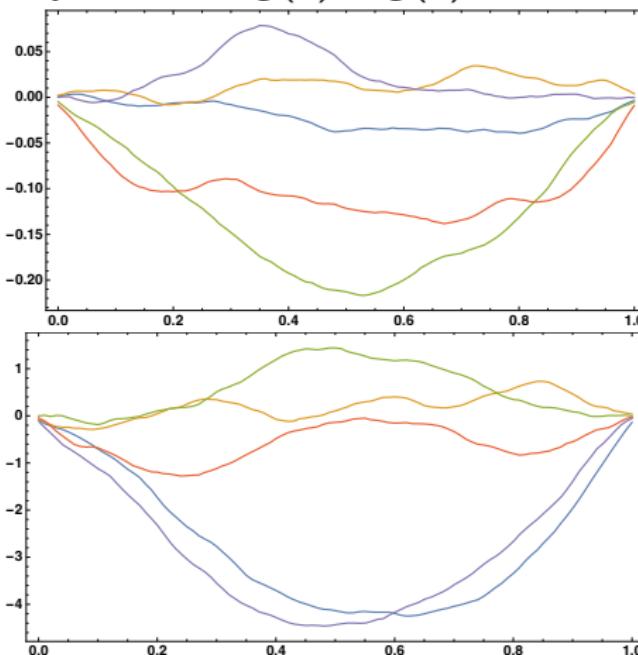


Figure: Top: 5 realizations of the Gaussian smoothness prior with $\alpha = 0.001$.
 Bottom: 5 realizations of the Gaussian smoothness prior with $\alpha = 0.02$.

Impulse priors

Assume that our prior information is that the signal contains small and well localized features in an almost constant background.

In such a case we could assume an impulse prior density, which means that it gives a low average amplitude but allows outliers. The tail of such a prior distribution is long, although the expected value is small.

Let $x \in \mathbb{R}^d$ represent the signal, where the component $x_j = f(t_j)$ is the values at the j^{th} coordinate. In what follows, x_i and x_j are assumed to be independent for $i \neq j$.

One example of an impulse prior is the ℓ^1 prior. It has the density

$$f(x) = \left(\frac{\alpha}{2}\right)^d \exp(-\alpha\|x\|_1)$$

with $\alpha > 0$, where the ℓ^1 -norm is defined as

$$\|x\|_1 = \sum_{j=1}^d |x_j|.$$

The impulse effect can be enhanced by choosing an even smaller power $p \in (0, 1)$ of the components of x , that is, using $\sum_{j=1}^d |x_j|^p$ instead of the ℓ^1 -norm.

Another choice that produces images with few distinctly different function values and a low-amplitude background is the **Cauchy density**

$$f(x) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^d \frac{1}{1 + \alpha^2 x_j^2}$$

with $\alpha > 0$.

Since we assumed that each component is independent of the others, random draws can be performed componentwise using, e.g., inverse transform sampling.

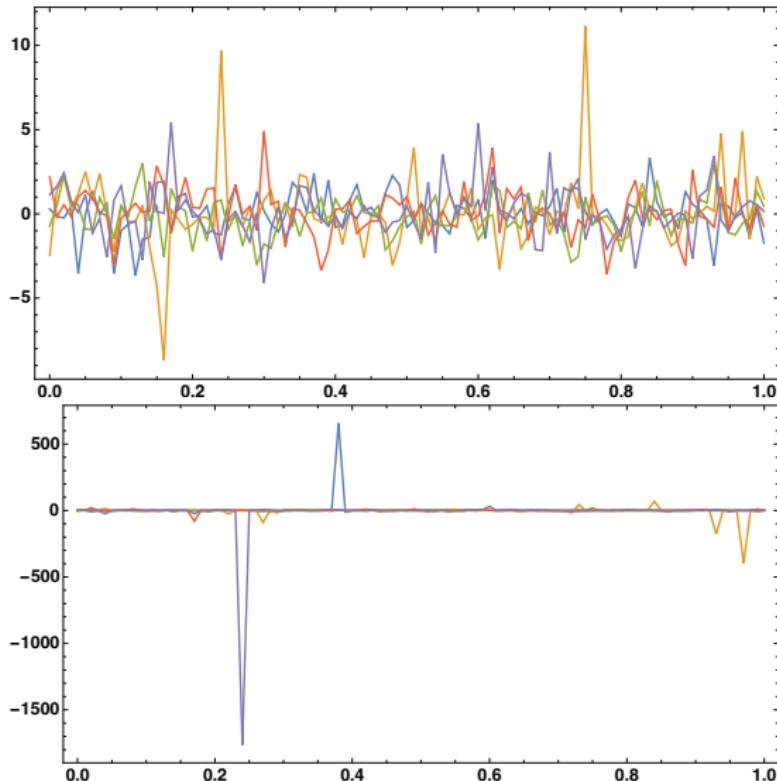


Figure: Top: 5 realizations of the ℓ^1 prior with $\alpha = 1$. Bottom: 5 realizations of the Cauchy prior with $\alpha = 1$.

Discontinuities

Assume still that we want to estimate a one-dimensional signal $g: [0, 1] \rightarrow \mathbb{R}$ with $g(0) = 0$ from indirect observations. Our prior knowledge is that the signal is usually relatively stable but can have large jumps every now and then. We may also have information on the size of the jumps or the rate of their occurrence.

We obtain one possible prior by taking the finite difference approximation of the derivative of g and assigning an impulsive noise distribution to it. Let us discretize the interval $[0, 1]$ by points $t_j = j/d$ and write $x_j = g(t_j)$. Consider the density

$$f(x) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^d \frac{1}{1 + \alpha^2(x_j - x_{j-1})^2}. \quad (2)$$

To draw samples from the above distribution we define new random variables for the jumps

$$u_j = x_j - x_{j-1}, \quad j = 1, \dots, d.$$

These each have the density

$$f(u) = \left(\frac{\alpha}{\pi}\right)^d \prod_{j=1}^d \frac{1}{1 + \alpha^2 u_j^2}.$$

In particular, the u_j are independent from each other, so that they can be drawn from a one-dimensional Cauchy density. Also note that $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$ satisfies $x = Lu$, where $L \in \mathbb{R}^{d \times d}$ is a lower triangular matrix with $L_{ij} = 1$ for $i \geq j$.[†] Generalizing the idea behind the above prior leads, e.g., to total variation priors.

[†]Note that in Python, it is more efficient to implement this as `x = numpy.cumsum(u)`.

Example: drawing realizations from the prior (2)

```
import numpy as np
import matplotlib.pyplot as plt

d = 1200
t = np.arange(1,d+1)/d
alpha = 1
quantile = lambda t: 1/alpha * np.tan(np.pi * (t-1/2))
unif = np.random.uniform(size=d)
draw = quantile(unif)
y = np.cumsum(draw)
plt.plot(t,y)
plt.xlabel('$t$', fontsize=14)
plt.ylabel('$g(t)$', fontsize=14)
plt.show()
```

Example: drawing realizations from the prior (2)

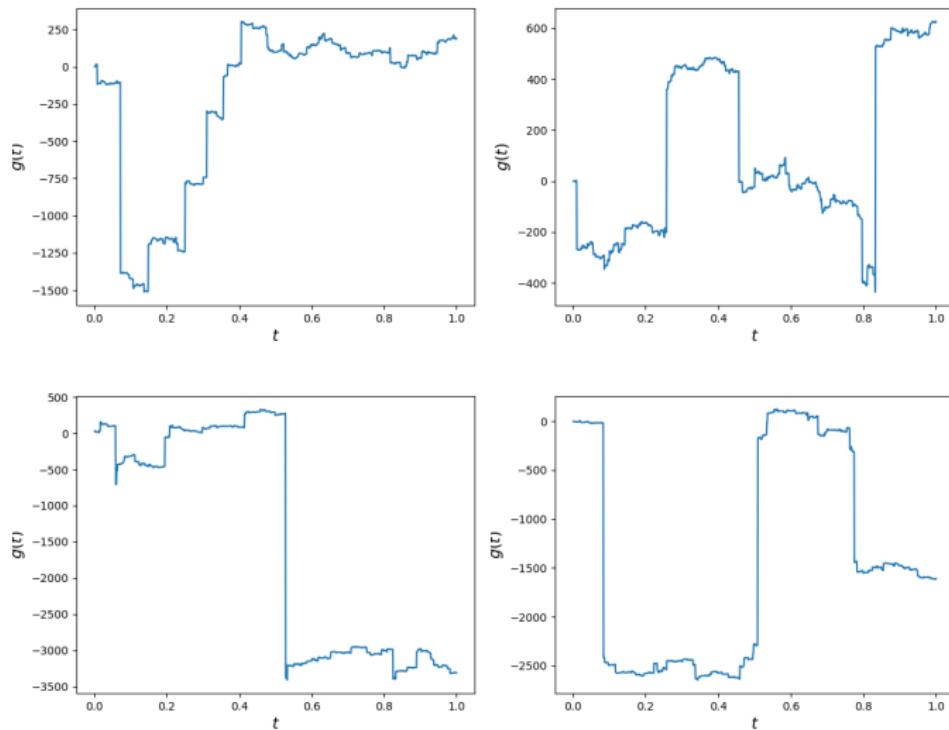


Figure: Four realizations drawn from the prior (2)

Hierarchical models

The prior density may depend on some parameter, such as variance or mean. So far we have assumed that these parameters are known. However, we often do not know how to choose them. If a parameter is not known, it can be estimated as a part of the statistical inference problem on the data. This leads to hierarchical models that include hypermodels for the parameters defining the prior density.

Assume that the prior distribution depends on a parameter α , which is assumed to be unknown. We then write the prior as a conditional density

$$f(x|\alpha).$$

We model the unknown α with a **hyperprior** $f_h(\alpha)$ and write the joint distribution of x and α as

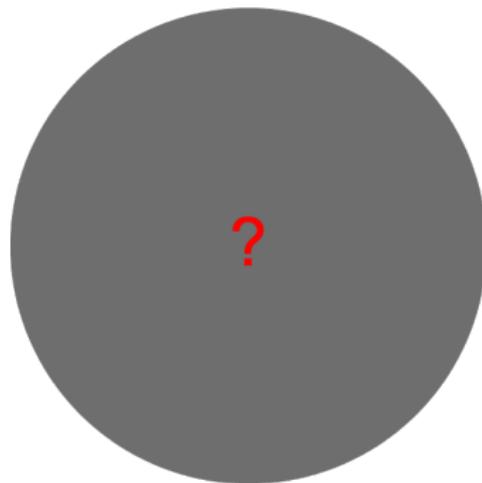
$$f(x, \alpha) = f(x|\alpha)f_h(\alpha).$$

Assuming we have a likelihood model $f(y|x)$ for the measurement y , we get the posterior density for x and α , given y , using Bayes' formula

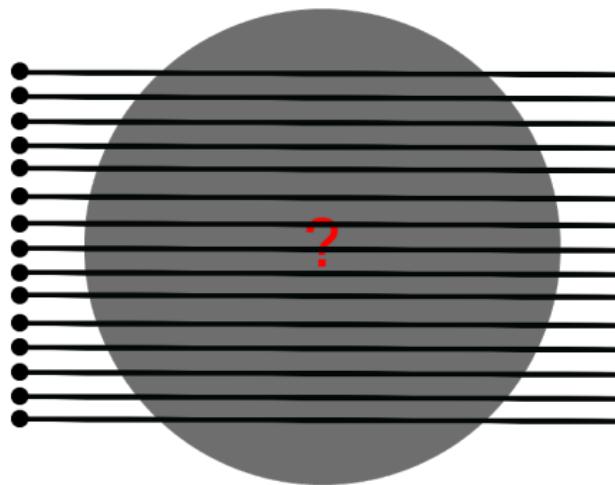
$$f(x, \alpha|y) \propto f(y|x, \alpha)f(x, \alpha) = f(y|x, \alpha)f(x|\alpha)f_h(\alpha).$$

The hyperprior density f_h may again depend on some hyperparameter α_0 . The main reason for the use of a hyperprior model is that the construction of the posterior is considered to be more robust with respect to fixing a value for the hyperparameter α_0 than fixing a value for α .

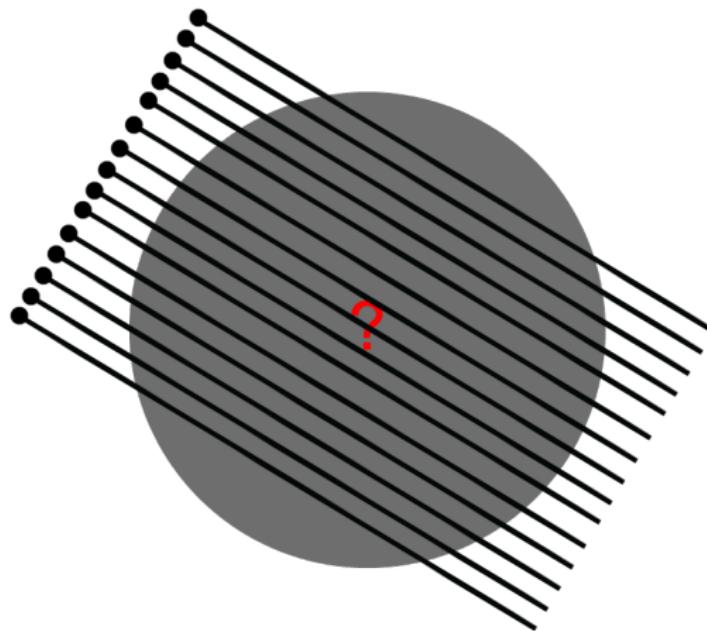
Case study: parallel-beam X-ray tomography



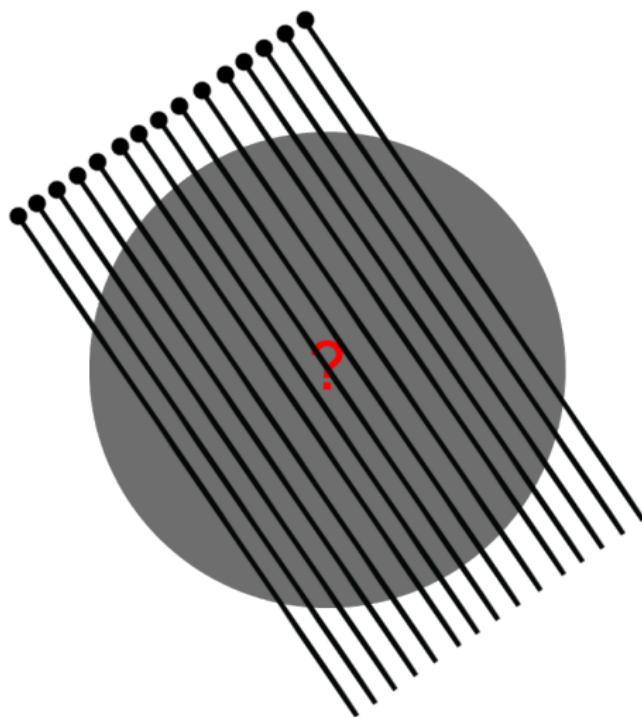
Case study: parallel-beam X-ray tomography



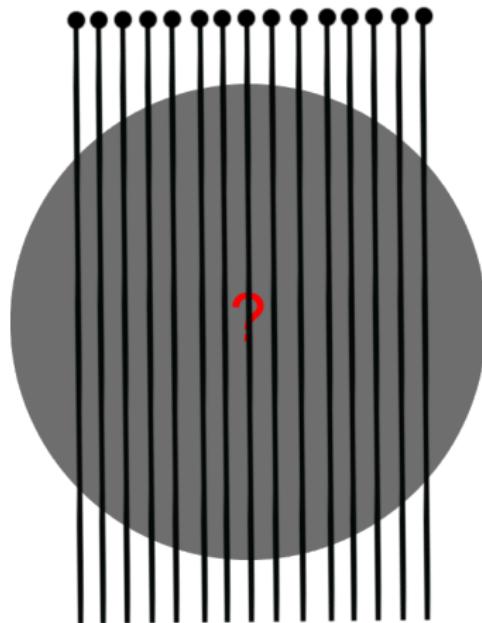
Case study: parallel-beam X-ray tomography



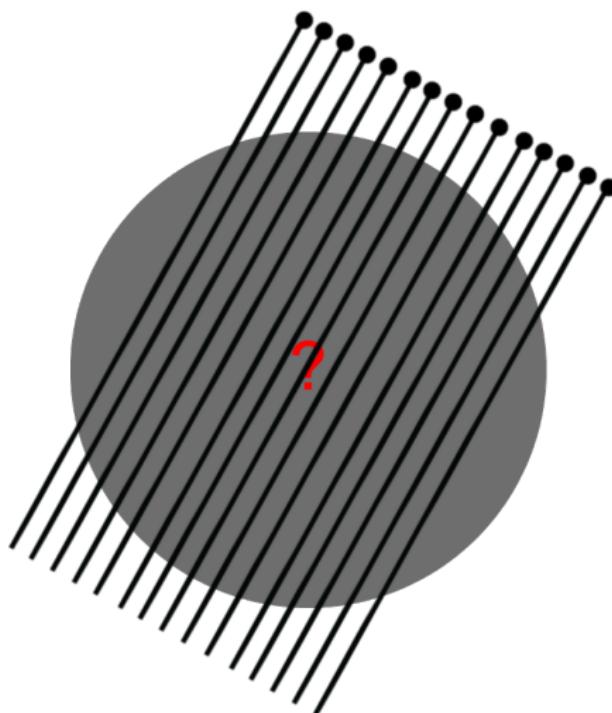
Case study: parallel-beam X-ray tomography



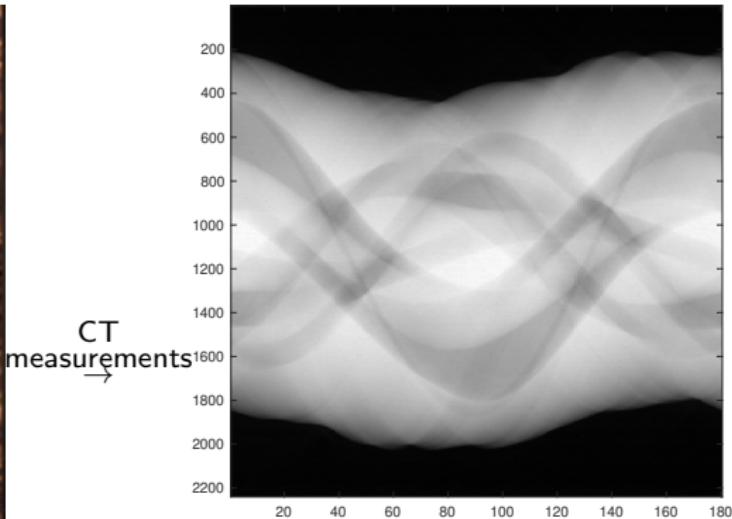
Case study: parallel-beam X-ray tomography



Case study: parallel-beam X-ray tomography



An object (left) is illuminated using a beam front of X-rays and the intensities of the X-rays are recorded after passing through the object. The measurements can be represented as a sinogram (right). In this case, the beam front consists of 2240 parallel X-rays (arranged as rows) taken at 180 equally spaced angles at 1° increments (arranged as columns). The goal is to reconstruct the interior density of the object based on the sinogram measurements.



Formation of a CT sinogram (by Samuli Siltanen):

https://www.youtube.com/watch?v=q7Rt_0Y_7tU

Let us consider the inference problem of recovering the attenuation coefficient (density) of an object given a set of X-ray measurements. The mathematical model can be expressed as

$$y = Ax,$$

where $y \in \mathbb{R}^Q$ denotes the (noisy) measurements for Q X-rays, $A \in \mathbb{R}^{Q \times n^2}$ is the projection matrix subject to an $n \times n$ pixel discretization of the computational domain, and $x \in \mathbb{R}^{n^2}$ denotes the (piecewise constant) discretization of the unknown attenuation inside the object of interest.

The data y can be reshaped into an $n \times n$ array, which is a graphical representation of the X-ray measurements (sinogram). The unknown can be reshaped into an $n \times n$ image of the density of the imaged object.

We use the FIPS open dataset of carved cheese available at
<https://doi.org/10.5281/zenodo.1254210>

The files `DataFull_128x15.mat` and `DataLimited_128x15.mat` contain sparse angle and limited angle tomography measurements, respectively. The data has been collected using 15 projections spanning either the full 360° circle in the first dataset, and 15 projections spanning a limited 90° angle of view in the second dataset. The computational domain is a 128×128 pixel grid in both cases. Each file contains a projection matrix A and a sinogram measurement matrix m .

By defining $y = m.reshape(m.size, 1)$, our naïve maximum likelihood (ML) reconstruction of the unknown x is precisely the least squares solution of the problem

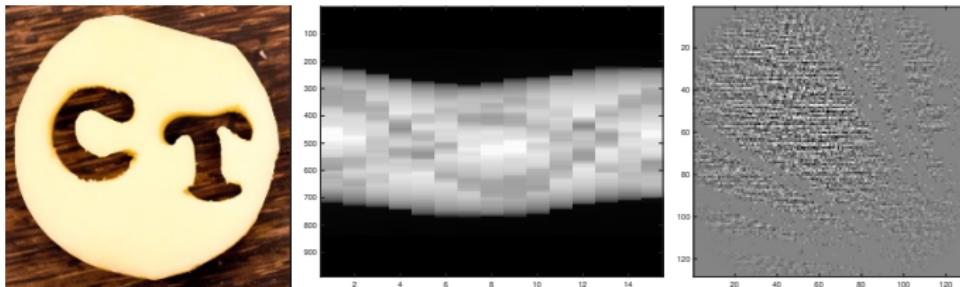
$$y = Ax.$$

The reconstruction is the image $x.reshape(128, 128)$.

In addition, we also consider the MAP estimators of the unknown x corresponding to a Gaussian white noise prior and a total variation prior (which gives a high probability to piecewise constant signals).

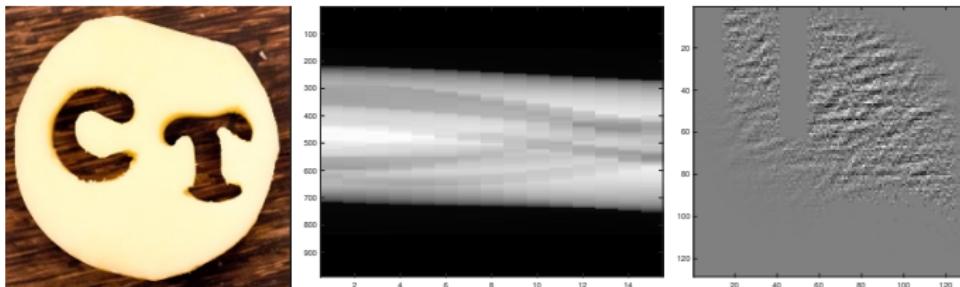
Maximum likelihood (ML) estimator

Sparse angle tomography data:



Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: ML estimator of the unknown density.

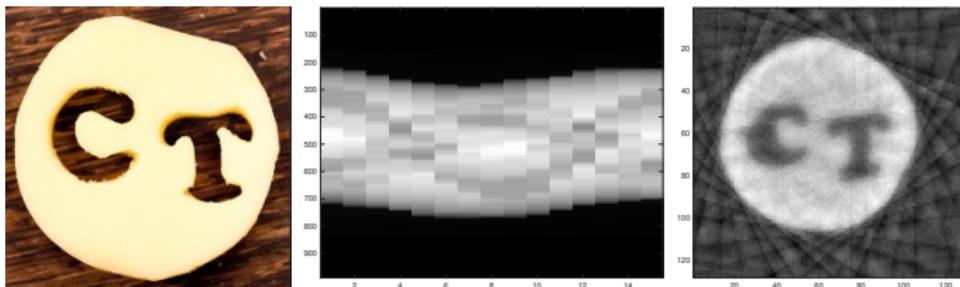
Limited angle tomography data:



Left: the actual object. Middle: sinogram data for limited angle tomography. Right: ML estimator of the unknown density.

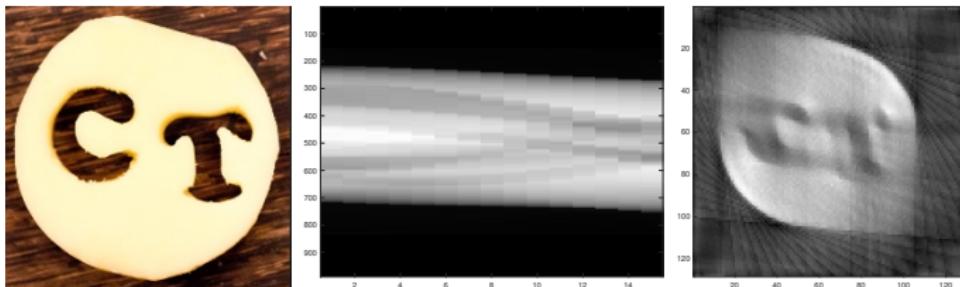
MAP estimator (Gaussian prior)

Sparse angle tomography data:



Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: MAP estimator with a Gaussian prior.

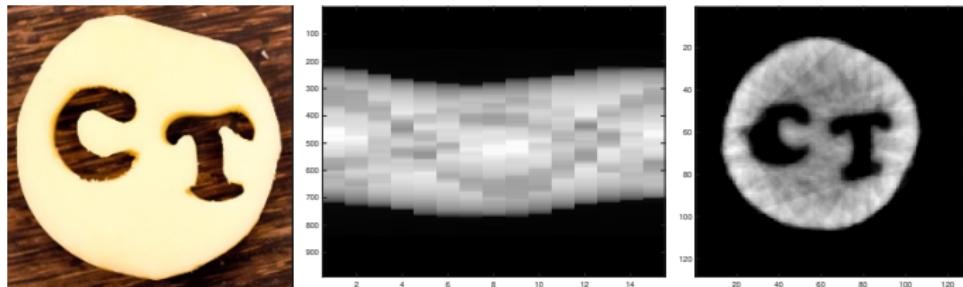
Limited angle tomography data:



Left: the actual object. Middle: sinogram data for limited angle tomography. Right: MAP estimator with a Gaussian prior.

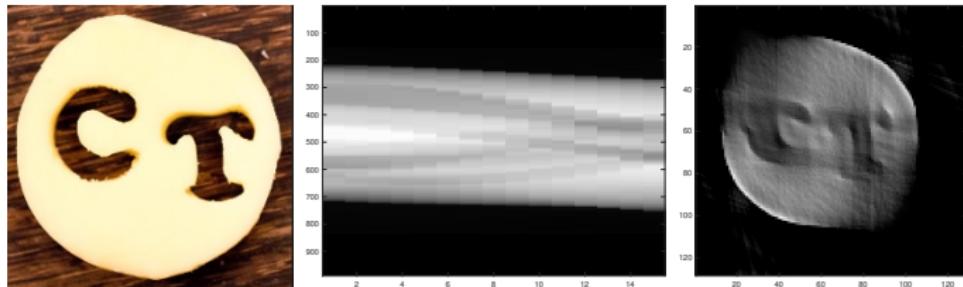
MAP estimator (total variation prior)

Sparse angle tomography data:



Left: the actual object. Middle: sinogram data for sparse angle tomography. Right: MAP estimator with a total variation prior.

Limited angle tomography data:



Left: the actual object. Middle: sinogram data for limited angle tomography. Right: MAP estimator with a total variation prior.

- In the previous example, using sparse or limited angle measurements means that the matrix system $y = Ax$ is underdetermined. Since we have very little data in the sparse or limited angle measurement settings, the ML estimator is useless in practice.
- In the Bayesian approach even a fairly weak Gaussian prior can produce a reconstruction. Using a more sophisticated prior such as a total variation prior improves the reconstruction quality even further (in this case, the density of the object can be well approximated using piecewise constant functions).
- The prior essentially compensates for the lack of data in the Bayesian approach to parameter recovery problems

Solution strategies

Bayes' formula produces an expression for the (in general high-dimensional) posterior distribution of the unknown parameter $x \in \mathbb{R}^d$, given the available data $y \in \mathbb{R}^k$. The main Bayesian estimators of the unknown parameter x are the MAP estimate \hat{x}_{MAP} (high-dimensional optimization problem) and the CM estimate \hat{x}_{CM} (high-dimensional integration problem). One may also be interested in quantifying the uncertainty in these estimates by computing the (co)variance of the posterior distribution or Bayesian credible sets (high-dimensional numerical integration problems). Typical solution strategies include the following.

- **Conjugate inference:** for a given likelihood, the prior is chosen such that the posterior is in the same probability distribution family as the prior (for example, if the likelihood and prior are both Gaussian, then the posterior is also Gaussian with known mean and covariance). In these cases, the MAP, CM, and (co)variance of the posterior have closed form solutions. This is an algebraic convenience, which avoids numerical difficulties otherwise associated with the computation of the MAP, CM, or other statistics of the posterior distribution.

- Numerical methods:

- The computation of the MAP estimate is a high-dimensional optimization problem. It is often convenient to work with the *negative log-posterior*

$$\hat{x}_{\text{MAP}} = \arg \min_{x \in \mathbb{R}^d} (-\log f(x|y)).$$

In some cases, the MAP estimator can be expressed as the solution to a Tikhonov functional. For example, consider the problem

$$y = F(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where $x \in \mathbb{R}^d$ is the unknown parameter, $y \in \mathbb{R}^k$ is the data, and $\sigma > 0$ is the noise level. If we endow x with a Gaussian prior, e.g., $x \sim \mathcal{N}(x_0, \gamma^2 I)$, $\gamma > 0$, then the MAP estimator can be found as the minimizer of the Tikhonov functional

$$\hat{x}_{\text{MAP}} = \arg \min_{x \in \mathbb{R}^d} (\|y - F(x)\|^2 + \lambda^2 \|x - x_0\|^2),$$

where $\lambda = \frac{\sigma}{\gamma}$. If $F: \mathbb{R}^d \rightarrow \mathbb{R}^k$ is linear, i.e., $F(x) = Ax$ for some matrix $A \in \mathbb{R}^{k \times d}$, then we can solve \hat{x}_{MAP} from the (invertible) linear system

$$(A^T A + \lambda^2 I) \hat{x}_{\text{MAP}} = A^T y + \lambda^2 x_0. \quad (\text{exercise})$$

- Numerical methods:

- The computation of the CM estimate is a high-dimensional numerical integration problem:

$$\hat{x}_{\text{CM}} = \int_{\mathbb{R}^d} x f(x|y) dx. \quad (3)$$

Typical solution strategies involve using high-dimensional cubatures or sampling-based methods. We will discuss the latter. Namely, if we are able to draw an i.i.d. sample x_1, \dots, x_n from the posterior $f(x|y)$, then we can in principle use the Monte Carlo method to approximate (3) as

$$\hat{x}_{\text{CM}} \approx \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n$$

and likewise for the posterior variance $\text{Var}(x|y) \approx \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$.

The difficulty with this approach lies in drawing a sample from a high-dimensional posterior distribution. To this end, we will discuss Markov Chain Monte Carlo (MCMC), which is an algorithm that can be used to draw a sample from a high-dimensional distribution with a known (unnormalized) density function.

Another approach is to use, e.g., importance sampling to obtain a (biased) estimate of the integral (3).

Appendix: Remark on Bayesian hypothesis testing

Remark on Bayesian hypothesis testing

It is possible to perform statistical hypothesis testing from a Bayesian point of view. We will only give a brief sketch of the main idea here.

The Bayesian approach to testing involves putting a prior on H_0 and on the parameter x and then computing $\mathbb{P}(H_0|y)$. Consider the case where x is a vector and we are testing

$$H_0: x = x_0 \quad \text{versus} \quad H_1: x \neq x_0.$$

It is usually reasonable to use the prior $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$ (although this is not essential in what follows). Under H_1 , we need a prior for x ; let us denote this prior density by $f(x)$. From Bayes' theorem,

$$\begin{aligned}\mathbb{P}(H_0|y) &= \frac{f(y|H_0)\mathbb{P}(H_0)}{f(y|H_0)\mathbb{P}(H_0) + f(y|H_1)\mathbb{P}(H_1)} = \frac{\frac{1}{2}f(y|x_0)}{\frac{1}{2}f(y|x_0) + \frac{1}{2}f(y|H_1)} \\ &= \frac{f(y|x_0)}{f(y|x_0) + \int_{\mathbb{R}^d} f(y|x)f(x) dx} = \frac{\mathcal{L}(x_0)}{\mathcal{L}(x_0) + \int_{\mathbb{R}^d} \mathcal{L}(x)f(x) dx}.\end{aligned}$$

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Thirteenth lecture, January 20, 2025

The linear Gaussian setting

In these notes we study the linear Gaussian setting, where the forward map F is linear and both the prior distribution and the distribution of the observational noise ε are Gaussian.

It arises frequently in applications, either directly or in the form of posterior distributions that are asymptotically Gaussian in the large data limit. It also allows computing explicit solutions which can be used to gain a general understanding. Apart from that, many methods employed in a nonlinear or non-Gaussian setting build on ideas from the linear Gaussian case by performing linearization or Gaussian approximation.

Let us suppose that the unknown $x \in \mathbb{R}^d$ and the data $y \in \mathbb{R}^k$ follow the relation

$$y = Ax + \varepsilon, \quad (1)$$

where

1. The forward model is linear, i.e., $A \in \mathbb{R}^{k \times d}$.
2. The prior distribution is Gaussian: $x \sim \mathcal{N}(x_0, \Gamma_{\text{pr}})$, where $x_0 \in \mathbb{R}^d$ and $\Gamma_{\text{pr}} \in \mathbb{R}^{d \times d}$ is symmetric and positive definite.
3. The noise is Gaussian: $\varepsilon \sim \mathcal{N}(\varepsilon_0, \Gamma_n)$, where $\varepsilon_0 \in \mathbb{R}^k$ and $\Gamma_n \in \mathbb{R}^{k \times k}$ is symmetric and positive definite.
4. x and ε are independent.

Theorem

Under assumptions 1–4, the posterior distribution corresponding to (1) is Gaussian with $x|y \sim \mathcal{N}(\mu_{\text{post}}, \Gamma_{\text{post}})$, where we have the posterior mean

$$\mu_{\text{post}} = (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A)^{-1} (A^T \Gamma_n^{-1} (y - \varepsilon_0) + \Gamma_{\text{pr}}^{-1} x_0)$$

and covariance

$$\Gamma_{\text{post}} = (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A)^{-1}.$$

Proof. Noting that $\Gamma_{\text{post}} = (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A)^{-1}$ and $\mu_{\text{post}} = \Gamma_{\text{post}}(A^T \Gamma_n^{-1}(y - \varepsilon_0) + \Gamma_{\text{pr}}^{-1} x_0)$, we obtain

$$\begin{aligned}
 f(x|y) &\propto \exp \left(-\frac{1}{2}(y - Ax - \varepsilon_0)^T \Gamma_n^{-1} (y - Ax - \varepsilon_0) \right) \exp \left(-\frac{1}{2}(x - x_0)^T \Gamma_{\text{pr}}^{-1} (x - x_0) \right) \\
 &= \exp \left(-\frac{1}{2} (y^T \Gamma_n^{-1} y - y^T \Gamma_n^{-1} A x - y^T \Gamma_n^{-1} \varepsilon_0 \right. \\
 &\quad \left. - x^T A^T \Gamma_n^{-1} y + x^T A^T \Gamma_n^{-1} A x + x^T A^T \Gamma_n^{-1} \varepsilon_0 \right. \\
 &\quad \left. - \varepsilon_0^T \Gamma_n^{-1} y + \varepsilon_0^T \Gamma_n^{-1} A x + \varepsilon_0^T \Gamma_n^{-1} \varepsilon_0 \right. \\
 &\quad \left. + x^T \Gamma_{\text{pr}}^{-1} x - 2x^T \Gamma_{\text{pr}}^{-1} x_0 + x_0^T \Gamma_{\text{pr}}^{-1} x_0 \right)
 \end{aligned}$$

Proof. Noting that $\Gamma_{\text{post}} = (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A)^{-1}$ and $\mu_{\text{post}} = \Gamma_{\text{post}}(A^T \Gamma_n^{-1}(y - \varepsilon_0) + \Gamma_{\text{pr}}^{-1} x_0)$, we obtain

$$\begin{aligned}
 f(x|y) &\propto \exp \left(-\frac{1}{2}(y - Ax - \varepsilon_0)^T \Gamma_n^{-1}(y - Ax - \varepsilon_0) \right) \exp \left(-\frac{1}{2}(x - x_0)^T \Gamma_{\text{pr}}^{-1}(x - x_0) \right) \\
 &\propto \exp \left(-\frac{1}{2} \left(\begin{array}{c} -x^T A^T \Gamma_n^{-1} y \\ -x^T A^T \Gamma_n^{-1} A x + x^T A^T \Gamma_n^{-1} \varepsilon_0 \\ + x^T A^T \Gamma_n^{-1} \varepsilon_0 \\ + x^T \Gamma_{\text{pr}}^{-1} x - 2x^T \Gamma_{\text{pr}}^{-1} x_0 \end{array} \right) \right)
 \end{aligned}$$

Proof. Noting that $\Gamma_{\text{post}} = (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A)^{-1}$ and $\mu_{\text{post}} = \Gamma_{\text{post}}(A^T \Gamma_n^{-1}(y - \varepsilon_0) + \Gamma_{\text{pr}}^{-1} x_0)$, we obtain

$$\begin{aligned}
 f(x|y) &\propto \exp \left(-\frac{1}{2}(y - Ax - \varepsilon_0)^T \Gamma_n^{-1}(y - Ax - \varepsilon_0) \right) \exp \left(-\frac{1}{2}(x - x_0)^T \Gamma_{\text{pr}}^{-1}(x - x_0) \right) \\
 &\propto \exp \left(-\frac{1}{2} \left(\begin{array}{c} -x^T A^T \Gamma_n^{-1} y \\ -x^T A^T \Gamma_n^{-1} A x + x^T A^T \Gamma_n^{-1} \varepsilon_0 \\ + x^T A^T \Gamma_n^{-1} \varepsilon_0 \\ + x^T \Gamma_{\text{pr}}^{-1} x - 2x^T \Gamma_{\text{pr}}^{-1} x_0 \end{array} \right) \right) \\
 &= \exp \left(-\frac{1}{2} \left(\underbrace{x^T (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A) x}_{=\Gamma_{\text{post}}^{-1}} - 2x^T \underbrace{(A^T \Gamma_n^{-1}(y - \varepsilon_0) + \Gamma_{\text{pr}}^{-1} x_0)}_{=\Gamma_{\text{post}}^{-1} \mu_{\text{post}}} \right) \right).
 \end{aligned}$$

On the previous slide, we arrived at

$$f(x|y) \propto \exp\left(-\frac{1}{2}(x^T \Gamma_{\text{post}}^{-1} x - 2x^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}})\right).$$

To finish the proof, we “complete the square” by multiplying and dividing by $\exp(-\frac{1}{2}\mu_{\text{post}}^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}})$. Since this term does not depend on x , we can absorb the denominator into the implied coefficient to obtain

$$\begin{aligned} f(x|y) &\propto \exp\left(-\frac{1}{2}(x^T \Gamma_{\text{post}}^{-1} x - 2x^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}})\right) \exp\left(-\frac{1}{2}\mu_{\text{post}}^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}}\right) \\ &= \exp\left(-\frac{1}{2}(x^T \Gamma_{\text{post}}^{-1} x - 2x^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}} + \mu_{\text{post}}^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}})\right) \\ &= \exp\left(-\frac{1}{2}((x - \mu_{\text{post}})^T \Gamma_{\text{post}}^{-1} (x - \mu_{\text{post}}) + 2x^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}} - 2x^T \Gamma_{\text{post}}^{-1} \mu_{\text{post}})\right) \\ &= \exp\left(-\frac{1}{2}((x - \mu_{\text{post}})^T \Gamma_{\text{post}}^{-1} (x - \mu_{\text{post}}))\right), \end{aligned}$$

as desired. □



Remark: The previous proof shows that if $x \sim \mathcal{N}(x_0, \Gamma_{\text{pr}})$ and $\varepsilon \sim \mathcal{N}(\varepsilon_0, \Gamma_n)$, then

$$x|y \sim \mathcal{N}(\mu_{\text{post}}, \Gamma_{\text{post}}),$$

where

$$\Gamma_{\text{post}} = (\Gamma_{\text{pr}}^{-1} + A^T \Gamma_n^{-1} A)^{-1} \quad (2)$$

$$\mu_{\text{post}} = \Gamma_{\text{post}}(A^T \Gamma_n^{-1} (y - \varepsilon_0) + \Gamma_{\text{pr}}^{-1} x_0). \quad (3)$$

One also has the following alternative representations for the posterior mean

$$\mu_{\text{post}} = x_0 + \Gamma_{\text{pr}} A^T (A \Gamma_{\text{pr}} A^T + \Gamma_n)^{-1} (y - Ax_0 - \varepsilon_0) \quad (4)$$

and the posterior covariance

$$\Gamma_{\text{post}} = \Gamma_{\text{pr}} - \Gamma_{\text{pr}} A^T (A \Gamma_{\text{pr}} A^T + \Gamma_n)^{-1} A \Gamma_{\text{pr}}. \quad (5)$$

Formula (5) can be proved, e.g., by using the **Sherman–Morrison–Woodbury formula** on (2). Formula (4) can be proved by plugging the formula (5) into (3) and simplifying the expression.

As the posterior distribution is Gaussian, its mean and its mode coincide. This means that the conditional mean estimator and the MAP estimator coincide in the linear Gaussian setting.

Corollary

The conditional mean estimator and the maximum a posteriori estimator coincide in the linear Gaussian setting and are given by

$$x_{CM} = x_{MAP} = \mu_{post}.$$

Example

Let $\Gamma_n = \sigma^2 I$, $\varepsilon_0 = 0$, $\Gamma_{pr} = \gamma^2 I$, $x_0 = 0$, and set $\lambda = \frac{\sigma}{\gamma}$. Then μ_{post} minimizes

$$J_\lambda(x) := \|y - Ax\|^2 + \lambda^2 \|x\|^2.$$

and therefore satisfies

$$(A^T A + \lambda^2 I) \mu_{post} = A^T y. \quad (6)$$

This example provides a connection between Bayesian inference and variational regularization: J_λ can be interpreted as the objective functional in a linear regression model with a regularization term $\lambda^2 \|x\|^2$. Equation (6) for μ_{post} is then exactly the normal equation.

In the general case, the formula

$$\mu_{post} = (\Gamma_{pr}^{-1} + A^T \Gamma_n^{-1} A)^{-1} (A^T \Gamma_n^{-1} (y - \varepsilon_0) + \Gamma_{pr}^{-1} x_0)$$

can thus be viewed as the solution to a generalized normal equation. This point of view helps to understand the structure of Bayesian inference by linking it to well-understood optimization approaches.

Numerical example: one-dimensional deconvolution

Suppose that we are interested in estimating a signal $g: [0, 1] \rightarrow \mathbb{R}$ from noisy, blurred observations modeled as

$$y_i = y(s_i) = \int_0^1 K(s_i, t)g(t) dt + \varepsilon_i, \quad i \in \{1, \dots, k\},$$

where the blurring kernel is

$$K(s, t) = \exp\left(-\frac{1}{2\omega^2}(s - t)^2\right), \quad \omega = 0.5,$$

and we have Gaussian measurement noise $\varepsilon \sim \mathcal{N}(0, \Gamma_{\text{noise}})$ with a symmetric, positive definite covariance matrix Γ_{noise} .

Discrete model

Midpoint rule:

$$y_i = \int_0^1 K(s_i, t)g(t) dt + \varepsilon_i \approx \frac{1}{d} \sum_{j=1}^d K(s_i, t_j)x_j + \varepsilon_i,$$

where $t_j = \frac{j}{d} - \frac{1}{2d}$ and $x_j = g(t_j)$ for $j \in \{1, \dots, d\}$.

If we have $s_i = \frac{i}{k} - \frac{1}{2k}$ for $i \in \{1, \dots, k\}$, then we have the discrete linear model

$$y = Ax + \varepsilon, \quad \text{where } A_{i,j} = \frac{1}{d}K(s_i, t_j).$$

To employ the Bayesian approach, we treat y , ε , and x as random variables. We assume that ε is Gaussian noise with variance $\sigma^2 I$,

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I), \quad \nu(\varepsilon) \propto \exp\left(-\frac{1}{2\sigma^2}\|\varepsilon\|^2\right).$$

The likelihood is then given by

$$f(y|x) = \nu(y - Ax) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Ax\|^2\right).$$

Next, we have to choose a prior distribution for the unknown. Assume that we know that $g(0) = g(1) = 0$ and that g is quite smooth, that is, the value of $g(t)$ in a point is more or less the same as in its neighbor. We will then model the unknown as

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + W_j, \quad j = 1, \dots, k, \quad (7)$$

where the term W_j follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$.

The variance γ^2 determines how much the reconstructed function x departs from the smoothness model $x_j = \frac{1}{2}(x_{j-1} + x_{j+1})$. We can write (7) as

$$Lx = W, \quad \text{where} \quad L := \frac{1}{2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}.$$

This leads to the so-called *smoothness prior*

$$f(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|Lx\|^2\right).$$

Let $L = \text{tridiag}(-1, 2, -1)$ and consider the following priors

$$f_{\text{pr},1}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|x - x_0\|^2\right) \quad \text{with covariance } \Gamma_{\text{pr},1} = \gamma^2 I;$$

$$f_{\text{pr},2}(x) \propto \exp\left(-\frac{1}{2\gamma^2}\|L(x - x_0)\|^2\right)$$

$$= \exp\left(-\frac{1}{2\gamma^2}(x - x_0)^T(L^T L)(x - x_0)\right) \quad \begin{matrix} \text{with covariance} \\ \Gamma_{\text{pr},2} = \gamma^2(L^T L)^{-1}, \end{matrix}$$

where $x_0 \in \mathbb{R}^d$ is the prior mean (assumed to be the same in both cases).
Hence

$$\bar{x}_j = x_0 + \Gamma_{\text{pr},j} A^T G_j^{-1} (y - Ax_0 - \varepsilon_0),$$

$$\Gamma_{\text{post},j} = \Gamma_{\text{pr},j} - \Gamma_{\text{pr},j} A^T G_j^{-1} A \Gamma_{\text{pr},j},$$

where $G_j = A \Gamma_{\text{pr},j} A^T + \Gamma_{\text{noise}}$ and $\Gamma_{\text{noise}} = \sigma^2 I$.

For the numerical experiment, we simulate measurements using the (smooth) ground truth signal

$$g(t) = 8t^3 - 16t^2 + 8t,$$

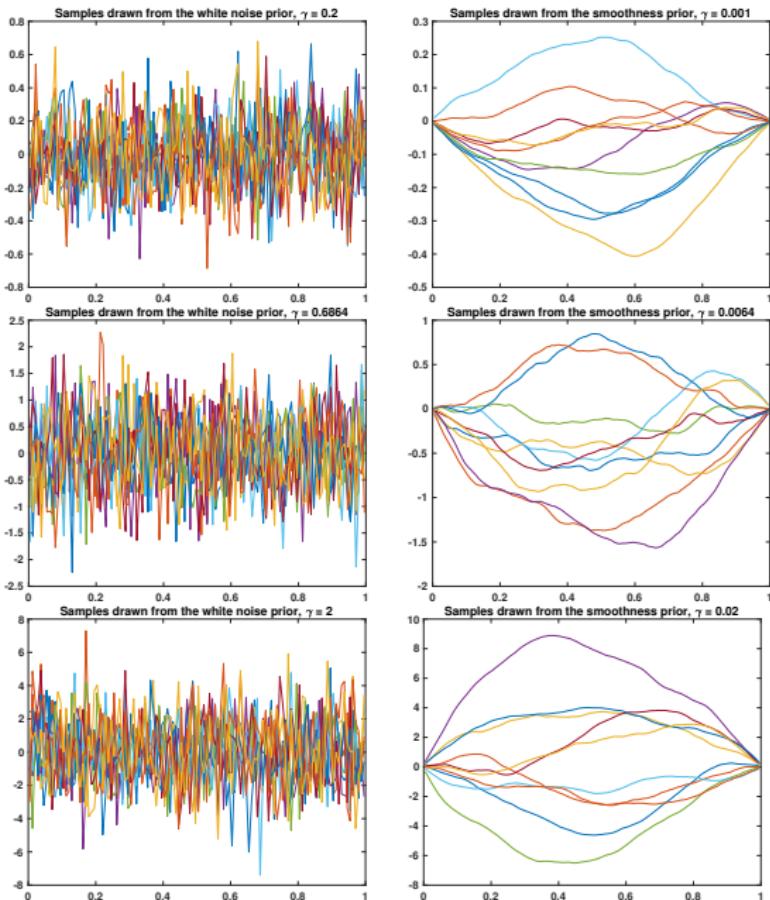
which satisfies $g(0) = g(1) = 0$. The measurements are contaminated with zero-mean 10% *relative* noise ($\sigma \approx 0.0618$) and we set $d = k = 120$.

Remark: We use a higher resolution model to simulate the measurement data. To achieve this, we generate the measurement data using a denser grid and then interpolate the forward solution onto a coarser computational grid, which is actually used to compute the reconstruction.

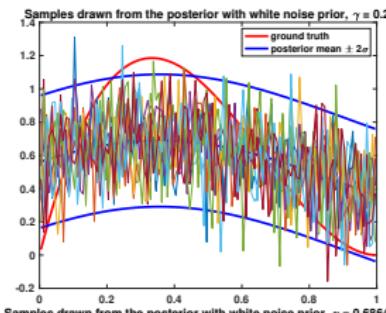
Since both the prior and the posterior are now Gaussian, we can use the coloring transformation to draw samples from the prior and posterior.

We also draw the posterior mean and the 2σ credibility envelopes.

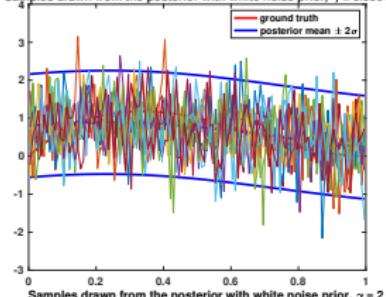
See the script deconv.py on the course webpage.



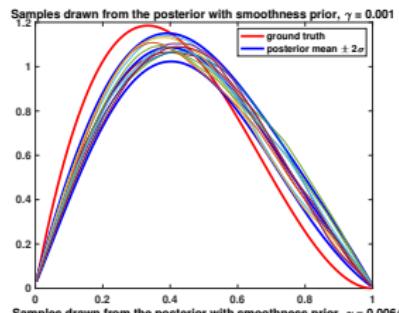
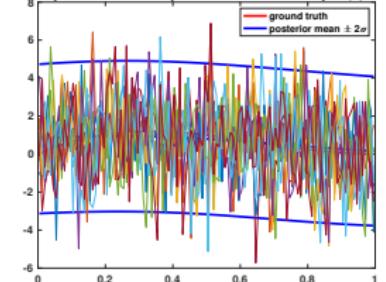
Samples drawn from the white noise prior and the smoothness prior for several values of γ .



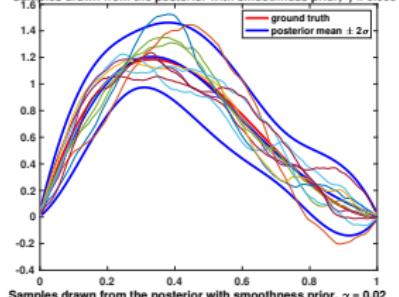
Samples drawn from the posterior with white noise prior, $\gamma \approx 0.6864$



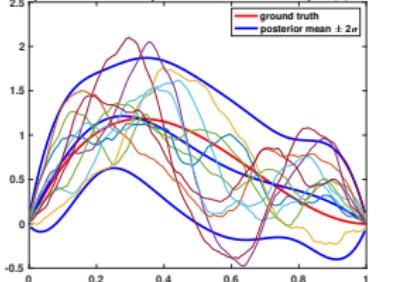
Samples drawn from the posterior with white noise prior, $\gamma = 2$



Samples drawn from the posterior with smoothness prior, $\gamma \approx 0.0064$



Samples drawn from the posterior with smoothness prior, $\gamma = 0.02$



Samples drawn from the posterior corresponding to both the white noise prior and the smoothness prior for several values of γ . We also plot the ground truth solution and the posterior mean.

A note on marginal Gaussian distributions

Let

$$f(x) \propto \exp\left(-\frac{1}{2}(x - \mu)^T \Gamma^{-1} (x - \mu)\right)$$

be a multivariate Gaussian PDF with mean μ and positive definite and symmetric covariance matrix Γ .

Q: What is Γ_{ii} ?

A: $\sigma_i^2 := \Gamma_{ii}$ is the variance of the marginal distribution with PDF

$$f(x_i) = \int_{\mathbb{R}^{n-1}} f(x_1, \dots, x_i, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n,$$

which is itself a (univariate) Gaussian PDF with mean μ_i .

This is why we can obtain the credibility envelopes by taking the square roots of the diagonal values of $\Gamma_{\text{post},j}$.

Relation to conjugate priors

The linear Gaussian setting is a special case of a more general technique, where the prior is chosen in such a way that, together with the likelihood function, the resulting posterior density belongs to the same probability distribution family as the prior. In this case, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function.

A conjugate prior is an algebraic convenience, giving a closed form expression for the posterior. In consequence, the CM estimator, MAP estimator, and variance typically also have closed form expressions and it is not necessary to use numerical integration or optimization to characterize the statistics of the posterior.

A list of the most commonly used conjugate priors can be found, e.g., at
https://en.wikipedia.org/wiki/Conjugate_prior

Kalman filter

So far we have discussed measurement models with a *static target*:

$$y_j = F(x) + \varepsilon_j, \quad \varepsilon_j \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \gamma^2 I).$$

Examples where the condition may not be valid:

- EEG
- Target tracking
- Weather forecasting

Dynamic observation models

More general observation model:

$$y_j = F(x_j) + \varepsilon_j, \quad j = 1, 2, \dots, J.$$

The observations cannot be integrated unless we have a *dynamic prior model*.

One of the simplest dynamic prior models is a 1-Markov evolution model

$$x_{j+1} = G(x_j) + \xi_{j+1}, \quad j = 0, 1, \dots, J - 1,$$

where $G: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is presumably known and ξ_{j+1} is an *innovation process*.

Examples

- Static measurement: $G(x) = x$, $\xi_{j+1} = 0$.
- Random walk model (often used in lack of anything more sophisticated):

$$x_{j+1} = x_j + \xi_{j+1}, \quad \xi_{j+1} \sim \mathcal{N}(0, \sigma^2 I).$$

- First order differential equation: assume that the unknown is a time-dependent vector $x(t) \in \mathbb{R}^d$ satisfying *ideally* the differential equation

$$x'(t) = f(x(t), t).$$

Time discretization: let $t_j = jh$, $j = 0, 1, \dots$, and write $x_j = x(t_j)$. Then we can use finite differences, e.g., forward Euler method

$$x_{j+1} = x_j + hf(x_j, t_j) + \xi_{j+1}$$

or backward Euler method

$$x_{j+1} = x_j + hf(x_{j+1}, t_{j+1}) + \xi_{j+1},$$

where ξ_{j+1} accounts for discretization errors as well as possible deviations from the ideal.

Basic form of Bayes filtering

Evolution-observation model:

$$x_{j+1} = G(x_j) + \xi_{j+1}, \quad j = 0, 1, \dots, J-1,$$

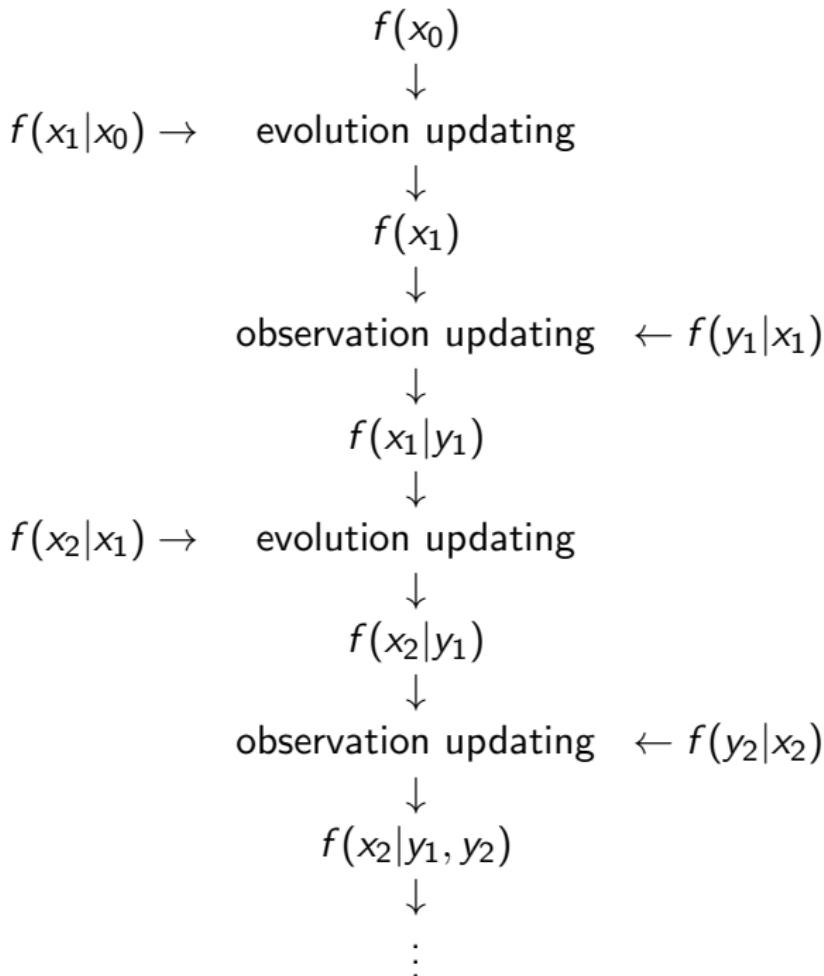
$$y_{j+1} = F(x_{j+1}) + \varepsilon_{j+1}, \quad j = 0, 1, \dots, J-1.$$

The observations y_1, \dots, y_J and the prior probability density of x_0 are given.

Adaptive algorithm

The goal is an algorithm which works as follows:

- Given the density of x_0 , *predict* the density of x_1 using the prior evolution model.
- Using the predicted density of x_1 as *prior*, calculate the posterior density of $x_1|y_1$.
- Using the posterior density of $x_1|y_1$, predict the density of x_2 .
- Using the predicted density of x_2 as *prior*, calculate the posterior density of $x_2|y_1, y_2$.
- Continue similarly.



- **Prediction step:** Given the density of x_j , calculate the density of x_{j+1} from

$$x_{j+1} = G(x_j) + \xi_{j+1}. \quad (\text{propagation problem})$$

- **Correction step:** Given the prior density of x_{j+1} , calculate the posterior density of $x_{j+1}|y_{j+1}$ using the observational model

$$y_{j+1} = F(x_{j+1}) + \varepsilon_{j+1}. \quad (\text{inference problem})$$

Particular approaches

- Linear model, Gaussian innovation and error: classical Kalman filtering.
- Linearization of non-linear evolution (or observation) model: extended Kalman filtering.
- Nonlinear and/or non-Gaussian models: particle filtering.

Kalman filter

Consider the linear ($G(\cdot) = M \cdot$, $F(\cdot) = H \cdot$) evolution-observation system

$$x_{j+1} = Mx_j + \xi_{j+1}, \quad \xi_{j+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma),$$

$$y_{j+1} = Hx_{j+1} + \varepsilon_{j+1}, \quad \varepsilon_{j+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma).$$

Prediction: Suppose $x_j \sim \mathcal{N}(m_j, C_j)$. Then

$$x_{j+1} = Mx_j + \xi_{j+1} \sim \mathcal{N}(\hat{m}_{j+1}, \hat{C}_{j+1}),$$

where $\hat{m}_{j+1} = Mm_j$ and $\hat{C}_{j+1} = MC_jM^T + \Sigma$.

Correction: Linear Gaussian setting implies $x_{j+1}|y_{j+1} \sim \mathcal{N}(m_{j+1}, C_{j+1})$, where

$$m_{j+1} = \hat{m}_{j+1} + \hat{C}_{j+1}H^T(H\hat{C}_{j+1}H^T + \Gamma)^{-1}(y_{j+1} - H\hat{m}_{j+1}),$$

$$C_{j+1} = \hat{C}_{j+1} - \hat{C}_{j+1}H^T(H\hat{C}_{j+1}H^T + \Gamma)^{-1}H\hat{C}_{j+1}.$$

Remark: The expensive step in Kalman filtering is the computation of the so-called *Kalman gain* matrix:

$$K_{j+1} = \hat{C}_{j+1}H^T(H\hat{C}_{j+1}H^T + \Gamma)^{-1}.$$

Kalman filter algorithm

Given: Initial distribution for $x_0 \sim \mathcal{N}(m_0, C_0)$, where $m_0 \in \mathbb{R}^d$ and $C_0 \in \mathbb{R}^{d \times d}$ is symmetric and positive definite.

for $j = 0, 1, 2, \dots, J - 1$, **do**

Prediction step:

$$\hat{m}_{j+1} = Mm_j$$

$$\hat{C}_{j+1} = MC_jM^T + \Sigma$$

Correction step:

$$K_{j+1} = \hat{C}_{j+1}H^T(H\hat{C}_{j+1}H^T + \Gamma)^{-1}$$

$$m_{j+1} = \hat{m}_{j+1} + K_{j+1}(y_{j+1} - H\hat{m}_{j+1})$$

$$C_{j+1} = \hat{C}_{j+1} - K_{j+1}H\hat{C}_{j+1}$$

end for

Output: Predicted distributions $\mathcal{N}(\hat{m}_{j+1}, \hat{C}_{j+1})$ and filtering distributions for $x_{j+1}|y_1, \dots, y_{j+1} \sim \mathcal{N}(m_{j+1}, C_{j+1})$, $j = 0, \dots, J - 1$.

Extended Kalman filter (non-linear evolution model)

Consider non-linear $G: \mathbb{R}^d \rightarrow \mathbb{R}^d$ and linear $F(\cdot) = H\cdot$ with

$$x_{j+1} = G(x_j) + \xi_{j+1}, \quad \xi_{j+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma),$$

$$y_{j+1} = Hx_{j+1} + \varepsilon_{j+1}, \quad \varepsilon_{j+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma),$$

with $x_0 \sim \mathcal{N}(m_0, C_0)$.

Prediction: Suppose $x_j \sim \mathcal{N}(m_j, C_j)$. We can linearize

$$x_{j+1} = G(x_j) + \xi_{j+1} \approx G(m_j) + DG(m_j)(x_j - m_j) + \xi_{j+1}.$$

An affine transformation is still Gaussian, so we obtain the approximations

$$\hat{m}_{j+1} = G(m_j), \quad \hat{C}_{j+1} = DG(m_j)C_jDg(m_j)^T + \Sigma.$$

Correction: Now that $x_{j+1} \sim \mathcal{N}(\hat{m}_{j+1}, \hat{C}_{j+1})$, we can use the linear Gaussian setting to obtain $x_{j+1}|y_{j+1} \sim \mathcal{N}(m_{j+1}, C_{j+1})$ with

$$m_{j+1} = \hat{m}_{j+1} + \hat{C}_{j+1}H^T(H\hat{C}_{j+1}H^T + \Gamma)^{-1}(y_{j+1} - B\hat{m}_{j+1}),$$

$$C_{j+1} = \hat{C}_{j+1} - \hat{C}_{j+1}H^T(H\hat{C}_{j+1}H^T + \Gamma)^{-1}H\hat{C}_{j+1}.$$

Ensemble Kalman filter (non-linear evolution model)

Consider

$$x_{j+1} = G(x_j) + \xi_{j+1}, \quad \xi_{j+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma),$$
$$y_{j+1} = Hx_{j+1} + \varepsilon_{j+1}, \quad \varepsilon_{j+1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma),$$

with $x_0 \sim \mathcal{N}(m_0, C_0)$.

The computation of the analytical predictive covariances and (in the non-linear setting) the Jacobi matrix become computationally inefficient and expensive for high-dimensional systems. The basic idea of ensemble Kalman filter is as follows:

- ① Draw a sample from the initial distribution of x_0 ("initial ensemble")
- ② Replace the predictive mean \hat{m}_{j+1} and covariance \hat{C}_{j+1} as well as the filtering mean m_{j+1} and covariance C_{j+1} with their corresponding sample means and covariances by propagating the initial ensemble through the evolution-observation model.

Ensemble Kalman filter algorithm

Given: Ensemble size N . Initial ensemble $\{x_0^{(i)}\}_{i=1}^N$ drawn from the initial distribution of $x_0 \sim \mathcal{N}(m_0, C_0)$, where $m_0 \in \mathbb{R}^d$ and $C_0 \in \mathbb{R}^{d \times d}$ is symmetric and positive definite.

Parameter $s \in \{0, 1\}$.

for $j = 0, 1, 2, \dots, J - 1$, **do**

Prediction step:

$$\text{draw } \xi_{j+1}^{(i)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \quad i = 1, \dots, N,$$

$$\hat{x}_{j+1}^{(i)} = G(x_j^{(i)}) + \xi_{j+1}^{(i)}, \quad i = 1, \dots, N,$$

$$\hat{m}_{j+1} = \frac{1}{N} \sum_{i=1}^N \hat{x}_{j+1}^{(i)} \quad \text{and} \quad \hat{C}_{j+1} = \frac{1}{N} \sum_{i=1}^N (\hat{x}_{j+1}^{(i)} - \hat{m}_{j+1})(\hat{x}_{j+1}^{(i)} - \hat{m}_{j+1})^T.$$

Correction step:

$$\text{draw } \varepsilon_{j+1}^{(i)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma), \quad i = 1, \dots, N,$$

$$y_{j+1}^{(i)} = y_{j+1} + s\varepsilon_{j+1}^{(i)}, \quad i = 1, \dots, N,$$

$$K_{j+1} = \hat{C}_{j+1} H^T (H \hat{C}_{j+1} H^T + \Gamma)^{-1},$$

$$x_{j+1}^{(i)} = \hat{x}_{j+1}^{(i)} + K_{j+1}(y_{j+1}^{(i)} - H \hat{x}_{j+1}^{(i)}), \quad i = 1, \dots, N.$$

end for

Output: Ensembles $\{x_j^{(i)}\}_{i=1}^N$, $j = 0, \dots, J$.

Remark:

- Setting the parameter $s = 1$ is suitable at approximating the Kalman filter in linear Gaussian settings: if each prediction particle $\tilde{x}_{j+1}^{(i)}$ is distributed according to a non-degenerate Gaussian distribution, then in the linear Gaussian setting the “corrected” particle $x_{j+1}^{(i)}$ will be Gaussian with mean and covariance that agree with the usual Kalman filter formulae.
- Setting the parameter $s = 0$ is natural if viewing the algorithm as a sequential optimizer in problems where the filtering distributions are not well approximated by Gaussians.

Numerical example

We wish to track the state $x_k = \begin{bmatrix} p_k \\ v_k \end{bmatrix} \in \mathbb{R}^2$ of a moving particle at discrete times $t_k = k\Delta t$, $k = 0, 1, 2, \dots$. The first component p_k corresponds to the position of the particle while the second component $v_k = \dot{p}_k$ is its velocity at time $k = 0, 1, 2, \dots$. Let us also denote the *unknown* acceleration of the particle as $a_k = \ddot{v}_k = \ddot{p}_k$ for $k = 0, 1, 2, \dots$. We have the following dynamics:

$$\begin{cases} p_k = p_{k-1} + v_{k-1}\Delta t + \frac{1}{2}a_{k-1}(\Delta t)^2 \\ v_k = v_{k-1} + a_{k-1}\Delta t \end{cases} \Leftrightarrow x_k = \underbrace{\begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}}_{=:M} x_{k-1} + \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} a_{k-1}.$$

If we assume that $a_{k-1} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, then the innovation process is

$$\xi_k := \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} a_{k-1} \sim \mathcal{N}(0, \Sigma), \text{ where } \Sigma := \begin{bmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{bmatrix} \begin{bmatrix} \frac{1}{2}(\Delta t)^2 & \Delta t \end{bmatrix}$$
$$= \begin{bmatrix} \frac{1}{4}(\Delta t)^4 & \frac{1}{2}(\Delta t)^3 \\ \frac{1}{2}(\Delta t)^3 & (\Delta t)^2 \end{bmatrix}. \text{ This yields the } \textit{evolution model}$$

$$x_k = Mx_{k-1} + \xi_k, \quad \xi_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma).$$

Meanwhile, we can only measure the location of the particle, so the *observation model* is given by

$$y_k = Hx_k + \varepsilon_k, \quad \varepsilon_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \gamma^2),$$

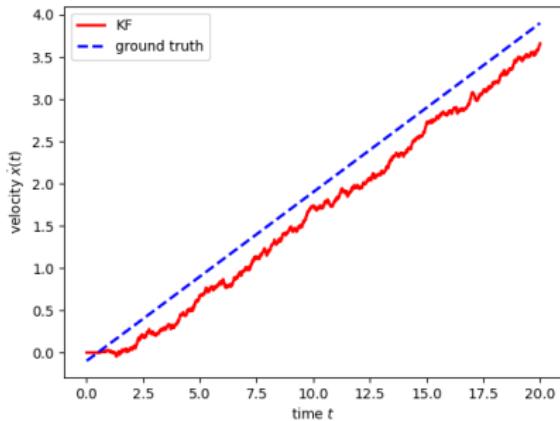
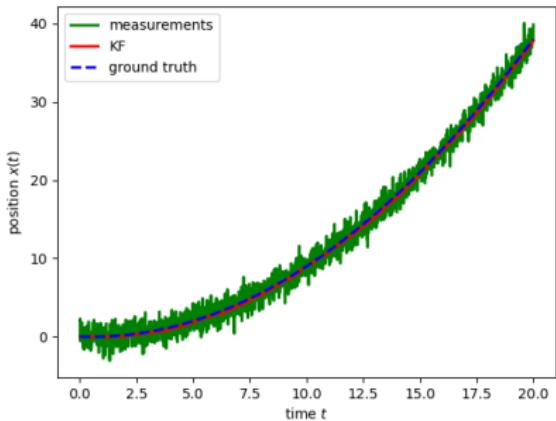
where $H := [1 \ 0]$ and y_k is a noisy measurement of the particle's location at time k , with the noise level assumed to be $\gamma := 1$.

We can now implement the Kalman filter for this model problem. We can assume that the initial position of the particle is perfectly known:

$$x_0 = \mathbb{E}[x_0] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad C_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{2 \times 2}.$$

The Kalman filter can be used to obtain the mean and covariance of the (Gaussian) filtering distribution $(p_k, v_k)|y_1, \dots, y_k$ for $k = 1, 2, 3, \dots$. We can plot the means of the filtered positions $(t_k, \mathbb{E}[p_k|y_1, \dots, y_k])$ and velocities $(t_k, \mathbb{E}[v_k|y_1, \dots, y_k])$.

The implementation is left as an exercise.



The true trajectory of the particle was $x(t) = 0.1(t^2 - t)$ (left figure) with velocity $\dot{x}(t) = 0.2t - 0.1$ (right figure). The observations at time points $t_k = k\Delta t$, with $\Delta t = 0.01$ and $k = 1, 2, \dots$, were contaminated with centered, unit-variance Gaussian noise (left figure). The red graphs correspond to the means of the filtered positions ($t_k, \mathbb{E}[p_k | y_1, \dots, y_k]$) (left figure) and velocities ($t_k, \mathbb{E}[v_k | y_1, \dots, y_k]$) (right figure).

Remarks:

- The Kalman filter is optimal in the sense that it gives the best estimator of the mean in an online setting.
- In the linear case ($G(\cdot) = M\cdot$), the ensemble Kalman filter converges to the Kalman filter. When applicable, the ensemble Kalman filter is much more efficient than particle filters. A primary advantage of ensemble methods is that they can provide good state estimation even when the number of particles is *not* large.

Appendix: General evolution-observation model and particle filters

General evolution-observation model and particle filters

Consider the more general model

$$x_{j+1} = G(x_j, \xi_{j+1}), \quad j = 0, 1, \dots, J-1,$$
$$y_{j+1} = F(x_{j+1}, \varepsilon_{j+1}), \quad j = 0, 1, \dots, J-1.$$

The functions F and G are assumed to be known. We also assume that $\xi_{j+1} \perp x_j$ and $\varepsilon_{j+1} \perp x_{j+1}$.

Observation and evolution models may be cumbersome or impossible to linearize (e.g., non-differentiable or no closed form). One may try Monte Carlo methods to simulate the distributions by random samples.

The goal in *particle filter* methods is to produce sequentially an ensemble of random samples $\{x_j^{(1)}, \dots, x_j^{(N)}\}$ distributed according to the conditional probability distributions $f(x_{j+1}|y_1, \dots, y_j)$ (prediction) or $f(x_j|y_1, \dots, y_j)$ (filtering). The vectors $x_j^{(i)}$ are called *particles* of the sample, hence the name particle filter.

One straightforward particle filter method is known as the *sampling importance resampling* filter (also known as *SIR* or *bootstrap filter*).

Sampling importance resampling

- ① Set $j = 0$ and generate an initial sample $S_0 = \{x_0^{(i)}\}_{i=1}^N$ by drawing from the density $f(x_0)$. (This may require MCMC if the initial density is complicated, e.g., non-Gaussian.)
- ② **Prediction:** Draw $\xi_{j+1}^{(i)}$ from the distribution of ξ_{j+1} and set $\hat{x}_{j+1}^{(i)} = G(x_j^{(i)}, \xi_{j+1}^{(i)})$ for $1 \leq i \leq N$. Let $\hat{S}_{j+1} = \{\hat{x}_{j+1}^{(i)}\}_{i=1}^N$.
- ③ **Correction:** Assume that from the observational model $y_j = F(x_j, \varepsilon_j)$, we can calculate the likelihood density $Cf(y_j|x_j)$, $j = 1, 2, \dots, J$, up to a multiplicative constant $C > 0$.[†] Calculate the importance of each propagated particle

$$\hat{w}_{j+1}^{(i)} = Cf(y_{j+1}|\hat{x}_{j+1}^{(i)}), \quad 1 \leq i \leq N,$$

and compute their *relative importance*

$$w_{j+1}^{(i)} = \frac{\hat{w}_{j+1}^{(i)}}{W}, \quad W = \sum_{i=1}^N \hat{w}_{j+1}^{(i)}.$$

Resampling: draw a new sample $S_{j+1} = \{x_{j+1}^{(1)}, \dots, x_{j+1}^{(N)}\}$ from the sample \hat{S}_{j+1} , with the probability of drawing $\hat{x}_{j+1}^{(i)}$ set equal to $w_{j+1}^{(i)}$. Set $j \leftarrow j + 1$ and return to step 2.

[†]E.g., if $y_j = F(x_j) + \varepsilon_j$, $\varepsilon_j \sim \mathcal{N}(0, \gamma^2 I)$, then $f(y_j|x_j) \propto \exp(-\frac{1}{2\gamma^2} \|y_j - F(x_j)\|^2)$.

Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fourteenth lecture, January 27, 2025

Why is sampling needed?

Recall that the CM estimator and the conditional covariance require solving integration problems involving the posterior density:

$$\hat{x}_{\text{CM}} = \mathbb{E}[x|y] = \int_{\mathbb{R}^d} x f(x|y) dx$$

$$\text{Cov}(x|y) = \int_{\mathbb{R}^d} (x - \hat{x}_{\text{CM}})(x - \hat{x}_{\text{CM}})^T f(x|y) dx.$$

In a non-Gaussian case, these integrals cannot typically be expressed in closed form. Therefore one must resort to numerical integration.

Suppose that our goal is to estimate some quantity of the form

$$\mathcal{I} = \mathbb{E}[G(X)] = \int_{\mathbb{R}^d} G(x)p(x) dx,$$

where $p: \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ is a probability density function and G is a quantity of interest.

For example, if p is a posterior density and $G(x) = x$, then \mathcal{I} would be precisely the CM estimator.

In principle, we could use a quadrature rule

$$\mathcal{I} = \int_{\mathbb{R}^d} G(x)p(x) dx \approx \sum_{j=1}^N w_j G(x_j)p(x_j)$$

with some suitable weights $\{w_j\}_{j=1}^N$ and nodes $\{x_j\}_{j=1}^N$. However, the design of efficient quadrature rules for high-dimensional problems is challenging. Moreover, the implementation of a quadrature rule would require reliable information about the location of the *support* of the probability density p .

Often it is more advisable to resort to sampling: draw a large enough sample $\{x_j\}_{j=1}^N$ from the probability distribution corresponding to p , and use these points to approximate the integral as

$$\mathcal{I} = \int_{\mathbb{R}^d} G(x)p(x) dx = \mathbb{E}[G(x)] \approx \frac{1}{N} \sum_{j=1}^N G(x_j) = \mathcal{I}_N.$$

According to the Law of Large Numbers, for any integrable G there holds

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N G(x_j) = \mathcal{I} \quad \text{almost surely.}$$

Furthermore, if G is square-integrable, then the Central Limit Theorem states that

$$\text{Var}(\mathcal{I} - \mathcal{I}_N) \approx \frac{\text{Var}(G(X))}{N},$$

i.e., the discrepancy between \mathcal{I} and \mathcal{I}_N should go to zero like $1/\sqrt{N}$.

Markov Chain Monte Carlo

Discrete time Markov chains

A sequence $\{X_k\}_{k=0}^{\infty}$ of random variables is called a *discrete time Markov chain* if the probability distribution of any X_{k+1} depends only on the previous state X_k :

$$\pi(x_{k+1} | x_0, \dots, x_k) = \pi(x_{k+1} | x_k).$$

Here, $\pi(x_{k+1}|x_0, \dots, x_k)$ (resp. $\pi(x_{k+1}|x_k)$) denotes the PDF of X_{k+1} conditioned on the previous states X_0, \dots, X_k (resp. X_k). Suppose in addition that there exists a *probability transition kernel* $q(x, y)$ such that

$$\pi(x_{k+1} | x_k) = q(x_k, x_{k+1}).$$

Then the Markov chain is called *time invariant* (or *time homogeneous*) since the kernel q is independent of the time k .

Remark. We assume that transition kernels satisfy the following:

- for each fixed $x \in \mathbb{R}^d$, the function $y \mapsto q(x, y)$ is a probability density. In particular, $\mathbb{P}(Y \in B | X = x) = \int_B q(x, y) dy$ and $\int_{\mathbb{R}^d} q(x, y) dy = 1$.

Example (Random walk in \mathbb{R}^d)

A *random walk* in \mathbb{R}^d is a process of moving around by taking random steps. Elementary random walk:

1. Choose a starting point $x_0 \in \mathbb{R}^d$ and a “step size” $\sigma > 0$. Set $k = 0$.
2. Draw a random vector $w_{k+1} \sim \mathcal{N}(0, I)$ and set $x_{k+1} = x_k + \sigma w_{k+1}$.
3. Set $k \leftarrow k + 1$ and return to step 2, unless your stopping criterion is satisfied.

The location of a random walk at time k is a realization of the random variable X_k , and we have an evolution model

$$X_{k+1} = X_k + \sigma W_{k+1}, \quad W_{k+1} \sim \mathcal{N}(0, I).$$

The conditional density of X_{k+1} , given $X_k = x_k$, is

$$\pi(x_{k+1}|x_k) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|x_k - x_{k+1}\|^2\right) = q(x_k, x_{k+1}),$$

where q is the (time invariant) transition kernel.

Let X be a random variable with probability density $p(x)$.

Let $q(x, y)$ be an arbitrary transition kernel used to generate a new random variable Y given $X = x$, i.e.,

$$\pi(y | x) = q(x, y).$$

By the law of total probability, the probability density of Y is

$$\pi(y) = \int_{\mathbb{R}^d} \pi(y | x)p(x) dx = \int_{\mathbb{R}^d} q(x, y)p(x) dx.$$

If the probability density of Y is equal to the probability density of X ,

$$\int_{\mathbb{R}^d} q(x, y)p(x) dx = p(y),$$

then we call p an *invariant density* of the transition kernel q .

Definition (Irreducible transition kernel)

The transition kernel q is *irreducible* if, regardless of the starting point, the Markov chain generated by q can visit any set of positive measure with positive probability.

Definition (Periodic transition kernel)

The transition kernel q is *periodic* if, for some integer $m \geq 2$, there is a set of disjoint nonempty sets $\{E_1, \dots, E_m\} \subset \mathbb{R}^d$ such that for all $j \in \{1, \dots, m\}$ and for all $x \in E_j$:

$$\mathbb{P}(Y \in E_{\text{mod}(j,m)+1} | X = x) = \int_{E_{\text{mod}(j,m)+1}} q(x, y) dy = 1.$$

That is, the Markov chain generated by q remains in a periodic loop forever.

Definition (Aperiodic transition kernel)

The transition kernel q is *aperiodic* if it is not periodic.

Theorem

Let $\{X_k\}_{k=0}^{\infty}$ be a time invariant Markov chain with the transition kernel q , i.e.,

$$\pi(x_{k+1} \mid x_k) = q(x_k, x_{k+1}).$$

Assume that p is an invariant density of q and the following technical conditions hold:

- q is irreducible;
- q is aperiodic.

Then for all $x_0 \in \mathbb{R}^d$ and any (measurable) $B \subseteq \mathbb{R}^d$, there holds

$$\lim_{N \rightarrow \infty} \mathbb{P}(X_N \in B \mid X_0 = x_0) = \int_B p(x) dx.$$

Moreover, for any integrable $G: \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N G(X_j) = \int_{\mathbb{R}^d} G(x)p(x) dx \quad \text{a.s.}$$

Suppose we want to sample some probability density p and we know that it is invariant with respect to transition kernel q . Then we can proceed as follows:

- ① Select starting point x_0 and set $k = 0$.
- ② Draw x_{k+1} from $q(x_k, x_{k+1})$.
- ③ Set $k \leftarrow k + 1$ and return to step 2.

The previous theorem implies that the sample $\{x_k\}_{k=0}^N$ is asymptotically distributed according to p as $N \rightarrow \infty$.

This raises the question: *given a probability density p , how do you find a kernel q such that p is its invariant density?*

The *Metropolis–Hastings algorithm* is a method to construct such a kernel!

Derivation of the Metropolis–Hastings algorithm

We are interested in obtaining samples from the probability density p . Consider the following Markov process: if you are currently situated at some $x \in \mathbb{R}^d$, either

- ① stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
- ② move away from x using a transition kernel $R(x, y)$ otherwise.

Here, both $R(x, y)$ and $r(x)$ are as yet undetermined—the trick will be to calibrate these in order to find a kernel such that p is its invariant density as discussed on the previous slide.

Since R is a transition kernel, $y \mapsto R(x, y)$ is a probability density and hence

$$\int_{\mathbb{R}^d} R(x, y) dy = 1 \quad \text{for all } x \in \mathbb{R}^d.$$

Denote by \mathcal{A} the event of moving away from x and by $\neg\mathcal{A}$ the event of not moving. Clearly

$$\mathbb{P}(\mathcal{A}) = 1 - r(x) \quad \text{and} \quad \mathbb{P}(\neg\mathcal{A}) = r(x).$$

Given a current state $X = x$, we want to know what is the probability density of Y generated by the aforementioned strategy. Let $B \subseteq \mathbb{R}^d$ and consider the probability of the event $Y \in B$. Then

$$\begin{aligned}\mathbb{P}(Y \in B \mid X = x) &= \mathbb{P}(Y \in B \mid X = x, \mathcal{A})\mathbb{P}(\mathcal{A}) \quad (\text{move away from } x) \\ &\quad + \mathbb{P}(Y \in B \mid X = x, \neg\mathcal{A})\mathbb{P}(\neg\mathcal{A}). \quad (\text{stay put at } x)\end{aligned}$$

The probability of arriving in B through a move is

$$\mathbb{P}(Y \in B \mid X = x, \mathcal{A}) = \int_B R(x, y) dy.$$

The only way to arrive in B without moving is if x is already in B :

$$\mathbb{P}(Y \in B \mid X = x, \neg\mathcal{A}) = \mathbf{1}_B(x) = \begin{cases} 1 & \text{if } x \in B, \\ 0 & \text{if } x \notin B. \end{cases}$$

Hence

$$\begin{aligned}\mathbb{P}(Y \in B \mid X = x) &= \int_B \overbrace{(1 - r(x))R(x, y)}^{=: K(x, y)} dy + r(x)\mathbf{1}_B(x) \\ &= \int_B K(x, y) dy + r(x)\mathbf{1}_B(x).\end{aligned}$$

The probability of $Y \in B$ can be obtained by marginalizing over x :

$$\begin{aligned}\mathbb{P}(Y \in B) &= \int_{\mathbb{R}^d} \mathbb{P}(Y \in B \mid X = x) p(x) dx \\&= \int_{\mathbb{R}^d} \left(\int_B K(x, y) dy \right) p(x) dx + \int_{\mathbb{R}^d} r(x) \mathbf{1}_B(x) p(x) dx \\&= \int_B \left(\int_{\mathbb{R}^d} K(x, y) p(x) dx \right) dy + \int_B r(x) p(x) dx \\&= \int_B \left(\int_{\mathbb{R}^d} K(x, y) p(x) dx + r(y) p(y) \right) dy \\&= \int_B \left(\int_{\mathbb{R}^d} K(x, y) p(x) dx - \int_{\mathbb{R}^d} K(y, x) p(y) dx + p(y) \right) dy,\end{aligned}$$

where we used $\int_{\mathbb{R}^d} K(y, x) dx = (1 - r(y)) \int_{\mathbb{R}^d} R(y, x) dx = 1 - r(y)$.

If the *balance equation*

$$\int_{\mathbb{R}^d} p(y) K(y, x) dx = \int_{\mathbb{R}^d} p(x) K(x, y) dx \tag{1}$$

holds, then

$$\mathbb{P}(Y \in B) = \int_B p(y) dy \quad \text{as desired.}$$

The Metropolis–Hastings algorithm is a technique for finding a kernel K that satisfies the *detailed balance equation*

$$p(y)K(y, x) = p(x)K(x, y),$$

which implies (1). Let us start with a *proposal density* $q(x, y)$, chosen so that generating a Markov chain with it is easy. (For this reason, a Gaussian kernel is a very popular choice.) There are three separate cases:

- ① If $p(y)q(y, x) = p(x)q(x, y)$, then set $r(x) = 0$,
 $R(x, y) = K(x, y) = q(x, y)$ and the previous analysis ensures that p is an invariant density for kernel q .
- ② If $p(y)q(y, x) < p(x)q(x, y)$, then define the kernel K to be

$$K(x, y) = \alpha(x, y)q(x, y),$$

where α is chosen s.t. $p(y)\alpha(y, x)q(y, x) = p(x)\alpha(x, y)q(x, y)$. We can make the selection

$$\alpha(y, x) = 1 \quad \text{and} \quad \alpha(x, y) = \frac{p(y)q(y, x)}{p(x)q(x, y)} < 1.$$

- ③ If $p(y)q(y, x) > p(x)q(x, y)$, then in complete analogy to the above:

$$\alpha(x, y) = 1 \quad \text{and} \quad \alpha(y, x) = \frac{p(x)q(x, y)}{p(y)q(y, x)} < 1.$$

In summary, we define K as

$$K(x, y) = \alpha(x, y)q(x, y), \quad \alpha(x, y) = \min \left\{ 1, \frac{p(y)q(y, x)}{p(x)q(x, y)} \right\}.$$

Even though the expression for K seems complicated, it turns out that the drawing can be performed according to the following procedure.

Metropolis–Hastings algorithm

- ① Choose $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
- ② Given $x = x^{(k)}$, draw y using the transition kernel $q(x, y)$ of your choosing.
- ③ Calculate the acceptance ratio

$$\alpha(x, y) = \min \left\{ 1, \frac{p(y)q(y, x)}{p(x)q(x, y)} \right\}.$$

- ④ Flip the α -coin: draw $t \sim \mathcal{U}([0, 1])$. If $\alpha > t$, set $x^{(k+1)} = y$, otherwise stay put at x and set $x^{(k+1)} = x^{(k)}$.
- ⑤ Set $k \leftarrow k + 1$ and return to step 2.

Remark. Note that due to the form of α , both the target p and the proposal density q can be *unnormalized* within the Metropolis–Hastings algorithm.

Why does this work?

Let us focus on the main loop of the Metropolis–Hastings algorithm:

- Given x , draw y using the transition kernel $q(x, y)$.
- Calculate the acceptance ratio $\alpha(x, y) = \min \left\{ 1, \frac{p(y)q(x,y)}{p(x)q(x,y)} \right\}$.
- Draw $t \sim \mathcal{U}([0, 1])$. If $\alpha > t$, accept y , otherwise stay put at x .

Recall that \mathcal{A} was the event of moving in the Markov chain. Then

$$\begin{aligned}\mathbb{P}(\mathcal{A}|y, x) &= \text{"probability of accepting transition"} = \alpha(x, y), \\ \mathbb{P}(y|x) &= \text{"probability of drawing } y\text{"} = q(x, y).\end{aligned}$$

Then

$$\begin{aligned}\text{"probability of accepted } y\text{"} &= \mathbb{P}(\mathcal{A}, y|x) \\ &= \mathbb{P}(\mathcal{A}|y, x)\mathbb{P}(y|x) \\ &= \alpha(x, y)q(x, y) = K(x, y),\end{aligned}$$

as desired.

Example

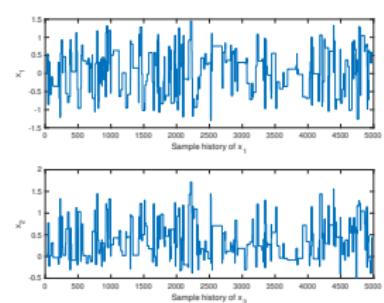
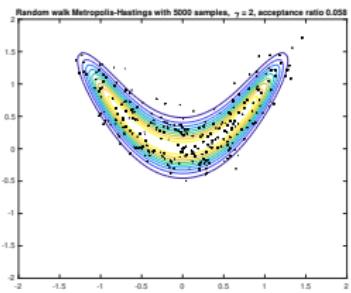
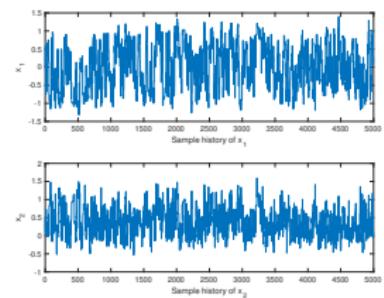
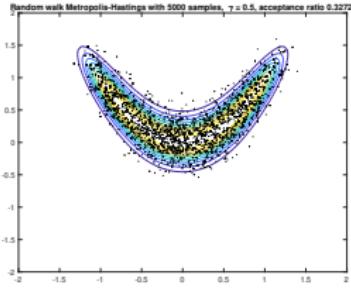
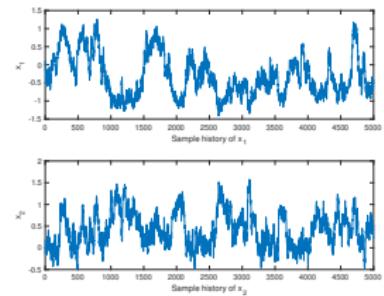
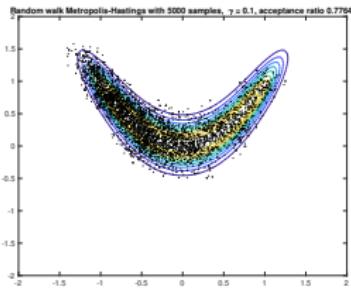
Let us consider sampling from the density

$$p(x_1, x_2) \propto \exp(-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4).$$

As the proposal distribution, we use the random walk model $Y = X + W$, $W \sim \mathcal{N}(0, \gamma^2 I)$, with the kernel

$$q(x, y) \propto \exp\left(-\frac{1}{2\gamma^2}\|x - y\|^2\right).$$

We draw 5000 samples from the probability distribution with density p using three different step sizes: $\gamma = 0.1$, $\gamma = 0.5$, and $\gamma = 2$.



Derivation of the single component Gibbs sampler

We continue to be interested in sampling the distribution with density $p(x)$. The single component Gibbs sampler is based on the same Markov process that was introduced in the derivation of Metropolis–Hastings: if you are currently situated at some $x \in \mathbb{R}^d$, either

- ① stay put at x with the probability $r(x)$, $0 \leq r(x) \leq 1$, or
- ② move away from x using a transition kernel $R(x, y)$ otherwise.

Recall also the definition we made in the Metropolis–Hastings derivation:

$$K(x, y) = (1 - r(x))R(x, y).$$

Suppose that x is a d -variate random variable. For the single component Gibbs sampler, we set $r(x) = 0$ (moving is obligatory) and define the transition kernel

$$K(x, y) = R(x, y) = \prod_{i=1}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d),$$

where $p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) = \frac{p(y_1, \dots, y_i, x_{i+1}, \dots, x_d)}{\int_{\mathbb{R}} p(y_1, \dots, y_i, x_{i+1}, \dots, x_d) dy_i}$.

This transition kernel K does not in general satisfy the detailed balance equation, but it does satisfy the standard balance equation, which is sufficient to ensure that p is the invariant density of the Markov chain (see derivation of the Metropolis–Hastings method).

Theorem

The transition kernel

$$K(x, y) = \prod_{i=1}^d p(y_i \mid y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d),$$

where $p(y_i \mid y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) = \frac{p(y_1, \dots, y_i, x_{i+1}, \dots, x_d)}{\int_{\mathbb{R}} p(y_1, \dots, y_i, x_{i+1}, \dots, x_d) dy_i}$,
satisfies

$$\int_{\mathbb{R}^d} p(y) K(y, x) dx = \int_{\mathbb{R}^d} p(x) K(x, y) dx.$$

Proof. We begin with the left-hand side of the balance equation and consider $\int_{\mathbb{R}^d} K(y, x) dx$. We integrate inductively over the variables in the order x_d, x_{d-1}, \dots, x_1 :

$$\int_{\mathbb{R}} K(y, x) dx_d = \int_{\mathbb{R}} \left(\prod_{i=1}^d p(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_d) \right) dx_d$$

$$= \left(\prod_{i=1}^{d-1} p(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_d) \right) \underbrace{\int_{\mathbb{R}} p(x_d | x_1, \dots, x_{d-1}) dx_d}_{=1}$$

$$= \prod_{i=1}^{d-1} p(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_d)$$

$$\Rightarrow \int_{\mathbb{R}} \int_{\mathbb{R}} K(y, x) dx_d dx_{d-1} = \int_{\mathbb{R}} \left(\prod_{i=1}^{d-1} p(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_d) \right) dx_{d-1}$$

$$= \left(\prod_{i=1}^{d-2} p(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_d) \right) \underbrace{\int_{\mathbb{R}} p(x_{d-1} | x_1, \dots, x_{d-2}, y_d) dx_{d-1}}_{=1}$$

$$= \prod_{i=1}^{d-2} p(x_i | x_1, \dots, x_{i-1}, y_{i+1}, \dots, y_d) \quad \Rightarrow \quad \dots$$

Proceeding by inductively integrating over $x_{d-2}, x_{d-3}, \dots, x_1$, we obtain

$$\int_{\mathbb{R}^d} K(y, x) dx = 1 \text{ and thus } \int_{\mathbb{R}^d} p(y) K(y, x) dx = p(y) \int_{\mathbb{R}^d} K(y, x) dx = p(y).$$

Next we consider the right-hand side of the balance equation. Recall that $K(x, y) = \prod_{i=1}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d)$. We integrate inductively over the variables, this time in the order x_1, \dots, x_d :

$$\begin{aligned}
& \int_{\mathbb{R}} p(x) K(x, y) dx_1 = K(x, y) \int_{\mathbb{R}} p(x_1, x_2, \dots, x_d) dx_1 && (K \text{ is independent of } x_1) \\
&= \left(\prod_{i=2}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) \right) \underbrace{p(y_1 | x_2, \dots, x_d)}_{=\frac{p(y_1, x_2, \dots, x_d)}{\int_{\mathbb{R}} p(x_1, x_2, \dots, x_d) dx_1}} \int_{\mathbb{R}} p(x_1, x_2, \dots, x_d) dx_1 \\
&= \left(\prod_{i=2}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) \right) p(y_1, x_2, \dots, x_d) \\
\Rightarrow & \int_{\mathbb{R}} \int_{\mathbb{R}} p(x) K(x, y) dx_1 dx_2 = \int_{\mathbb{R}} \left(\prod_{i=2}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) \right) p(y_1, x_2, \dots, x_d) dx_2 \\
&= \left(\prod_{i=3}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) \right) \underbrace{p(y_2 | y_1, x_3, \dots, x_d)}_{=\frac{p(y_1, y_2, x_3, \dots, x_d)}{\int_{\mathbb{R}} p(y_1, x_2, x_3, \dots, x_d) dx_2}} \int_{\mathbb{R}} p(y_1, x_2, \dots, x_d) dx_2 \\
&= \left(\prod_{i=3}^d p(y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_d) \right) p(y_1, y_2, x_3, \dots, x_d) \quad \Rightarrow \quad \dots
\end{aligned}$$

Proceeding by inductively integrating over x_3, \dots, x_d , we eventually obtain $\int_{\mathbb{R}^d} p(x) K(x, y) dx = p(y)$. Therefore the balance equation holds.



Single component Gibbs sampler

- ① Choose the initial value $x^{(0)} \in \mathbb{R}^d$ and set $k = 0$.
- ② Draw the next sample as follows:

- (i) Set $x = x^{(k)}$ and $j = 1$.
- (ii) Draw $t \in \mathbb{R}$ from the one-dimensional distribution

$$p(t | y_1, \dots, y_{j-1}, x_{j+1}, \dots, x_d) \propto p(y_1, \dots, y_{j-1}, t, x_{j+1}, \dots, x_d)$$

and set $y_j = t$.

- (iii) If $j = d$, set $y = (y_1, \dots, y_d)$ and terminate the inner loop. Otherwise, set $j \leftarrow j + 1$ and return to step (ii).
- ③ Set $x^{(k+1)} = y$, increase $k \leftarrow k + 1$ and return to step 2.

Example

Let us consider the density from before

$$p(x_1, x_2) = \frac{1}{Z} \exp(-10(x_1^2 - x_2)^2 - (x_2 - \frac{1}{4})^4),$$

where the normalizing constant is $Z = 1.1813\dots$

This time we use the Gibbs sampler. To sample the univariate densities that arise in the process, we use inverse transform sampling. In this case, the explicit algorithm we use is written below.

Fix $x^{(0)} \in \mathbb{R}^2$ and set $x = x^{(0)}$;

For $k = 1, \dots, N$, do

Calculate $\Phi_1(t) = \int_{-\infty}^t p(x_1, x_2) dx_1$;

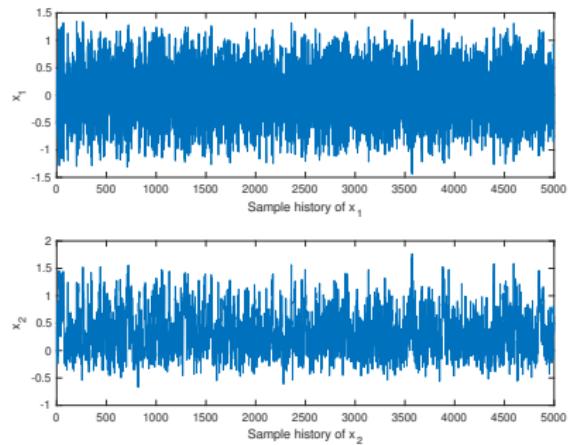
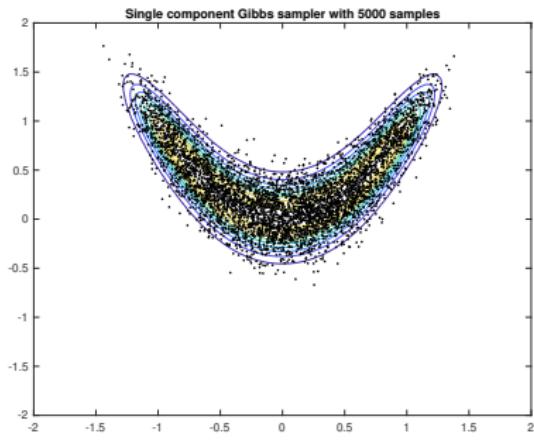
Draw $u \sim \mathcal{U}([0, 1])$, set $x_1 = \Phi_1^{-1}(u)$;

Calculate $\Phi_2(t) = \int_{-\infty}^t p(x_1, x_2) dx_2$;

Draw $u \sim \mathcal{U}([0, 1])$, set $x_2 = \Phi_2^{-1}(u)$;

Set $x^{(k)} = x$.

End

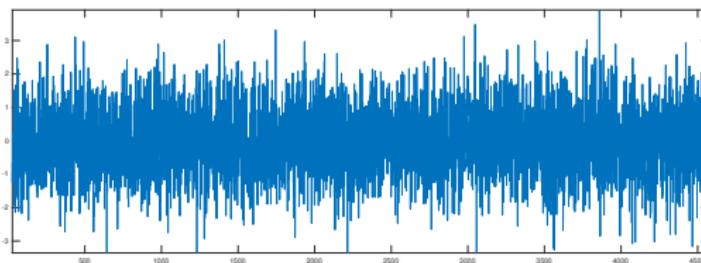


Computational remarks about MCMC

- As a general rule of thumb, one should aim at roughly 30% acceptance rates when using Gaussian (or close to Gaussian) proposal and target densities with MH.
- It usually takes the Markov chain a number of iterations to reach the steady state. To this end, it is usually advisable to discard the first N_0 obtained samples since they may not be representative of the target distribution—this is the so-called “burn-in” period. The length of the burn-in period varies depending on the application, but one might consider throwing away the first $\sim 5 - 10\%$ steps for a sufficiently large sample size as an example.
- In MH, using a Gaussian kernel (e.g., random walk Metropolis–Hastings) is a popular choice due to the ease of implementation. While it is a safe choice, it does not take into account the form of the posterior density. To increase efficiency, it is advisable to take the shape of the density into account when designing the proposal density. In the high-dimensional setting, this is especially useful if the posterior density is *anisotropic* (stretched in some directions).

Computational remarks about MCMC

- The proposal distribution in MH can also be updated while the sampling algorithm moves around the posterior density. This process is called *adaptation*.
- Visual assessment: we are aiming for independent sample points, where the sample histories look like a “fuzzy worm”. One could aim at something like the Gaussian white noise signal below:



- More quantitatively, the independence of consecutive draws can be estimated from the sample itself by computing its (sample-based) autocovariance.

A note on convergence

The convergence of the Metropolis–Hastings and Gibbs sampler algorithms depends on whether they satisfy the ergodicity conditions from before. There are known sufficient conditions concerning the density p that guarantee the ergodicity of these methods. For example, the following proposition gives some relatively general conditions.

Proposition

- (a) Let $p: \mathbb{R}^d \rightarrow \mathbb{R}_+$ and let $q: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}_+$ be a candidate-generating kernel. If the Markov chain corresponding to q is aperiodic, then the Metropolis–Hastings chain is also aperiodic. Further, if the Markov chain corresponding to q is irreducible and $\alpha(x, y) > 0$ for all $(x, y) \in E_+ \times E_+$, where $E_+ := \{x \in \mathbb{R}^d \mid p(x) > 0\}$, then the Metropolis–Hastings chain is irreducible.
- (b) Let p be a lower semicontinuous density and E_+ as above. The Gibbs sampler defines an irreducible and aperiodic transition kernel if E_+ is connected and each $(d - 1)$ -dimensional marginal $p(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d) = \int_{\mathbb{R}} p(x) dx_j$ is locally bounded.

Autocovariance and correlation length

The independence of consecutive draws can be estimated from the sample itself. Suppose that we are interested in the convergence of the integral of $G(x)$ with respect to the probability density $p(x)$. Let us denote $z_j = G(x_j)$, where $\{x_1, \dots, x_N\} \subset \mathbb{R}^d$ is a MCMC sample and let $\bar{z} = N^{-1} \sum_{j=1}^N z_j$. Then we define the normalized autocovariance of the sample as

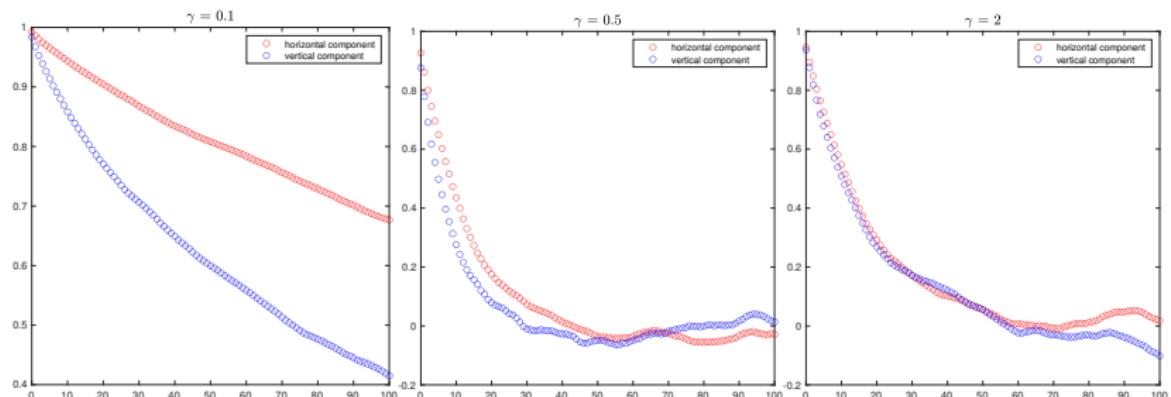
$$\gamma_k = \frac{1}{(N - k)\gamma_0} \sum_{j=1}^{N-k} (z_j - \bar{z})(z_{j+k} - \bar{z}), \quad k \geq 1,$$

where $\gamma_0 = N^{-1} \sum_{j=1}^N z_j^2$.

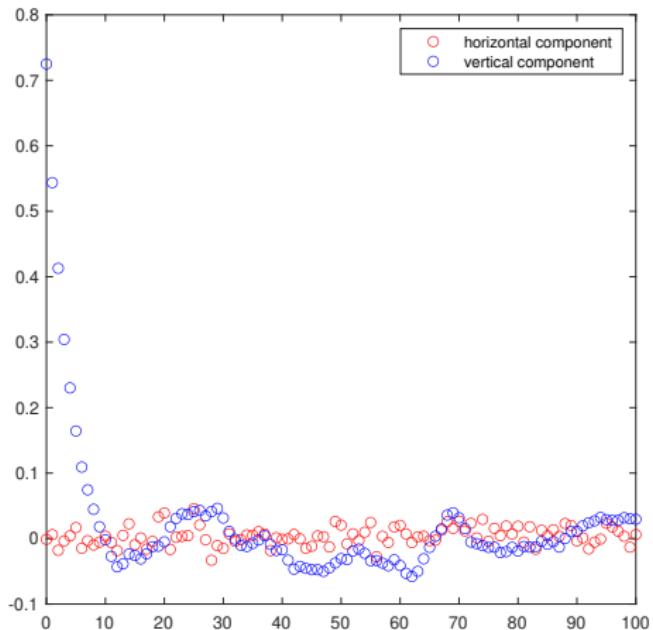
The correlation length of the sample $\{z_j\}_{j=1}^N$ can be estimated based on the decay of the normalized autocovariance sequence of the sample.

If every k^{th} sample point is independent, one might expect the discrepancy to behave as $1/\sqrt{N/k} = \sqrt{k/N}$ instead of $1/\sqrt{N}$. In consequence, one should try to choose the proposal distribution so that the *correlation length* is as small as possible.

Normalized autocovariance sequences for the Metropolis–Hastings example



Normalized autocovariance sequences for the Gibbs example



Preconditioned Crank–Nicolson algorithm

- The preconditioned Crank–Nicolson (pCN) algorithm is an instance of the Metropolis–Hastings algorithm with a specially chosen proposal density.
- The proposal is drawn using the model $Y = \sqrt{1 - \beta^2}X + \beta W$, where $W \sim \mathcal{N}(0, C_0)$, C_0 is a symmetric and positive definite matrix, with the (*non-symmetric!*) kernel

$$q(x, y) \propto \exp \left(-\frac{1}{2\beta^2} (y - \sqrt{1 - \beta^2}x)^T C_0^{-1} (y - \sqrt{1 - \beta^2}x) \right).$$

Here, the step size $0 < \beta < 1$ is a free parameter (which can be optimized for statistical efficiency).

- The pCN method is *dimension robust*: the acceptance probability does not degenerate to zero as the dimension $d \rightarrow \infty$. Contrast this with, e.g., random walk Metropolis, whose acceptance probability degenerates to zero as the dimension $d \rightarrow \infty$.

Further variations of MCMC

We have only scratched the surface of some basic ideas surrounding MCMC methods. In the literature and practical applications, one can find many variations of these ideas to boost the performance of MCMC for “difficult” / “high-dimensional” problems. To list a couple of notable ones:

- Adaptive Metropolis: as the proposal density $q(x, y)$, use a random walk model $Y = X + W$ with $W \sim \mathcal{N}(0, \Gamma)$, where the covariance Γ is replaced by the *sample covariance* (plus some small perturbation of identity) computed using the sample history. The updating can happen either at every step or after every M steps of the Metropolis iteration. The main theoretical challenge is proving the ergodicity of the chain—this was proved by Haario, Saksman, and Tamminen (2001). Computationally, stable updating formulae for the sample means and covariances are needed in practice.
- Independence Metropolis: as the proposal density $q(x, y)$, use a probability density that is independent of the previous sample x , i.e., $q(x, y) = q(y)$. The proposal density should be similar to the target density.
- Metropolis-within-Gibbs, Delayed rejection adaptive Metropolis, . . .

Software: <https://mjlaine.github.io/mcmcstat/>
<https://mc-stan.org/>