Introduction to Statistics – Exercises

Matthias REITZNER, Berhard HAFER, Germain POULLOT

April 2026 – July 2026

You do not need to do all the exercises! Please follow the instructions given during the Übung about which exercises to prepare for which: Usuually, you should follow the red stars ⋆ . Harder exercises are indicated with blue points • .

As I have a lot of material in French, and not that much time to re-write everything by hand, some exercises & corrections have been translated using ChatGPT: everything has been re-read, of course, but I hereby apologize for the small mistakes that might remain.

Contents

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 0

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Exercise 0 [⋆ ⋆ General Information]

If you have understood the following information, then you can go to the next exercises.

1. Lectures (Vorlesung) are Mondays and Wednesdays, 8am to 10am, room 69/125 (start 8:30).
2. Exercise sessions (Übungen) are Fridays, 8am to 10am, room 69/125.
3. Tutorium sessions (Tutorium) are Mondays, 10am to 12am, room 32/107 (to be confirmed).
4. The exercise sheets will be updated weekly, please think about downloading the new file on StuD.IP every so often. Solutions will be uploaded from time to time.
5. On each exercise sheet, there are far too many exercises. You can do all of them if you want (they are interesting), but you will need to prepare the ones with a red star ⋆ . Especially, prepare Exercises 2, 6, 8, 10 & 11 (of sheet 1) for Friday 17.04.2026; and prepare Exercises 12, 13, 15, 16 & 17 (of sheet 1) for Friday 24.04.2026; the next pack of to-be-prepared exercises will be announced during the next exercise sessions (Übungen). Yes, the numbers are random: the aim is to force you to have a quick look at the other exercises. If you feel there are too many exercises to prepare, or their writing is incomprehensible, please complain (politely) by sending an email to the teaching team (see 9.). If you don't understand a word, please remember Wikipedia exists and that you are now old enough to use it.
6. During the exercise sessions, some students will come to the board and present their solutions. Students will be chosen according to a certain (non-uniform) probability distribution. Berhard Hafer will check your solution and ask you some questions: exercise sessions are the opportunity to make mistakes and to learn stuffs! Each student must present several times during the semester. The first exercise session will take place on 17.04.2026.
7. The final examination will be an oral examination. The scheduling will be made precise later. To be allowed to take the exam, you need to prepare $\geq 50\%$ of the exercises each week, i.e. prepare at least between 2 and 3 exercises among the ones marked with ⋆ each week. Checking that you have indeed prepared $\geq 50\%$ is based mainly on trust, but we will conduct more serious checks if we realize you are trying to dodge the preparation of exercises. We are all doing statistics: running a precise test takes time and energy that we would rather spend in improving the course, but if we are forced to do it, we will find a way to test you precisely.
8. For the Tutorium, we will alternate between Tutorium sessions and learning sessions of the software R. We start on the 13.04.2026 by a Tutorium session, then a R session on 20.04, then a Tutorium session on 27.04, etc. This schedule is negotiable, depending on your needs.
9. You can send emails to the teaching team, please remain polite and try to be clear. Especially, please indicate which course you are talking about (each of us teach several courses at the same time, we tend to forget who is who). Putting several people in copy increases the probability of getting an answer. English, French or Spanish emails can be sent to Germain Poullot (who will not feel a strong need to answer if you write in German).
10. The use of ChatGPT (or any LLM or any generative AI system) is not forbidden, however, remember that you are graded via oral examinations: if you do not master what you are talking about, it will be very hard to fool the examiners... An oral examination is not a recitation of what you have written on a sheet of paper, and we will challenge you into showing your deeper understanding of the problem at stake. We are counting on you for making a clever use of these new tools, do not confuse gaining points and gaining knowledge!

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 1

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Exercise 1 [Bonferroni inequality]
Let $(\Omega, \mathbb{P})$ be a probability space, and let $A_1, \dots, A_n$ be events. Show that

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) \geq \left(\sum_{i=1}^{n} \mathbb{P}(A_i) \right) - (n-1).$$

Exercise 2 [⋆ Roots of random polynomials]
We roll a six-sided dice three times, and denote by W_1, W_2, and W_3 the successive outcomes obtained. We define

$$Q(x) = W_1 x^2 + W_2 x + W_3.$$

Determine the probability that:

- Q has two distinct real roots,
- Q has a double real root,
- Q has no real roots.

Exercise 3 [Birthday bet]
You are in a class of 30 students. Your math teacher wants to bet 10 euros with you that at least two people in the class share the same birthday. Do you accept the bet?

Exercise 4 [• Organizing a tournament]
To organize a tournament, a random draw is held involving n first-division basketball teams and n second-division teams, such that each team plays exactly one match.

- Compute the probability p_n that all matches are between a first-division team and a second-division team.
- Compute the probability q_n that all matches are between two teams from the same division.
- Show that for all $n \geq 1$,
$$\frac{2^{2n-1}}{n} \leq \binom{2n}{n} \leq 2^{2n}.$$
- Deduce $\lim_{n \to +\infty} p_n$ and $\lim_{n \to \infty} q_n$.

Exercise 5 [Find me a probability]
For $n \geq 1$, find a probability on $\{1, \ldots, n\}$ such that the probability of $\{1, \ldots, k\}$ is proportional to k^2 for every k. ("Find a probability" = compute $\mathbb{P}(\{k\})$ for all k.)

Exercise 6 [⋆ Probability of independent union]
Let $A_1, \ldots, A_n$ be events in a probability space $(\Omega, \mathbb{P})$. They are assumed to be mutually independent, with respective probabilities $p_i = \mathbb{P}(A_i)$.

Give a simple expression for

$$\mathbb{P}(A_1 \cup \cdots \cup A_n)$$

in terms of $p_1, \ldots, p_n$.

Application 1: Suppose a person is subjected to n independent experiments, and in each experiment they have a probability p of having an accident. What is the probability that they have at least one accident?

(**Application 2:** Re-do Exercise 3 on the birthday paradox.)

Exercise 7 [Impossible to be independent]
Suppose we have a probability space such that the sample space Ω is a finite set with cardinality equal to a prime number p, and that the model is uniform (all outcomes are equally likely).

Prove that two non-trivial events A and B (i.e., different from $\varnothing$ and Ω) cannot be independent.

Exercise 8 [⋆ Electrical circuits]
Let A, B, and C be three events. Show that:

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

We are given three electrical components C_1, C_2, and C_3, whose probabilities of functioning are p_i, and which operate independently of each other. Determine the probability that the circuit functions:

- if the components are arranged in series,
- if the components are arranged in parallel,
- if the circuit is mixed: C_1 is in series with the subcircuit formed by C_2 and C_3 in parallel.

Exercise 9 [• Independent events and size of the universe]
Let $(\Omega, \mathbb{P})$ be a finite sample space. Suppose that there exist n mutually independent events $A_1, \ldots, A_n$ such that $\mathbb{P}(A_k) \in (0, 1)$ for all $k = 1, \ldots, n$.

- Let $B_1, \ldots, B_n \in \mathcal{P}(\Omega)$ such that for all $k = 1, \ldots, n$, $B_k = A_k$ or $B_k = A_k^c$ (the complement of A_k). Show that
$$B_1 \cap \cdots \cap B_n \neq \varnothing.$$
- Let $B_1, \ldots, B_n, B_1', \ldots, B_n' \in \mathcal{P}(\Omega)$ such that for all $k = 1, \ldots, n$, $B_k, B_k' \in \{A_k, A_k^c\}$. Assume that $(B_1, \ldots, B_n) \neq (B_1', \ldots, B_n')$. Show that
$$B_1 \cap \cdots \cap B_n \quad \text{and} \quad B_1' \cap \cdots \cap B_n'$$
are disjoint.
- Deduce that
$$\mathrm{card}(\Omega) \geq 2^n.$$

Exercise 10 [⋆ • Euler totient function]

Let $n > 1$ be a fixed integer. We choose uniformly at random an integer x from $\{1, \dots, n\}$. For each integer $m \leq n$, let A_m denote the event "m divides x". Let B denote the event "x is coprime with n". Finally, let $p_1, \dots, p_r$ be the prime divisors of n.

- Express B in terms of the events A_{p_k}.
- For any natural number m that divides n, compute the probability of A_m.
- Show that the events $A_{p_1}, \dots, A_{p_r}$ are mutually independent.
- Deduce the probability of B.
- **Application:** Let $\phi(n)$ denote the number of integers between 1 and n that are coprime with n. Show that

$$\phi(n) = n \prod_{k=1}^{r} \left(1 - \frac{1}{p_k}\right).$$

Exercise 11 [⋆ Are you sick?]

You are the chief of staff for the Minister of Health. A disease is present in the population at a rate of one sick person per 10,000. A representative from a major pharmaceutical laboratory comes to promote their new screening test: if a person is sick, the test is positive 99% of the time; if a person is not sick, the test is positive 0.1% of the time.

Would you authorize the commercialization of this test?

Exercise 12 [⋆ Liar!]

You are playing heads or tails with another player. He bets on heads, flips the coin, and gets heads. What is the probability that he is a cheater?

Exercise 13 [⋆ • König-Huygens formula and Jensen inequality]

- For a random variable X, (re)prove König-Huygens formula, i.e. $\mathbb{V}X = \mathbb{E}X^2 - (\mathbb{E}X)^2$. Deduce that $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.
- More generally, let f be a convex function, i.e. for all $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$, we have

$$f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$

  Show Jensen inequality, that is (say, for a random variable taking finitely many values):

$$\mathbb{E}f(X) \geq f(\mathbb{E}X)$$

  Remember that if $f'' \geq 0$, then f is a convex function, and deduce again that $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

Exercise 14 [Expectation versus probability]

For a Bernoulli variable, i.e. a random variable X taking values in $\{0, 1\}$, show that $\mathbb{P}(X = 1) = \mathbb{E}X$.

Let X be a discrete and positive random variable. Show that $\mathbb{E}X = \sum_{k=1}^{+\infty} \mathbb{P}(X \geq k)$.

Let X be a continuous and positive random variable. Show that $\mathbb{E}X = \int_0^{+\infty} \mathbb{P}(X \geq t)\, \mathrm{d}t$.

Exercise 15 [⋆ • Maximum of random variables]

Let $X_1, \dots, X_n$ be independent and identically distributed random variables (continuous or discrete). Let $M = \max(X_1, \dots, X_n)$. Compute the distribution function of M (i.e. $F_M : t \mapsto \mathbb{P}(M \leq t)$).

Now, suppose that $X_1, \dots, X_n$ are independent and identically distributed random variables distributed uniformly on $[0, 1]$. Compute the k^{th} moments of $M = \max(X_1, \dots, X_n)$ and its variance. Study their limits (for k fixed, when $n \to +\infty$).

A friend said to me "if you want to simulate the maximum of two independent random variables uniformly distributed in $[0,1]$, don't sample twice: just sample once and take the square root.". Is he correct?

Exercise 16 [⋆ • Random variable without moments (Martingale & Cauchy)]

- We are playing heads and tails on a coin. If my first coin ends on head, I'll give you 1€, and we stop here. Otherwise, we continue playing: if my second coin ends on head, I'll give you 2€, and we stop here. Otherwise, we continue playing: if my third coin ends on head, I'll give you 4€, and we stop here. And so on... if we go to the k^{th} coin and it ends on head, then I'll give you 2^k€, and we stop here. How many euros do you expect to win? (i.e. how many euros should you pay me in advance to play this game?)
- The density function of the Cauchy distribution is:

$$f(x) = \frac{1}{\pi(1+x^2)}$$

  Check that this is a well-defined density function (remember that arctan exists).

  Compute the first, second, and third moments of a Cauchy distributed random variable. You should run into a problem.

(You can also have a look at Lévy distribution.)

Exercise 17 [⋆ Markov inequality]
Prove Markov inequality, that is, for a non-negative random variable X, we have $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}$ (*Hint*: introduce the random variable $Y = (X \geq a)$ which takes values in $\{0,1\}$).

Deduce that $\mathbb{P}\left(\frac{X}{\mathbb{E}X} \geq \frac{1}{a}\right) \leq a$ and that $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X^p}{a^p}$ for a non-negative random variable X and all $a \geq 0$, $p \geq 1$.

Deduce Chebychev inequality, that is for **all** X (non-negative or not) and $a \geq 0$:

$$\mathbb{P}\big(|X - \mathbb{E}X| \geq a\big) \leq \frac{\mathbb{V}X}{a^2}$$

(Look at exercise 21 on Cantelli inequality for a continuation.)

Exercise 18 [• Are you winning, son?]
You and an opponent play a (video) game: in each round, you have a probability p_1 of winning, and your opponent has a probability p_2 (independent of you). The first player to win a round wins the stake. What is the probability that you win? What is the probability of a tie?

What happens when the two players have the same level (i.e. $p_1 = p_2$)?

Exercise 19 [• Coupon Collector Problem]
You are collecting cards (e.g., Pokémon). You buy them one by one, but you do not know which card you are purchasing at the time of purchase (they are in sealed packs). There are n cards in total, and each purchase costs 1 euro. Show that, on average, completing your collection will cost $M_n \sim n \ln n$ euros. Show that the probability that the total cost deviates from the mean by more than εn is less than $2/\varepsilon^2$. Show that the probability of paying more than $\beta n \ln n$ euros (for $\beta > 0$) is less than $n^{-\beta+1}$.

Exercise 20 [• Hoeffding inequality]
We admit Hoeffding's lemma: if Y is a random variable bounded between $-M$ and M, then for all $t > 0$, we have

$$\mathbb{E}(e^{tY}) \leq e^{t^2 M^2/2}.$$

Let $(X_i)_i$ be independent random variables, each bounded between $-M$ and M (i.e., $\forall i$, $\mathbb{P}(-M \leq X_i \leq M) = 1$). Let

$$S_n = \sum_{k=1}^{n} X_k.$$

We assume for simplicity that $\forall n$, $\mathbb{E}(S_n) = 0$. By applying Markov's inequality to $\mathbb{P}(e^{tS_n} \geq e^{t\lambda})$, show that

$$\mathbb{P}(S_n \geq \lambda) \leq e^{-\lambda^2/(2nM^2)} \quad \text{for all } \lambda > 0.$$

Exercise 21 [• Cantelli inequality]

Let Y be a random variable with $\mathbb{E}(Y) = 0$ and $\mathrm{Var}(Y) = \sigma^2 < \infty$. Let λ and u be positive real numbers. Show that

$$\mathbb{P}(Y \geq \lambda) \leq \mathbb{P}\big((Y+u)^2 \geq (\lambda+u)^2\big).$$

Deduce that, for all $u > 0$,

$$\mathbb{P}(Y \geq \lambda) \leq \frac{\sigma^2 + u^2}{(\lambda+u)^2} \quad \text{(using Markov's inequality)}.$$

Prove that, for $\lambda > 0$:

$$\mathbb{P}(Y \geq \lambda) \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}.$$

Similarly, show that, for $\lambda < 0$:

$$\mathbb{P}(Y \leq \lambda) \geq 1 - \frac{\sigma^2}{\sigma^2 + \lambda^2}.$$

Exercise 22 [• Chernoff inequality]

Let X be a random variable and $t > 0$ such that $\mathbb{E}(e^{tX})$ is finite. Show that for all $a \geq 0$:

$$\mathbb{P}(X \geq a) \leq e^{-ta}\mathbb{E}(e^{tX}) \quad \text{and} \quad \mathbb{P}(X \leq -a) \leq e^{-ta}\mathbb{E}(e^{tX}).$$

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p$. Define $S = \sum_{i=1}^{n} X_i$. Show that, for $k \in \mathbb{R}$,

$$\mathbb{P}(S \geq knp) \leq \frac{1}{e^{np}} \left(\frac{e}{k}\right)^{knp}.$$

Exercise 23 [• The kurtosis is not linear]

Let X and Y be independent random variables with $\mathbb{E}X = \mathbb{E}Y = 0$.

Show that the three first moments are additive, i.e. that $\mathbb{E}(X+Y)^k = \mathbb{E}X^k + \mathbb{E}Y^k$ for $k \in \{1, 2, 3\}$.

Show that the fourth moment is not additive, i.e. that $\mathbb{E}(X+Y)^4 \neq \mathbb{E}X^4 + \mathbb{E}Y^4$ for well-chosen independent X and Y.

(The first moment is called the expectation, the second one yields the variance, the third one is called the skewness, and the fourth one is called the kurtosis.)

Introduction to Statistics – Solutions Sheet 1

Exercise 1 [Bonferroni inequality]

We proceed by induction on n, the key point being the formula

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

The property is true for $n = 1$. Let $n \geq 2$, and assume the property holds for $n-1$. We prove it for n.

Set $A = A_1 \cap \cdots \cap A_{n-1}$ and $B = A_n$. Then, by the previous formula,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B).$$

Now, we use the induction hypothesis to bound $\mathbb{P}(A)$ from below, and the fact that $\mathbb{P}(A \cup B) \leq 1$, which yields

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) \geq \sum_{i=1}^{n-1} \mathbb{P}(A_i) - (n-2) + \mathbb{P}(A_n) - 1,$$

which is exactly the desired result.

Exercise 2 [⋆ Roots of random polynomials]

We associate to the random experiment the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}^3$, with the uniform probability. Thus, the probability of an event A is

$$\mathbb{P}(A) = \frac{\text{card}(A)}{6^3}.$$

We first consider the event

$$A = \{(W_1, W_2, W_3) \in \Omega \mid W_2^2 - 4W_1W_3 > 0\}.$$

It suffices to count the elements of A. We start by establishing a small table of values of $4ac$:

$W_3 \backslash W_1$	1	2	3	4	5	6
1	4	8	12	16	20	24
2	8	16	24	32	40	48
3	12	24	36	48	60	72
4	16	32	48	64	80	96
5	20	40	60	80	100	120
6	24	48	72	96	120	144

We calculate the cardinality of A by counting the number of pairs (W_1, W_3) such that $W_2^2 > 4W_1W_3$ for each of the 6 possible values of W_2. We find:

$$\text{card}(A) = 0 + 0 + 3 + 5 + 14 + 16 = 38.$$

Hence,

$$\mathbb{P}(A) = \frac{38}{216} = \frac{19}{108}.$$

Similarly, let

$$B = \{(W_1, W_2, W_3) \in \Omega \mid W_2^2 - 4W_1W_3 = 0\}, \quad C = \{(W_1, W_2, W_3) \in \Omega \mid W_2^2 - 4W_1W_3 < 0\}.$$

Counting in the same way, we get:

$$\mathbb{P}(B) = \frac{5}{216}.$$

Finally, since A, B, and C form a complete system of events, we deduce:

$$\mathbb{P}(C) = 1 - \mathbb{P}(A) - \mathbb{P}(B) = \frac{173}{216}.$$

Exercise 3 [Birthday bet]

We want to find the probability that two people in a class share the same birthday. If this probability is greater than $\frac{1}{2}$, one should not accept the bet.

It is easier to calculate the probability of the complementary event. So, we first compute the probability that no two people share the same birthday, ignoring leap years for simplicity.

Formalizing the problem: for all 30 people to have different birthdays, we need to construct an injective map from $\{1, \ldots, 30\}$ to $\{1, \ldots, 365\}$. There are

$$365 \times 364 \times \cdots \times 336$$

such injective maps. The total number of functions from $\{1, \ldots, 30\}$ to $\{1, \ldots, 365\}$ is 365^{30}.

Hence, the probability that no two people share the same birthday is

$$\frac{365 \times 364 \times \cdots \times 336}{365^{30}} = \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{336}{365} \simeq 0.293.$$

Therefore, the probability that at least two people share a birthday is

$$1 - 0.293 \simeq 0.706.$$

Exercise 4 [• Organizing a tournament]

We count the draws while taking into account the order of the matches in the draw (i.e., if we have 4 teams A, B, C, D, the draws $(A-B, C-D)$ and $(C-D, A-B)$ are considered two different draws because the first match is different). One could do without this convention and obtain the same results, but the method would be slightly different.

A draw can thus be seen as an n-list of matches, i.e., an n-list of disjoint 2-combinations. There are

$$\binom{2n}{2}$$

ways to choose the first combination. Once this combination is chosen, there are

$$\binom{2n-2}{2}$$

ways to choose the next combination, and so on. Therefore, the total number of draws is

$$\binom{2n}{2} \times \binom{2n-2}{2} \times \cdots \times \binom{4}{2} \times \binom{2}{2} = \frac{(2n)!}{2^n}.$$

Among these draws, let us count those where only teams from different divisions face each other. For the first match, there are n ways to choose a first-division team and n ways to choose a second-division team, giving n^2 choices. For the second match, there remain $n-1$ teams in each division, giving $(n-1)^2$ choices, and so on. Finally, we obtain:

$$p_n = \frac{(n!)^2}{\frac{(2n)!}{2^n}} = \frac{2^n\,(n!)^2}{(2n)!}.$$

First, if n is odd, such a draw is clearly impossible, so $q_n = 0$. Assume n is even and write $n = 2k$. We first choose the k matches among the $2k$ first-division teams to play each other; this gives $\binom{2k}{k}$

choices. Once this choice is made, we count the number of arrangements among the first-division teams, which is $(2k)!/2^k$, and similarly for the second-division teams. Thus,

$$q_{2k} = \frac{\binom{2k}{k}\left(\frac{(2k)!}{2^k}\right)^2}{\frac{(4k)!}{2^{2k}}} = \frac{\binom{2k}{k}}{\binom{4k}{2k}}.$$

We can write

$$(2n)! = 2^n(2n-1)(2n-3)\cdots 3\cdot 1\cdot n!.$$

Hence,

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2} = 2^n\frac{(2n-1)(2n-3)\cdots 3\cdot 1}{n!}.$$

Using the bounds

$$2n-k-1 \leq 2n-k \leq 2n-k+1,$$

we obtain

$$2^n\frac{(2n-2)(2n-4)\cdots 4\cdot 2}{n!} \leq \binom{2n}{n} \leq 2^n\frac{2n(2n-2)\cdots 4\cdot 2}{n!},$$

which gives the requested result.

We thus have

$$0 \leq p_n \leq \frac{n}{2^{n-1}},$$

which shows that $p_n \to 0$ as $n \to +\infty$. Similarly for q_n.

Exercise 5 [Find me a probability]

The statement tells us that we are looking for a probability P such that

$$P(\{1,\ldots,k\}) = \lambda k^2.$$

Then, for $k = 1,\ldots,n$,

$$P(\{k\}) = P(\{1,\ldots,k\}) - P(\{1,\ldots,k-1\}) = \lambda k^2 - \lambda(k-1)^2 = \lambda(2k-1).$$

We determine λ by noting that

$$P(\{1,\ldots,n\}) = 1,$$

which gives

$$\lambda n^2 = 1 \quad \Longrightarrow \quad \lambda = \frac{1}{n^2}.$$

Hence, the probability is defined by

$$P(\{k\}) = \frac{2k-1}{n^2}.$$

It is easily verified that, conversely, this probability satisfies $P(\{1,\ldots,k\}) \propto k^2$.

Exercise 6 [⋆ Probability of independent union]

If the events $A_1,\ldots,A_n$ are mutually independent, then their complements $\overline{A_1},\ldots,\overline{A_n}$ are also independent. We deduce that

$$\mathbb{P}(A_1\cup\cdots\cup A_n) = 1 - \mathbb{P}(\overline{A_1}\cap\cdots\cap\overline{A_n}) = 1 - \prod_{i=1}^n \mathbb{P}(\overline{A_i}) = 1 - \prod_{i=1}^n (1-p_i).$$

In the particular case given, the probability of having at least one accident is therefore

$$1-(1-p)^n.$$

Exercise 7 [Impossible to be independent]
Suppose there exist two nontrivial events A and B that are independent. Let m be the cardinality of A and n the cardinality of B. Then

$$\mathbb{P}(A) = \frac{m}{p} \quad \text{and} \quad \mathbb{P}(B) = \frac{n}{p}.$$

Since A and B are assumed independent, and using the uniform probability model, we have

$$\operatorname{card}(A \cap B) = p\,\mathbb{P}(A \cap B) = p\,\mathbb{P}(A)\mathbb{P}(B) = \frac{mn}{p}.$$

Since the cardinality must be an integer, p divides the product mn, and by Gauss's theorem, it divides one of the factors, say n. Moreover, since $n \leq p$, this is possible only if $n = 0$ or $n = p$.

In other words, B must be either the certain event or the impossible event, i.e., a trivial event.

Exercise 8 [⋆ Electrical circuits]
We proceed in two steps. First, we have

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

But

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

and

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

This is also called the Poincaré sieve formula, which generalizes to multiple events by induction.

Let F_i denote the event: "component C_i works". By hypothesis, the events F_i are mutually independent. We need to compute:

1. $\mathbb{P}(F_1 \cap F_2 \cap F_3)$ — the circuit formed by three components in series works if and only if all three components work. 2. $\mathbb{P}(F_1 \cup F_2 \cup F_3)$ — the circuit formed by three components in parallel works if and only if at least one component works. 3. $\mathbb{P}(F_1 \cap (F_2 \cup F_3))$ — a mixed configuration where F_1 is in series with the parallel sub-circuit of F_2 and F_3.

By independence of events:

$$\mathbb{P}(F_1 \cap F_2 \cap F_3) = p_1 p_2 p_3.$$

Using the formula above and independence:

$$\mathbb{P}(F_1 \cup F_2 \cup F_3) = p_1 + p_2 + p_3 - p_1p_2 - p_1p_3 - p_2p_3 + p_1p_2p_3.$$

The event $F_2 \cup F_3$ is independent of F_1. Hence,

$$\mathbb{P}(F_1 \cap (F_2 \cup F_3)) = \mathbb{P}(F_1)\,\mathbb{P}(F_2 \cup F_3) = p_1\,(p_2 + p_3 - p_2p_3).$$

Exercise 9 [• Independent events and size of the universe]
Since the events $B_1, \ldots, B_n$ are independent, we have

$$\mathbb{P}(B_1 \cap \cdots \cap B_n) = \prod_{i=1}^{n} \mathbb{P}(B_i) > 0$$

because $\mathbb{P}(B_i) \in (0, 1)$. Thus,

$$B_1 \cap \cdots \cap B_n \neq \emptyset.$$

Let k be such that $B_k \neq B'_k$. Then necessarily,

$$B_k \cap B'_k = \emptyset,$$

and therefore

$$(B_1 \cap \cdots \cap B_n) \cap (B_1' \cap \cdots \cap B_n') = \emptyset.$$

For each n-tuple $B = (B_1, \ldots, B_n)$, there exists an element $x_B \in B_1 \cap \cdots \cap B_n$. These elements x_B are pairwise distinct according to the previous argument, and since there are 2^n such n-tuples, the cardinality of Ω satisfies

$$\mathrm{card}(\Omega) \geq 2^n.$$

Exercise 10 [⋆ • Euler totient function]

We know that x is coprime with n if and only if none of the prime divisors of n divide x. Hence,

$$B = \overline{A_{p_1}} \cap \cdots \cap \overline{A_{p_r}}.$$

Since we are in the equiprobable situation, it suffices to compute the cardinality of A_m. If $n = k \cdot m$, then the multiples of m less than or equal to n are $m, 2m, \ldots, km$. Therefore,

$$\mathbb{P}(A_m) = \frac{k}{n} = \frac{1}{m}.$$

Let $i_1 < \cdots < i_m$ be distinct integers chosen from $\{1, \ldots, r\}$. We need to show that

$$\mathbb{P}(A_{p_{i_1}}) \cdots \mathbb{P}(A_{p_{i_m}}) = \mathbb{P}(A_{p_{i_1}} \cap \cdots \cap A_{p_{i_m}}).$$

But

$$\mathbb{P}(A_{p_{i_1}}) \cdots \mathbb{P}(A_{p_{i_m}}) = \prod_{j=1}^{m} \frac{1}{p_{i_j}}.$$

On the other hand, since $p_{i_1}, \ldots, p_{i_m}$ are pairwise coprime, an integer is a multiple of $p_{i_1} \ldots p_{i_m}$ if and only if it is divisible by each p_{i_j} for $j = 1, \ldots, m$. We deduce that

$$A_{p_{i_1}} \cap \cdots \cap A_{p_{i_m}} = A_{p_{i_1} \ldots p_{i_m}},$$

so that

$$\mathbb{P}(A_{p_{i_1}} \cap \cdots \cap A_{p_{i_m}}) = \frac{1}{p_{i_1} \ldots p_{i_m}},$$

which proves the desired result.

The events $\overline{A_{p_1}}, \ldots, \overline{A_{p_r}}$ are also independent. We deduce that

$$\mathbb{P}(B) = \prod_{j=1}^{r} \mathbb{P}(\overline{A_{p_j}}) = \prod_{j=1}^{r} \left(1 - \frac{1}{p_j}\right).$$

Finally, by the equiprobable model, and since the cardinality of B equals $\varphi(n)$, we have

$$\mathbb{P}(B) = \frac{\varphi(n)}{n},$$

which, together with the previous question, gives the desired result:

$$\varphi(n) = n \prod_{j=1}^{r} \left(1 - \frac{1}{p_j}\right).$$

Exercise 11 [⋆ Are you sick?]

The numbers given may look excellent, but they actually give the opposite of what we want. The problem is rather the following: if a person tests positive, are they actually sick? Bayes' formula allows us to trace this probability.

Specifically, let M be the event "the person is sick" and T be the event "the test is positive". The data we have are

$$\mathbb{P}(M) = 10^{-4}, \quad \mathbb{P}(T \mid M) = 0.99, \quad \mathbb{P}(T \mid \overline{M}) = 0.001.$$

We want to compute $\mathbb{P}(M \mid T)$. Bayes' formula gives:

$$\mathbb{P}(M \mid T) = \frac{\mathbb{P}(T \mid M)\,\mathbb{P}(M)}{\mathbb{P}(T \mid M)\,\mathbb{P}(M) + \mathbb{P}(T \mid \overline{M})\,\mathbb{P}(\overline{M})} = \frac{10^{-4} \times 0.99}{10^{-4} \times 0.99 + 10^{-3} \times 0.9999} \simeq 0.09.$$

This is disastrous! The probability that a person who tests positive is actually sick is less than 10%. The test therefore produces a lot of false positives (people who test positive but are not sick). This is exactly the problem with relatively rare diseases: screening tests must be extremely reliable.

Moreover, this is a good illustration of the old statisticians' saying: one can make numbers say anything, cf. the pharmaceutical laboratory!

Exercise 12 [⋆ Liar!]

Let x be the proportion of cheaters in the population. Let P, F, H, T denote the events "the player gets heads", "the player gets tails", "the player is honest", and "the player is a cheater", respectively.

It seems reasonable to assume

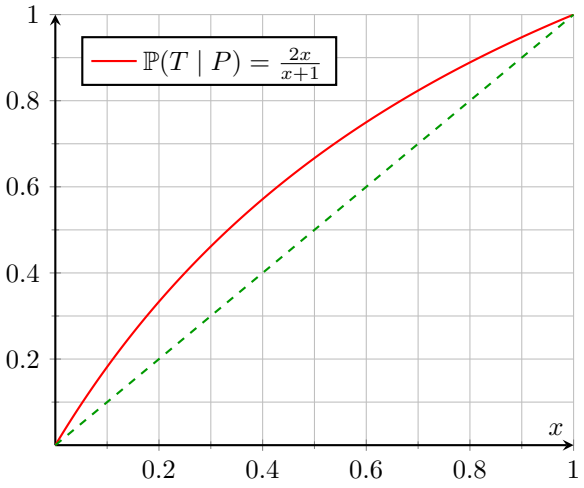
$$\mathbb{P}(P \mid H) = \frac{1}{2}, \quad \mathbb{P}(F \mid H) = \frac{1}{2}, \quad \mathbb{P}(P \mid T) = 1$$

(a cheater truly gets whatever they want!).

We want to compute $\mathbb{P}(T \mid P)$. By Bayes' formula, we have:

$$\mathbb{P}(T \mid P) = \frac{\mathbb{P}(P \mid T)\,\mathbb{P}(T)}{\mathbb{P}(P \mid T)\,\mathbb{P}(T) + \mathbb{P}(P \mid H)\,\mathbb{P}(H)} = \frac{x}{x + \frac{1}{2}(1-x)} = \frac{2x}{x+1}.$$

The result is more or less reassuring depending on the proportion of cheaters x in the population! Note that $\mathbb{P}(T \mid P) \geq \mathbb{P}(T)$. Indeed, seeing someone winning should increase your belief that he/she is a cheater.



Exercise 13 [⋆ • König-Huygens formula and Jensen inequality]

- $\mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}\big(X^2 - 2X(\mathbb{E}X) + (\mathbb{E}X)^2\big) = \mathbb{E}X^2 - 2(\mathbb{E}X)(\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ by linearity of the expectation.

  As $\mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2$ is the expectation of a **positive** variable, we have $\mathbb{V}X \geq 0$, hence $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

- Suppose X takes finitely many values, say $\mathbb{P}(X \in A) = 1$ for a finite set $A = \{a_1, \dots, a_m\} \subseteq \mathbb{R}$, and let $p_k = \mathbb{P}(X = a_k)$. Recall that $\sum_{k\in A} p_k = 1$ Then we have, the middle inequality is just the general form of the definition of a convex function:

$$f(\mathbb{E}X) = f\left(\sum_{k\in A} p_k a_k\right) \geq \sum_{k\in A} p_k f(a_k) = \mathbb{E}f(X)$$

  The function $f : t \mapsto t^2$ is convex because its second derivative satisfies $f''(t) = 2 \geq 0$ for all t. Hence, by Jensen inequality: $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

Exercise 14 [Expectation versus probability]

If X is a Bernoulli variable, then:

$$\mathbb{E}X = 0 \cdot \mathbb{P}(X = 0) + 1 \cdot \mathbb{P}(X = 1) = \mathbb{P}(X = 1)$$

This sounds easy and stupid, but it has a lot of applications!
Besides, for a discrete variable:

$$\mathbb{E}X = \sum_{k=0}^{+\infty} k\mathbb{P}(X = k) = \sum_{k=0}^{+\infty} \sum_{1\leq j\leq k} \mathbb{P}(X = k) = \sum_{j=1}^{+\infty} \sum_{k\geq j} \mathbb{P}(X = k) = \sum_{j=0}^{+\infty} \mathbb{P}(X \geq j)$$

For a continuous and positive random variable, the reasoning is the same.

Exercise 15 [⋆ • Maximum of random variables]

As M is a maximum, the event $(M \leq t)$ is equivalent to $\big((X_1 \leq t) \text{ and } \dots \text{ and } (X_n \leq t)\big)$. As the variables are independent, we get: $\mathbb{P}(M \leq t) = \mathbb{P}\big((X_1 \leq t) \cap \dots \cap (X_n \leq t)\big) = \mathbb{P}(X_1 \leq t) * \dots * \mathbb{P}(X_n \leq t) = \prod_i \mathbb{P}(X_i \leq t)$.

As all the variables X_i follow the same distribution, we get $\mathbb{P}(M \leq t) = \mathbb{P}(X_1 \leq t)^n$, so $F_M = F_X^n$, where F_X is the distribution function of these variables.

If the X_i are uniformly distributed $[0, 1]$, then $F_X(t) = t$ for $t \in [0, 1]$ (and $F_X(t) = 0$ for $t \leq 0$, and $F_X(t) = 1$ for $t \geq 1$). Thus, $F_M(t) = t^n$ and the density function is $f_M(t) = F'_M(t) = n \cdot t^{n-1}$. We get in this case:

$$\mathbb{E}M^k = \int_0^1 t^k \cdot n\, t^{n-1}\, \mathrm{d}t = n \int_0^1 t^{n+k-1}\, \mathrm{d}t = \frac{n}{n+k} \xrightarrow[n\to+\infty]{k \text{ fixed}} 1$$

In this case, the variance is $\mathbb{V}M = \mathbb{E}M^2 - (\mathbb{E}M)^2 = \frac{n}{n+2} - \frac{n^2}{(n+1)} = \frac{n}{(n+1)^2(n+2)} \to 0$ when $n \to +\infty$.

To conclude, your friend is correct: If you want to sample the maximum of two variables (i.i.d. and uniform in $[0, 1]$), then $\mathbb{P}(M \leq t) = \mathbb{P}(X \leq t)^2 = t^2$. If you want to sample the square root of one variable (uniform in $[0, 1]$), then $\mathbb{P}(\sqrt{X} \leq t) = \mathbb{P}(X \leq t^2) = t^2$. As both variables M and $\sqrt{X}$ have the same distribution function, you'll get the same samples. (As n^{th} roots are harder to compute than maxima, one may use samples of maxima to mimic samples of n^{th} roots.)

Exercise 16 [⋆ • Random variable without moments (Martingale & Cauchy)]

- The probability that you've not won after $k-1$ attempt is $(1/2)^{k-1}$, so the probability that you win 2^k€ is $(1/2)^{k-1} \cdot (1/2) = 1/2^k$. Hence, the expectation is $\sum_k 2^k \cdot 1/2^k = \sum_k 1 = +\infty$€. You should pay me in advance an infinite amount of money to play this game!

- There are so many problems with the moments of the Cauchy distribution. First of all, remember that $\arctan' x = \frac{1}{1+x^2}$, so $\int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{1+x^2} = \arctan(+\infty) - \arctan(-\infty) = \frac{\pi}{2} + \frac{\pi}{2} = \pi$, so $\int_{-\infty}^{+\infty} f(x)\, \mathrm{d}x = 1$, and f is a well defined density function (it is continuous).

However, the k^{th} moment is given by $\frac{1}{\pi}\int_{-\infty}^{+\infty}\frac{x^k\,\mathrm{d}x}{1+x^2}$.

For even moments, note that $x\mapsto\frac{x^{2k}}{1+x^2}$ is a positive function which does not converge to 0 in $\pm\infty$, hence $\mathbb{E}X^{2k}=+\infty$.

For odd moments, the case is worse. The function $x\mapsto\frac{x^{2k+1}}{1+x^2}$ is odd, positive for $x\geq 0$, negative for $x\leq 0$. The integral $\int_0^{+\infty}\frac{x^{2k+1}\,\mathrm{d}x}{1+x^2}$ is infinite (for $2k+1=1$, a primitive is $\frac{1}{2}\ln(1+x^2)$ which tends to $+\infty$ when $x\to+\infty$; while for $2k+1\geq 3$, we have $\frac{x^{2k+1}}{1+x^2}\to+\infty$ when $x\to+\infty$). So $\int_{-\infty}^{+\infty}\frac{x^{2k+1}\,\mathrm{d}x}{1+x^2}$ does not have a well-defined value, e.g. $\lim_{a\to+\infty}\int_{-a}^{+a}\frac{x^1\,\mathrm{d}x}{1+x^2}=0$ and $\lim_{a\to+\infty}\int_{-a}^{+2a}\frac{x^1\,\mathrm{d}x}{1+x^2}=\ln 2\neq 0$ would not yield the same number.

The Cauchy distribution has **infinite even moments**, and **undefined odd moments**.

Exercise 17 [⋆ Markov inequality]

Consider that random variable $Y=\mathbb{1}(X\geq a)$. If $X\geq a$, then $Y=1$, implying in particular $X\geq aY$. If $X\leq a$, then $Y=0$, implying in particular $X\geq aY$.

As the expectancy is an increasing function, we get $\mathbb{E}X\geq\mathbb{E}(aY)=a\mathbb{E}Y=a\mathbb{P}(X\geq a)$, which is Markov inequality.

Apply Markov inequality to the random variable $X/\mathbb{E}X$ with parameter $1/a$ to get $\mathbb{P}\left(\frac{X}{\mathbb{E}X}\geq a\right)\leq a$.

Note that $x\mapsto x^p$ is an increasing function and that $Y^p=Y$ because it is a 0/1 variable, to get that $a^p\mathbb{P}(X\geq a)=a^p\mathbb{E}Y=a^p\mathbb{E}Y^p\leq\mathbb{E}X^p$.

For Chebyshev inequality, recall that $|x|\geq a$ if and only if $x^2\geq a^2$ for all $x\in\mathbb{R}$ (positive or not), hence: $\mathbb{P}\big(|X-\mathbb{E}X|\geq a\big)=\mathbb{P}\big((X-\mathbb{E}X)^2\geq a^2\big)\leq\frac{1}{a^2}\mathbb{E}(X-\mathbb{E}X)^2=\frac{\mathbb{V}X}{a^2}$.

Exercise 18 [• Are you winning, son?]

Let $q_1=1-p_1$ and $q_2=1-p_2$. Let X_1 be the random variable representing the number of rounds you need to play before winning, and X_2 the corresponding variable for your opponent. Then X_1 follows a geometric distribution $\mathcal{G}(p_1)$ and X_2 follows a geometric distribution $\mathcal{G}(p_2)$. The two variables are independent.

Recall that if $X\sim\mathcal{G}(p)$, then denoting $q=1-p$ we get

$$\mathbb{P}(X\geq k)=q^{k-1}.$$

This can be seen in two ways: either by noting that winning after k rounds is equivalent to failing in the first $k-1$ rounds, or by computing

$$\mathbb{P}(X\geq k)=\sum_{j=k}^{+\infty}\mathbb{P}(X=j)=\sum_{j=k}^{+\infty}pq^{j-1}=pq^{k-1}\sum_{j=0}^{+\infty}q^j=\frac{pq^{k-1}}{1-q}=q^{k-1}.$$

Thus, winning the game amounts to requiring fewer attempts than the opponent, i.e., $X_1<X_2$. By independence of X_1 and X_2, we have:

$$\mathbb{P}(X_1<X_2)=\sum_{k=1}^{+\infty}\mathbb{P}(X_1=k)\mathbb{P}(X_2\geq k+1)=\sum_{k=1}^{+\infty}p_1q_1^{k-1}q_2^k=p_1q_2\sum_{j=0}^{+\infty}(q_1q_2)^j=\frac{p_1q_2}{1-q_1q_2}.$$

There are two ways to compute the probability of a tie. Either directly:

$$\mathbb{P}(X_1=X_2)=\sum_{k=1}^{+\infty}\mathbb{P}(X_1=k)\mathbb{P}(X_2=k)=\sum_{k=1}^{+\infty}p_1q_1^{k-1}p_2q_2^{k-1}=p_1p_2\sum_{k=0}^{+\infty}(q_1q_2)^k=\frac{p_1p_2}{1-q_1q_2},$$

or by using total probability:

$$\mathbb{P}(X_1=X_2)=1-\mathbb{P}(X_1<X_2)-\mathbb{P}(X_1>X_2)=1-\frac{p_1q_2}{1-q_1q_2}-\frac{p_2q_1}{1-q_1q_2}=\frac{p_1p_2}{1-q_1q_2}.$$

If both players have the same skill level, let $p = p_1 = p_2$. Then the probability of a tie $E(p)$ becomes

$$E(p) = \frac{p^2}{1-q^2} = \frac{p^2}{p(2-p)} = \frac{p}{2-p}.$$

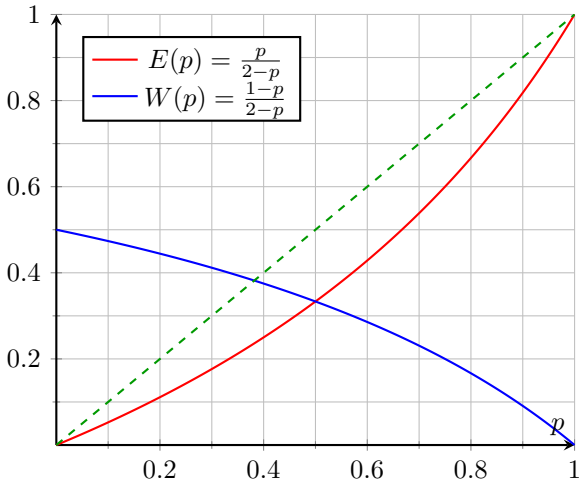
Since the probability that player 1 wins is the same as the probability that player 2 wins, denote this probability by $W(p)$. The total probability (player 1 wins + player 2 wins + tie) equals 1:

$$W(p) + W(p) + E(p) = 1.$$

Hence,

$$W(p) = \frac{1}{2}\left(1 - \frac{p}{2-p}\right) = \frac{1-p}{2-p}.$$

The game is highly balanced when both players are equally weak ($W(p) \approx 1/2$ for $p \approx 0$), and very prone to ties when both players are equally strong ($E(p) \to 1$ as $p \to 1$). This last statement is worth reading several times to fully grasp the difference between the two situations! When $p = 1/2$, then we have an equi-probability between a tie, first player wins, and second player wins (i.e. $E(1/2) = W(1/2) = 1/3$).



Exercise 19 [• Coupon Collector Problem]
Let T_n denote the "time", that is, the number of blister packs one needs to buy before obtaining all n cards in the collection.

Suppose we have already obtained $i-1$ distinct cards. Let t_i be the additional time needed to obtain the i-th new card. Then

$$T_n = \sum_{i=1}^{n} t_i.$$

The random variable t_i follows a geometric law of parameter $p_i = \frac{n-(i-1)}{n}$. Indeed, t_i is the waiting time until the first success, where the probability of success is p_i, namely the probability of drawing one of the $n-(i-1)$ cards not yet collected among the n possible cards.

Thus,

$$\mathbb{E}(t_i) = \frac{1}{p_i} = \frac{n}{n-i+1}.$$

Since $T_n = \sum_{i=1}^{n} t_i$, we obtain

$$\mathbb{E}(T_n) = \sum_{i=1}^{n} \mathbb{E}(t_i) = n \sum_{i=1}^{n} \frac{1}{n-i+1} = nH_n,$$

where $H_n = \sum_{k=1}^{n} \frac{1}{k}$ is the n-th harmonic number. It is well known that $H_n \sim \ln n$. Therefore,

$$M_n = \mathbb{E}(T_n) = nH_n \sim n \ln n.$$

Next, the second part amounts to proving that

$$\mathbb{P}(|T_n - M_n| \geq \varepsilon n) \leq \frac{2}{\varepsilon^2}.$$

Since $M_n = \mathbb{E}(T_n)$, this is an instance of Chebyshev's inequality. It remains to compute $\mathrm{Var}(T_n)$. As the t_i are independent (they are waiting times depending only on how many cards have already been collected), we have

$$\mathrm{Var}(T_n) = \sum_{i=1}^{n} \mathrm{Var}(t_i).$$

For a geometric random variable with parameter p_i, we have

$$\mathrm{Var}(t_i) = \frac{1 - p_i}{p_i^2} \leq \frac{1}{p_i^2}.$$

Thus,

$$\mathrm{Var}(T_n) \leq \sum_{i=1}^{n} \frac{1}{p_i^2} = n^2 \sum_{i=1}^{n} \frac{1}{(n - i + 1)^2}.$$

Since $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ and all terms are positive,

$$\mathrm{Var}(T_n) \leq n^2 \frac{\pi^2}{6} \leq 2n^2.$$

Applying Chebyshev's inequality,

$$\mathbb{P}(|T_n - nH_n| \geq \varepsilon n) \leq \frac{\mathrm{Var}(T_n)}{(\varepsilon n)^2} \leq \frac{\pi^2}{6\varepsilon^2} \leq \frac{2}{\varepsilon^2}.$$

Finally, we want to prove that

$$\mathbb{P}(T_n > \beta n \ln n) \leq n^{-\beta+1}.$$

To do so, consider the event Z_i^r that the i-th card has not been obtained by time r, i.e. $Z_i^r = (t_i > r)$. Clearly, $t_i > r$ implies $T_n > r$. Hence,

$$(T_n > r) \subseteq \bigcup_{i=1}^{n} (t_i > r),$$

so by the union bound,

$$\mathbb{P}(T_n > r) \leq \sum_{i=1}^{n} \mathbb{P}(t_i > r).$$

It remains to estimate $\mathbb{P}(t_i > r)$ for $r = \beta n \ln n$. Not having obtained the i-th card by time r means having failed r times. The failure probability is at most $\frac{n-1}{n}$ (since at least one card is still missing), hence

$$\mathbb{P}(t_i > r) \leq \left(\frac{n-1}{n}\right)^r \leq e^{-r/n}.$$

Since this bound does not depend on i, we obtain

$$\mathbb{P}(T_n > \beta n \ln n) \leq \sum_{i=1}^{n} \mathbb{P}(t_i > \beta n \ln n) \leq n e^{-\beta \ln n} = n^{-\beta+1}.$$

Exercise 20 [• Hoeffding inequality]
Since the exponential function is increasing, the inequalities $tS_n \geq \lambda t$ and $e^{tS_n} \geq e^{\lambda t}$ are equivalent. It follows that we can apply Markov's inequality, since the random variable e^{tS_n} is nonnegative:

$$\mathbb{P}(S_n \geq \lambda) = \mathbb{P}(e^{tS_n} \geq e^{\lambda t}) \leq \frac{1}{e^{\lambda t}}\mathbb{E}(e^{tS_n}) \leq e^{-\lambda t}\,\mathbb{E}\left(\prod_{k=1}^{n} e^{tX_k}\right) \leq e^{-\lambda t}\prod_{k=1}^{n}\mathbb{E}(e^{tX_k}) \quad \text{(by independence)}$$

$$\leq e^{-\lambda t}\left(e^{\frac{t^2M^2}{2}}\right)^n \quad \text{(by Hoeffding's lemma)} \leq \exp\left(-\lambda t + \frac{nM^2}{2}t^2\right).$$

Since this inequality holds for all $t > 0$, we can optimize over t to obtain the best bound. The optimal choice is $t = \frac{\lambda}{nM^2} > 0$, which yields, for all $\lambda > 0$,

$$\mathbb{P}(S_n \geq \lambda) \leq \exp\left(-\frac{\lambda^2}{2nM^2}\right).$$

Remark. This inequality holds in a more general setting: if $\mathbb{P}(X_i \in [a_i, b_i]) = 1$, then

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geq \lambda) \leq \exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

and

$$\mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq \lambda) \leq 2\exp\left(-\frac{2\lambda^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

This inequality is much, much stronger than Markov's or Chebyshev's inequalities. One of the reasons is that it is often advantageous to decompose random variables into sums of simpler components.

Exercise 21 [• Cantelli inequality]
Since λ and u are positive, we observe that $(Y+u)^2 \geq (\lambda+u)^2$ is implied by $Y + u \geq \lambda + u$. Hence, the probability of the second event is smaller than that of the first. Thus (recall that $\mathbb{E}(Y) = 0$):

$$\mathbb{P}(Y \geq \lambda) = \mathbb{P}(Y + u \geq \lambda + u) \leq \mathbb{P}\big((Y+u)^2 \geq (\lambda+u)^2\big) \leq \frac{\mathbb{E}\big[(Y+u)^2\big]}{(\lambda+u)^2}$$

$$\leq \frac{\mathbb{E}(Y^2) + 2u\mathbb{E}(Y) + u^2}{(\lambda+u)^2} = \frac{\sigma^2 + u^2}{(\lambda+u)^2}.$$

This inequality holds for all $u > 0$, so we can optimize by differentiating

$$f(u) = \frac{\sigma^2 + u^2}{(\lambda+u)^2}.$$

A computation gives

$$(\lambda+u)^4 f'(u) = 2u(\lambda+u)^2 - (2u+2\lambda)(\sigma^2+u^2) = 2\lambda u^2 + 2(\lambda - \sigma^2)u - 2\lambda\sigma^2.$$

In particular, $f'\left(\frac{\sigma^2}{\lambda}\right) = 0$. Taking $u = \frac{\sigma^2}{\lambda}$, we obtain

$$\mathbb{P}(Y \geq \lambda) \leq \frac{\lambda^2\sigma^2 + \sigma^4}{\lambda^2} \Big/ \left(\frac{\lambda^2 + \sigma^2}{\lambda}\right)^2 = \frac{\sigma^2}{\lambda^2 + \sigma^2}.$$

This proves Cantelli's inequality for $\lambda > 0$.

Next, note that

$$\mathbb{P}(Y \leq \lambda) = 1 - \mathbb{P}(Y \geq \lambda) = 1 - \mathbb{P}(-Y \leq -\lambda).$$

Thus, if $\lambda < 0$, since $-Y$ has the same expectation and variance as Y, we obtain

$$\mathbb{P}(Y \leq \lambda) \geq 1 - \frac{\sigma^2}{\lambda^2 + \sigma^2}.$$

Remark. For a random variable X with finite (not necessarily zero) expectation and variance σ^2, one can replace Y by $X - \mathbb{E}(X)$ everywhere. Combining the two inequalities yields

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) \leq \frac{2\sigma^2}{\lambda^2 + \sigma^2}.$$

This is slightly weaker than Chebyshev's inequality (asymptotically), but useful in some cases. In particular, Cantelli's inequalities work without absolute values (just like Hoeffding's or Chernoff's inequalities!).

Exercise 22 [• Chernoff inequality]
We observe that $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta})$ for all $t > 0$ (since the function $x \mapsto e^{tx}$ is increasing). Hence, since e^{tX} is a nonnegative random variable, we can apply Markov's inequality:

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}} = e^{-ta}\,\mathbb{E}(e^{tX}).$$

This inequality holds for all $t > 0$ such that $\mathbb{E}(e^{tX})$ is finite.
For $S = \sum_i X_i$, we have $\mathbb{E}(S) = \sum_i \mathbb{E}(X_i) = np$ and, in particular,

$$\mathbb{E}(e^{tS}) = \mathbb{E}\Big(\prod_i e^{tX_i}\Big) = \prod_i \mathbb{E}(e^{tX_i}).$$

Now, $e^{tX_i} \in \{1, e^t\}$ since $|X_i| \in \{0, 1\}$, so $\mathbb{E}(e^{tS})$ is finite (as a product of finite expectations). In fact,

$$\mathbb{E}(e^{tX_i}) = (1-p)e^0 + pe^t = 1 + p(e^t - 1) \leq e^{p(e^t-1)}.$$

It follows that

$$\mathbb{P}(S \geq a) \leq e^{-ta}\mathbb{E}(e^{tS}) \leq e^{-ta}\prod_i e^{p(e^t-1)} \leq e^{-ta}e^{np(e^t-1)} = e^{-ta+np(e^t-1)}.$$

In particular, for $a = k\,\mathbb{E}(S) = knp$, we have

$$\mathbb{P}(S \geq knp) \leq e^{-knpt+np(e^t-1)} = e^{np(e^t-kt-1)}.$$

Since this formula holds for all $t > 0$, we can choose t such that $e^t = k$ (this is the zero of the derivative of $t \mapsto e^t - kt - 1$), giving

$$e^t - kt - 1 = k(1 - \ln k) - 1.$$

Finally, we obtain

$$\mathbb{P}(S \geq knp) \leq e^{(k-1)np}/k^{knp}$$

Exercise 23 [• The kurtosis is not linear]
Expectation is additive (i.e. $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$) and multiplicative (i.e. $\mathbb{E}(X \cdot Y) = (\mathbb{E}X) \cdot (\mathbb{E}Y)$ for independent random variables). If you have any doubt, you should re-prove it!

Now, for $k = 2$, we have (remember that $\mathbb{E}X = \mathbb{E}Y = 0$):

$$\mathbb{E}(X+Y)^2 = \mathbb{E}(X^2+2XY+Y^2) = \mathbb{E}X^2+2\mathbb{E}(XY)+\mathbb{E}Y^2 = \mathbb{E}X^2+2(\mathbb{E}X)(\mathbb{E}Y)+\mathbb{E}Y^2 = \mathbb{E}X^2+\mathbb{E}Y^2$$

For $k = 3$, we have (remember that $\mathbb{E}X = \mathbb{E}Y = 0$):

$$\mathbb{E}(X+Y)^3 = \mathbb{E}(X^3+3X^2Y+3XY^2+Y^3) = \mathbb{E}X^2+2\mathbb{E}X^2\mathbb{E}Y+3\mathbb{E}X\mathbb{E}Y^2+\mathbb{E}Y^3 = \mathbb{E}X^3+\mathbb{E}Y^3$$

For $k = 4$, there will be some mixed terms $(\mathbb{E}X^2)(\mathbb{E}Y^2)$ that won't dispear. For example, let X and Y be random variable following a Rademacher distribution, that is $\mathbb{P}(X = -1) = \mathbb{P}(X = +1) = 1/2 = \mathbb{P}(Y = +1) = \mathbb{P}(Y = -1)$. We have $\mathbb{E}X = \mathbb{E}Y = 0$, and $X^4 = 1$, so $\mathbb{E}X^4 = \mathbb{E}Y^4 = 1$. But the distribution of $X + Y$ is:

k	-2	0	2
$\mathbb{P}(X+Y=k)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Hence $\mathbb{P}\big((X+Y)^4=0\big)=\mathbb{P}\big((X+Y)^4=16\big)=\frac{1}{2}$, and $\mathbb{E}(X+Y)^4=8\neq 1+1$.

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 2

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Exercise 1 [Grouping groups]

In one room, 9 people are seated and their average age is 25 years. In another room, 11 people are gathered and their average age is 45 years. The two groups are brought together. Compute the average age of the combined group.

Exercise 2 [Averaging averages]

In a group, 25% of the families have one child, 40% have two children, 20% have three children, and 15% have four children. What is the average number of children in this group?

Exercise 3 [Artificially increasing salaries]

In a company of 360 employees, managers earn on average 2000€ and workers earn on average 1200€. The average salary of all employees is 1400€.

1. What is the number of managers?
2. If all workers receive a 5% raise:
   (a) What is the new average salary of the employees in the company?
   (b) Does the range of salaries increase?
   (c) Does the median salary increase?
3. Answer the same questions if, instead of increasing the workers' salaries, all managers receive a 5% raise.

Exercise 4 [⋆ Discrete versus continuous variables]

Here are the 72 results of an English exam (in France, grade are from 0/20 (bad) to 20/20 (good)):

12	8	15	11	4	7	13	2	9	10	17	13	14	3	6	6	8	12	9	16	12	9	4	15
5	3	13	2	18	5	6	11	10	14	6	8	17	10	14	11	16	10	8	10	9	11	10	14
7	13	19	14	10	15	12	13	6	12	11	9	13	16	15	13	5	10	7	16	10	8	16	11

- **Discrete quantitative variable:**

1. Complete the following table:

Scores	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency																					

2. Determine the mode, the mean, and the standard deviation of the dataset using this representation.

- **Continuous quantitative variable:**

1. Complete the following table:

Score intervals	[0,4[	[4,8[	[8,12[	[12,16[	[16,20]
Frequency					

2. Determine the mode, the mean, and the standard deviation of the dataset using this representation.

Exercise 5 [Basic discrete statistics]

A population of households has been classified according to the number of family shares used to calculate income tax.

Number of shares	1	1.5	2	2.5	3	3.5	4	4.5
Number of households	48	58	136	184	210	122	62	12

1. What is the variable studied and what are its modalities?
2. Determine the range and the mode of this dataset.
3. Provide a graphical representation of this population.
4. Plot the cumulative frequency curve.
5. Calculate the median, the mean, and the standard deviation of this variable.

Exercise 6 [Basic continuous statistics]

Distribution of monthly incomes in a population of 4000 people (in euros):

Salaries	[0,1000[	[1000,2000[	[2000,3000[	[3000,4000[	[4000,5000[	[5000,6000[
Frequencies	1431	685	1180	468	204	32

1. What is the variable studied and what are its modalities?
2. Determine the range and the mode of this dataset.
3. After calculating the frequencies, represent the dataset graphically.
4. Calculate the mean and the standard deviation of this dataset.
5. Draw the cumulative frequency diagram.
6. Deduce the median, the quartiles, and draw the corresponding box plot.
7. Among salaries below 1000 euros, two-thirds are less than 500 euros. How many employees does this represent?

Exercise 7 [Comparing people]

Two shooters, X and Y, compete for selection in an event consisting of 25 shots on target. The results are recorded in the following table:

Points	50	30	20	10	0
X	5	7	6	5	2
Y	7	4	6	4	4

1. Does the average score per shot allow us to distinguish between the two competitors? What happens if we remove the five worst shots of each?
2. Calculate the median of the two datasets. Can the competitors be distinguished?
3. Represent the two datasets with two histograms, one below the other. In your opinion, which shooter is more consistent?
4. Calculate the standard deviations of the two datasets.

Exercise 8 [Why is the mean a "good" measure?]

Let $(x_1, \dots, x_n)$ be n real observed values with $n \in \mathbb{N}^*$, and let $\overline{x}$ be their arithmetic mean. Show that $\overline{x}$ is the unique minimum of the function

$$f(y) = \sum_{i=1}^{n} (x_i - y)^2, \quad y \in \mathbb{R}.$$

Exercise 9 [⋆ Did the teacher forget me?]

Info: You can either use the form of Glivenko–Cantelli theorem seen during the lecture (that is $\mathbb{P}\big(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \geq \varepsilon\big) \leq \frac{1}{4n\varepsilon^2}$), or use Dvoretzky–Kiefer–Wolfowitz inequality: for samples $x_1, \dots, x_n$ from a distribution $F(t) = \mathbb{P}(X \leq t)$, and $F_n(t) = \frac{1}{n}\sum_k \mathbf{1}(x_k \leq t)$, then when $n \to +\infty$, we have for all $\varepsilon > 0$: $\mathbb{P}\big(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \geq \varepsilon\big) \leq 2\mathrm{e}^{-2n\varepsilon^2}$.

There n (≤ 13 in reality) exercise sessions in the semester. There are s (≤ 15) students in the course. At each exercise session, k (≤ 5) students are chosen to present an exercise on the board. We will assume that this choice is made uniformly at random and independently.

1. How many exercises each student is expected to have presented on the board at the end of the semester?
2. You are a student of this course. As you love the course, you have prepared all the exercises and are willing to present them, but and you have currently presented 0 exercise, even though the m^{th} exercise session is just finished (*schaaaaaaaaaade*). Using Glivenko–Cantelli theorem, compute the smallest m such that you should be surprised (i.e., such that the probability of having been asked to present 0 time is less than 10%). Apply with $s = 15$ and $k = 5$.
3. After m^{th} sessions, you remark that your friend (who also prepare all the exercise and is willing to present them) had to present an exercise at least once during each session. What is the minimum m that would surprise you? Apply with $s = 15$ and $k = 5$.

Exercise 10 [Gini index]

The *Gini index* G for the incomes of a population of n persons is defined[1] as follows: let $y_1, \dots, y_n$ be the incomes of the n persons at stake, sorted so that $y_1 \leq \cdots \leq y_n$; let $z_k = \sum_{j=1}^{k} y_j$ be the associated cumulated incomes; plot the points $\left(\frac{1}{n}\cdot k, \frac{1}{z_n}\cdot z_k\right)$ and draw the broken line passing through these points (called the Lorenz curve); compute the integral under this curve and divide by $\frac{1}{2}$ to get $1 - G$.

1. (Draw it) Show that the Lorenz curve C is contained inside the triangle with vertices $(0,0)$, $(1,0)$ and $(1,1)$. Deduce that $0 \leq G \leq 1$.

[1] This definition is weirdly stated on purpose to force you to understand non-mathematician people writing math.

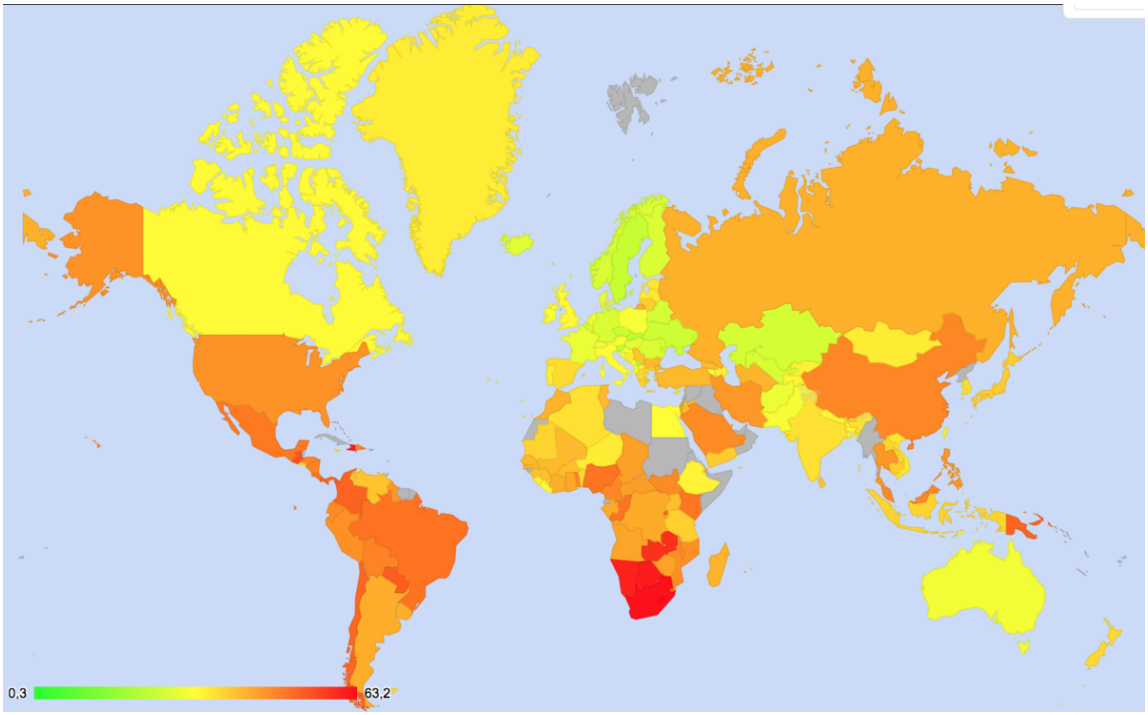


Figure 1: Gini index expressed in % (data have been collected over various years, from 1980 to 2020)

2. Show that $G = 0$ if and only if there are no inequalities of incomes (i.e. everyone is paid the same).
3. Show that $G = 1$ if and only if there is a dictator (i.e. a single person gets all the money).
4. Show that:
$$G = \frac{2}{n} \frac{\sum_{j=1}^{n} j \cdot y_j}{\sum_{j=1}^{n} y_j} - \frac{n+1}{n}$$
5. (Spearman–Brown formula) Instead of giving you clean data, someone gives you the cumulative data $\big((c_1, d_1), \ldots, (c_r, d_r)\big)$, i.e. someone tells you that "the 100 poorest persons earn 5€ in total", "the 1000 poorest persons earn 7€ in total", ..., "the c_k poorest persons earn d_k€ in total". Show that:
$$G = 1 - \sum_{k=0}^{n-1} (c_{k+1} - c_k)(d_{k+1} - d_k)$$

Exercise 11 [• Atkinson index]

For $\varepsilon \geq 0$ with $\varepsilon \neq 1$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the Atkinson index is defined as (recall the $\overline{\boldsymbol{y}}$ is the mean of $\boldsymbol{y}$):
$$A_\varepsilon(\boldsymbol{y}) = 1 - \frac{1}{\overline{\boldsymbol{y}}} \left(\frac{1}{n} \sum_{k=1}^{n} y_k^{1-\varepsilon} \right)^{1/(1-\varepsilon)}$$

1. Show that $A_\varepsilon(\boldsymbol{y}) = 0$ if and only if either $\varepsilon = 0$ or $y_j = \overline{\boldsymbol{y}}$ for all j.
2. Suppose $\boldsymbol{y}$ describes the incomes of a population. The State is writing a bill which will to decide to whom some money shall be given. The goal is to keep the Atkinson index unchanged. Show that if ε is small (no aversion for income inequality), then money shall be given equally to everyone, whether poor or rich. Show that the higher ε is (high aversion for income inequality), the more money shall be given to poor people and the less to rich people.
3. Show that Atkinson index satisfies the *principle of transfers*: if a rich individual with income y_i transfers x€ (with $x > 0$) to a poor individual with income y_j such that $y_i - x > y_j + x$,

then Atkinson index decreases. Moreover, the bigger x, the more the index decreases; and the bigger ε is, the more the index decreases.

Atkinson index is used (with ε between 0.5 and 1.5 typically) instead of Gini index, in order to detect inequalities between parts of the population with high incomes versus with low incomes. The smaller the ε, the more one consider as acceptable that inequalities of income exists (i.e. when $\varepsilon = 0$, increasing the income of anyone leads to the same increases in the happiness of the whole population). On the opposite, when $\varepsilon = +\infty$, one consider that the social welfare is driven by the income of the poorest individual (i.e. if you want to make your country happier, you should increase the income of the poorest individual).

Exercise 12 [• Various means]

There are other ways to define a notion of "mean" of two (positive) numbers x and y, explicitly:

- arithmetic mean: $\frac{x+y}{2}$
- geometric mean: $\sqrt{xy}$
- harmonic mean: $\frac{2}{\frac{1}{x}+\frac{1}{y}}$
- quadratic mean: $\sqrt{\frac{x^2+y^2}{2}}$
- logarithmic mean: $\frac{\ln y - \ln x}{y-x}$

A *mean function* is a function $m : \mathbb{R}^2 \to \mathbb{R}$ satisfying the following 3 assumptions:

a. if $x \leq y$, then $x \leq m(x, y) \leq y$
b. $m(x, y) = m(y, x)$
c. $m(x, y) = x$ if and only if $x = y$

1. Prove that the five means presented above are mean functions.

2. Which of these five means functions are 1-homogeneous (i.e. for all x, y, t, one have $m(t \cdot x, t \cdot y) = t \cdot m(x, y)$)?

3. Which of these five means functions are increasing (i.e. for x, x', y, if $x \leq x'$, then $m(x, y) \leq m(x', y)$)?

4. In general, the *Hölder mean* or *power mean* for $p \geq 1$ is defined as $\left(\frac{x^p+y^p}{2}\right)^{\frac{1}{p}}$ for some $p \geq 1$. Show that four of these five means are Hölder means, and study the Hölder means for $p = -\infty$ and $p = +\infty$.

5. For fixed x, y, study the function $f : p \mapsto \left(\frac{x^p+y^p}{2}\right)^{\frac{1}{p}}$ for $p \in \mathbb{R}$, especially its variation (you are allowed to use your knowledge on derivation...). Conclude that the usual means are ordered:

$$\min(x, y) \leq \frac{2}{\frac{1}{x} + \frac{1}{y}} \leq \sqrt{xy} \leq \frac{x+y}{2} \leq \sqrt{\frac{x^2+y^2}{2}} \leq \max(x, y)$$

6. If you are still motivated, show that the logarithmic mean is always in-between the geometric and the arithmetic mean.

You can also have a look at the Wikipedia pageS on meanS.

Exercise 13 [• Means as best approximations]

Let $f : I \to \mathbb{R}$ be a continuous and injective function. The f-mean (or Kolmogorov–Nagumo–de Finetti mean) of $\boldsymbol{x} = (x_1, \ldots, x_n)$ is defined as:

$$M_f(\boldsymbol{x}) = f^{-1}\left(\frac{1}{n}\sum_{k=1}^{n} f(x_k)\right)$$

1. Show that the Hölder means (defined in the previous exercise) are f-means for $f = x^p$, in particular that the arithmetic (respectively geometric, harmonic, quadratic) means are f-means for $f(t) = at + b$ (respectively $f(t) = \log_c t$, $f(t) = \frac{1}{t}$, $f(t) = t^2$) for any a, b, c with $a \neq 0$, $c \neq 1$.
2. A semi-distance[2] is a function $d : \mathbb{R}^2 \to \mathbb{R}_+$ such that $d(x, y) = 0$ if and only if $x = y$, and $d(x, y) = d(y, x)$ for all x, y. For any continuous and injective function f, show that $d_f : x, y \mapsto \big(f(x) - f(y)\big)^2$ is a semi-distance.
3. For some fixed continuous and injective function f with associated semi-distance d_f, and numbers $\boldsymbol{x} = (x_1, \dots, x_n)$, show that the f-mean $M_f(\boldsymbol{x})$ minimizes the function:

$$y \longmapsto \sum_{k=1}^{n} d_f(y, x_k)$$

4. Fixed numbers $\boldsymbol{x} = (x_1, \dots, x_n)$. For the trivial semi-distance $\delta(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{else} \end{cases}$, show that the minimizer of $y \mapsto \sum_{k=1}^{n} \delta(y, x_k)$ is the mode of $\boldsymbol{x}$.
5. Fixed numbers $\boldsymbol{x} = (x_1, \dots, x_n)$. For the absolute semi-distance $d(x, y) = |x - y|$, show that the minimizer of $y \mapsto \sum_{k=1}^{n} d(y, x_k)$ is a median of $\boldsymbol{x}$.

(Open problem) Find a semi-distance d such that the minimizer of $y \mapsto \sum_{k=1}^{n} d(y, x_k)$ is the logarithmic mean of $\boldsymbol{x}$ (as far as I know, the existence of non-existence of such a semi-distance has not been shown).

Exercise 14 [⋆ • Delta method, Kolmogorov means, length bias]

Recall that:

- The mean value theorem exists.
- X_n converges in probability towards Y, denoted $X_n \xrightarrow{p} Y$, if $\mathbb{P}\big(|X_n - Y| \geq \varepsilon\big) \xrightarrow[n \to +\infty]{} 0$ for all $\varepsilon > 0$.
- X_n converges in distribution towards Y, denoted $X_n \xrightarrow{d} Y$, if $\sup_t \big|\mathbb{P}(X_n \leq t) - \mathbb{P}(Y \leq t)\big| \xrightarrow[n \to +\infty]{} 0$.
- In general, convergence in probability implies convergence in distribution. When X_n converges towards a deterministic variable (i.e. a fixed number), then both convergences are equivalent.

That been said, let's proceed.

Let X_n be a sequence of (discrete of continuous) real valued random variables satisfying a central limit theorem, i.e. $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that $g'(x)$ is defined on a neighborhood of θ, and $g'(\theta) \neq 0$. We want to prove that $\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\Big(0, \big(\sigma \cdot g'(\theta)\big)^2\Big)$.

1. Prove that for all n there is $\overline{\theta}_n \in [X_n, \theta]$ such that $\sqrt{n}\big(g(X_n) - g(\theta)\big) = g'(\overline{\theta}_n) \cdot \sqrt{n}\big(X_n - \theta\big)$.
2. Show that $\overline{\theta}_n \xrightarrow{p} \theta$.
3. Conclude that $\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\Big(0, \big(\sigma \cdot g'(\theta)\big)^2\Big)$.

Application: The Kolmogorov mean associated to a continuous and injective function $f : I \to J$ (I and J two intervals of $\mathbb{R}$) is defined as $M_f(\boldsymbol{x}) = f^{-1}\left(\frac{1}{n}\sum_{k=1}^{n} f(x_k)\right)$. Similarly, the Kolmogorov expectation associated to f is $\mathbb{E}_f(X) = f^{-1}\big(\mathbb{E}f(X)\big)$.

4. Show that if f is differentiable (on a neighborhood of $\mathbb{E}_f X$), then $\sqrt{n}\big(M_f(\boldsymbol{X}) - \mathbb{E}_f X\big)$ converges in distribution towards a normal distribution. Give the standard deviation of this normal distribution.

Application of the application You try to measure the expected duration of cancer treatments at your local hospital. To do so, you go to the hospital on a given day (say, today), you get the contact of all the patients and wait for them to finish their treatment (either by attaining remission or not, sadly). The problem is that you cannot just take the mean of your sample, because you have not seen all the patients who go under treatment at this hospital, only the ones who were undergoing a treatment on the day you were there. Especially, if a patient was treated on a very short period of

[2] To create a distance, one needs to require the triangular inequality, that is $d(x, y) \leq d(x, z) + d(z, y)$ for all x, y, z.

time, you are very likely to have missed him/her in your data. We want to correct this *length bias* of your *cross-section study*.

Let f be the density function of the true durations of a cancer treatment at your local hospital. Let g be the density function of the durations you observed. As "the shorter the length, the less likely you are to observe it", you can suppose that $g(t)$ is proportional to $t \cdot f(t)$ (of course, $f(t) = g(t) = 0$ for $t \leq 0$).

5. Compute $g(t)$.
6. Show that the expectation of the true duration is estimated by the harmonic mean of the observed durations (the harmonic mean is the Kolmogorov mean for the function $x \mapsto 1/x$).

Exercise 15 [⋆ Quantile convergence]

Let F be a continuous distribution function, and $p \in [0, 1]$. Let $x_p = \inf\{x \in \mathbb{R} \ ; \ F(x) \geq p\}$.

Let $x_1, \ldots, x_n$ be samples for F. Show that, almost surely, there exists a permutation $\sigma \in S_n$ such that $x_{\sigma(1)} < \cdots < x_{\sigma(n)}$. Without loss of generality, we can thus assume that $x_1 < \cdots < x_n$. Let $Q_{p,n} = x_{\lfloor np \rfloor + 1}$.

Show (using Glivenko–Cantelli) that $Q_{n,p} \xrightarrow[n \to +\infty]{\text{alm. surely}} x_p$.

Exercise 16 [⋆ Body temperature]

It is known that a healthy human body has an average temperature of $36.5°C$, with a standard deviation of $0.5°C$. Sixty healthy humans are selected at random. What is the probability that their temperatures average at least $37°C$. (In reality, the human temperature depends on the time of the day, the hormonal cycle, the location in the body, etc. and the tools to measure it are biased.)

Exercise 17 [• Light bulbs are dying]

A string of 10 light bulbs is connected in series, which means that the entire string will not light up if any one of the light bulbs fails. Assume that the lifetimes of the bulbs, $\tau_1, \ldots, \tau_n$, are independent random variables that are exponentially distributed with mean λ (say $\lambda = 10$ years). Find the distribution of the life length of this string of light bulbs. Express it via the density function. Show that the expected life time of the string can give you an (unbiased, consistent) estimator for λ.

The constructor claims that the life time of his bulbs is in average 10 years (and follow an exponential distribution). You have 1 000 000 light bulbs, but your lab will shut down in 1 month (due to lack of funding, because you bought too many light bulbs). Discuss the best strategy to test the claim of the constructor (i.e. how may strings of how many light bulbs should you make?).

Introduction to Statistics – Solutions Sheet 2

Exercise 1 [Grouping groups]

Exercise 2 [Averaging averages]

Exercise 3 [Artificially increasing salaries]

Exercise 4 [⋆ Discrete versus continuous variables]

Here are the 72 results of an English exam (in France, grade are from 0/20 (bad) to 20/20 (good)):

12 8 15 11 4 7 13 2 9 10 17 13 14 3 6 6 8 12 9 16 12 9 4 15
5 3 13 2 18 5 6 11 10 14 6 8 17 10 14 11 16 10 8 10 9 11 10 14
7 13 19 14 10 15 12 13 6 12 11 9 13 16 15 13 5 10 7 16 10 8 16 11

- **Discrete quantitative variable:**

1. Complete the following table:

Scores	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Frequency	0	0	2	2	2	3	5	3	5	5	9	6	5	7	5	4	5	2	1	1	0

2. mode: 10; mean: $\frac{757}{72} \approx 10.51$; standard deviation: $\sqrt{\frac{76\,895}{5184}} \approx 3.85$

- **Continuous quantitative variable:**

1. Complete the following table:

Score intervals	[0,4[	[4,8[	[8,12[	[12,16[	[16,20]
Frequency	4	13	25	21	9

2. mode : $[8, 12]$ (it indeed contains the discrete mode); to compute the mean and the standard deviation, you need to consider that the distribution is uniform on each interval. Equivalently, we can identify each interval with its mid-point. We get: mean: $(2 \cdot 4 + 13 \cdot 6 + 25 \cdot 10 + 21 \cdot 14 + 9 \cdot 18)/72 = 11$; standard deviation: $\sqrt{\frac{163}{9}} \approx 4.26$. Note that the mean and the standard deviation are different for continuous and for discrete representation. Going to continuous, we have lost information. This is a good thing in that it helps visualizing, but it is a bad thing because we loose precision. To understand how to properly "pack" the data, please read the Wikipedia page on Histograms.

Exercise 5 [Basic discrete statistics]

Exercise 6 [Basic continuous statistics]

Exercise 7 [Comparing people]

Exercise 8 [Why is the mean a "good" measure?]

Note that, for fixed $x_1, \dots, x_n$, the function $f : y \mapsto \sum_i (x_i - y)^2$ is just a $\mathcal{C}^\infty$ function $\mathbb{R} \to \mathbb{R}_+$, hence we can derivate to get:

$$f'(y) = 2\sum_i (y - x_i) = 2\left(ny - \sum_i x_i\right)$$

Thus, $f'(y) = 0$ if and only $y = \frac{1}{n}\sum_i x_i = \overline{x}$. This value is indeed a minimum because the second derivative is $f''(y) = 2n > 0$ for all $y \in \mathbb{R}$.

Exercise 9 [⋆ Did the teacher forget me?]

Info: You can either use the form of Glivenko–Cantelli theorem seen during the lecture (that is $\mathbb{P}\big(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \geq \varepsilon\big) \leq \frac{1}{4n\varepsilon^2}$), or use Dvoretzky—Kiefer–Wolfowitz inequality: for samples $x_1, \dots, x_n$ from a distribution $F(t) = \mathbb{P}(X \leq t)$, and $F_n(t) = \frac{1}{n}\sum_k \mathbf{1}(x_k \leq t)$, then when $n \to +\infty$, we have for all $\varepsilon > 0$: $\mathbb{P}\big(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \geq \varepsilon\big) \leq 2\mathrm{e}^{-2n\varepsilon^2}$.

There n (≤ 13 in reality) exercise sessions in the semester. There are s (≤ 15) students in the course. At each exercise session, k (≤ 5) students are chosen to present an exercise on the board. We will assume that this choice is made uniformly at random and independently.

1. There are nk exercises in total, and a fixed student has probability $\frac{1}{s}$ to present each exercise, so a fixed student will present $\frac{nk}{s}$ exercises in average.
2. Consider the Bernoulli variable X with $\mathbb{P}(X = 1) = \frac{1}{s} = 1 - \mathbb{P}(X = 0)$, i.e. $F(t) = 0$ for $t < 0$, and $F(t) = \frac{1}{15}$ for $t \in [0, 1[$, and $F(t) = 1$ for $t \geq 1$. For a fixed student, presenting an exercise should be distributed according to the same distribution as X. Each exercise is the occasion of a sample of X (with $x = 1$ if we are chosen to present the exercise). Hence, having presented 0 times after m session (hence $m \cdot k$ exercises) means that $F_{mk}(t) = 0$ for $t < 1$ and $F_{mk}(t) = 1$ for $t \geq 1$. Hence we have: $\sup_t \big|F_{mk}(t) - F(t)\big| = \frac{1}{s}$. By Glivenko–Cantelli and Dvoretzky–Kiefer–Wolfowitz inequalities, we have:

   $$\mathbb{P}\Big(\sup_t \big|F_{mk}(t) - F(t)\big| \geq \frac{1}{s}\Big) \leq \min\left(\frac{s^2}{4mk},\ 2\mathrm{e}^{-\frac{2mk}{s^2}}\right)$$

   We would start to be surprised if the left-hand side is smaller than $10\% = 0.1$, so let's solve (k and s are fixed, the variable is m):

   $$0.1 = \frac{s^2}{4mk} \quad \text{or} \quad 0.1 = 2\mathrm{e}^{-\frac{2mk}{s^2}}$$

   We get:

   $$m = \frac{5}{2} \cdot \frac{s^2}{k} \quad \text{or} \quad m = \frac{\ln 20}{2} \cdot \frac{s^2}{k}$$

   As $\ln 20 \approx 3$, the right-hand expression is smaller so we should keep this one. Note that m increase with s and decrease with k, which is very reasonable: in an amphitheater of thousands of students, you would expect to never present an exercise (so $m \to +\infty$ when $s \to +\infty$); if we do a thousand exercises per session, you would expect to present each time (so $m \to 0$ when $k \to +\infty$).
   For $s = 15$ and $k = 5$, we get $m = 112.5$ and $m \approx 67.4$ for these two methods. The correct conclusion is that the professor should not ask student to present at random, otherwise it wouldn't be surprising that you never present an exercise!
3. The idea is the same, but with a different true and observed distributions: For a fixed student, the probability to present at least one exercise per session is $1 - \big(1 - \frac{1}{s}\big)^k = (1-$ probability of presenting none of the exercises). Hence $F(t) = 0$ for $t < 0$, and $F(t) = \big(1 - \frac{1}{s}\big)^k$ for $t \in [0, 1[$, and $F(t) = 1$ for $t \geq 1$. The observed distribution is $F_m(t) = 0$ for $t < 0$ and $F_m(t) = 1$ for $t \geq 0$. Be careful, here we observe each session, not each exercise.
   Thus $\sup_t \big|F(t) - F_m(t)\big| = \big(1 - \frac{1}{s}\big)^k$.
   We get (with Dvoretzky–Kiefer–Wolfowitz): $m = \frac{\ln 20}{2}\big(1 - \frac{1}{s}\big)^{-2k} \approx 3$ for $s = 15$ and $k = 5$. This situation is far more surprising than the fact of not presenting any exercise.

Exercise 10 [Gini index]

Exercise 11 [• Atkinson index]

Exercise 12 [• Various means]

There are other ways to define a notion of "mean" of two (positive) numbers x and y, explicitly:

- arithmetic mean: $\frac{x+y}{2}$
- geometric mean: $\sqrt{xy}$
- harmonic mean: $\frac{2}{\frac{1}{x}+\frac{1}{y}}$
- quadratic mean: $\sqrt{\frac{x^2+y^2}{2}}$
- logarithmic mean: $\frac{\ln y - \ln x}{y-x}$

A *mean function* is a function $m : \mathbb{R}^2 \to \mathbb{R}$ satisfying the following 3 assumptions:

a. if $x \leq y$, then $x \leq m(x,y) \leq y$
b. $m(x,y) = m(y,x)$
c. $m(x,y) = x$ if and only if $x = y$

1. Clear
2. All but logarithmic (because $\frac{\ln(t\cdot y)-\ln(t\cdot x)}{t\cdot y - t\cdot x} = \frac{1}{t}\frac{\ln y - \ln x}{y-x}$)
3. All but logarithmic (the logarithmic mean is increasing when $x, y \geq 1$, thought): just compute the derivative for x when y is fixed.
4. All but logarithmic: $p = -\infty$ gives Hölder mean $= \min(x,y)$; $p = -1$ gives harmonic mean; $p \to 0$ gives geometric mean; $p = 1$ gives arithmetic mean; $p = 2$ gives quadratic mean; $p \to +\infty$ gives maximum.
5. $\ln f(p) = \frac{1}{p}\ln\frac{x^p+y^p}{2}$, so:

$$p^2(\ln f)'(p) = -\ln\frac{x^p+y^p}{2} + p\frac{x^p\ln x + y^p \ln y}{x^p+y^p}$$

   To study the sign, the easiest way is to derivate this expression again, to get:

$$(p^2(\ln f)')'(p) = p\frac{(xy)^p(\ln x - \ln y)^2}{(x^p+y^p)^2}$$

   Hence, $p^2(\ln f)'(p)$ admits a global minimum for $p = 0$, which is 0. Thus, $p^2(\ln f)'(p)$ (and thus $(\ln f)'(p)$ itself) is positive for all p, so $\ln f$ and f itself are increasing.

   Consequently, $\lim_{p\to-\infty} f(p) \leq f(-1) \leq \lim_{p\to 0} f(p) \leq f(1) \leq f(2) \leq \lim_{p\to+\infty} f(p)$, which gives the claimes inequalities.
6. I am not motivated: please read the Wikipedia page on logarithmic means.

Exercise 13 [• Means as best approximations]

Exercise 14 [⋆ • Delta method, Kolmogorov means, length bias]

Recall that:

- The mean value theorem exists.

- X_n converges in probability towards Y, denoted $X_n \xrightarrow{p} Y$, if $\mathbb{P}\big(|X_n - Y| \geq \varepsilon\big) \xrightarrow[n\to+\infty]{} 0$ for all $\varepsilon > 0$.
- X_n converges in distribution towards Y, denoted $X_n \xrightarrow{d} Y$, if $\sup_t \big|\mathbb{P}(X_n \leq t) - \mathbb{P}(Y \leq t)\big| \xrightarrow[n\to+\infty]{} 0$.
- In general, convergence in probability implies convergence in distribution. When X_n converges towards a deterministic variable (i.e. a fixed number), then both convergences are equivalent.

That been said, let's proceed.

Let X_n be a sequence of (discrete of continuous) real valued random variables satisfying a central limit theorem, i.e. $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that $g'(x)$ is defined on a neighborhood of θ, and $g'(\theta) \neq 0$. We want to prove that $\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{d} \mathcal{N}\Big(0, \big(\sigma \cdot g'(\theta)\big)^2\Big)$.

1. Fix an event $\omega \in \Omega$, and consider the interval $[\theta, X_n(\omega)]$. By the mean value theorem for g on this interval, there exists some $\overline{\theta}_n$ in this interval such that $\frac{g(X_n(\omega)) - g(\theta)}{X_n(\omega) - \theta} = g'(\overline{\theta}_n)$. Multiplying by $\sqrt{n}(X_n(\omega) - \theta)$ gives the claim.
2. For all events $\omega \in \Omega$, we have $\overline{\theta}_n \in [X_n(\omega), \theta]$. Hence $\mathbb{P}(|\overline{\theta}_n - \theta| \geq \varepsilon) \leq \mathbb{P}(|X_n - \theta| \geq \varepsilon) \to 0$ when $n \to \infty$ for all $\varepsilon > 0$. The limits comes from the fact that θ is a deterministic number, so as X_n converges in distribution towards θ, it also converges in probability towards it. Thus $\overline{\theta}_n \xrightarrow{p} \theta$
3. For any t and any $\eta_n \to 0$, we have (for n large enough):

$$\mathbb{P}\big(\sqrt{n}(g(X_n) - g(\theta)) \leq t\big) = \mathbb{P}\left(\sqrt{n}(X_n - \theta) \leq \frac{t}{g'(\overline{\theta}_n)}\right) \sim \mathbb{P}\left(\sqrt{n}(X_n - \theta) \leq \frac{t}{g'(\theta)} + \eta_n\right) \sim \Phi\left(\frac{t}{\sigma g'(\theta)} + \eta_n\right) \to \Phi\left(\frac{t}{\sigma g'(\theta)}\right)$$

   Thus $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(0, (\sigma \cdot g'(\theta))^2\right)$
4. It is literally an application with $g = f^{-1}$ and $\theta = \mathbb{E}_f(X)$ for the sequence of variables $f(X_n)$. This leads to a central limit theorem, with mean 0 and standard deviation $\frac{\sigma}{\big(f'(\mathbb{E}_f X)\big)}$, by computing the derivative of f^{-1} at $\mathbb{E}_f X$).
5. We just need to normalize $g(t) = \lambda t f(t)$:

$$1 = \int_{-\infty}^{+\infty} g(t)\,\mathrm{d}t = \lambda \int_{-\infty}^{+\infty} t f(t)\,\mathrm{d}t = \lambda \cdot \mathbb{E}X$$

   Thus $g(t) = \frac{1}{\mathbb{E}X} t f(t)$.
6. Let's compute the harmonic expectation for a random variable Y with density g:

$$\mathbb{E}_{\text{harmo}} Y = \int_{-\infty}^{+\infty} \frac{1}{t} g(t)\,\mathrm{d}t = \int_{-\infty}^{+\infty} \frac{1}{t} t f(t)\,\mathrm{d}t = \mathbb{E}X$$

   Where X is a random variable distributed according to the true duration. We get $\mathbb{E}_{\text{harmo}} Y = \mathbb{E}X$. This, combined with our central limit theorem, ensures that taking the harmonic mean of our observations yields an unbiased and convergent estimator of the expectation of the duration of the treatment of the patients. We managed to correct for the non-observation of short treatments in our data. Note that computing the harmonic mean (of observed duration of treatments) in practice is easy: it is not some fancy stuffs, but just:

$$\text{Hamonic mean}(x_1, \ldots, x_n) = \frac{n}{\sum_k \frac{1}{x_k}}$$

Exercise 15 [⋆ Quantile convergence]

$$|F(Q_{n,p}) - F(x_p)| \leq |F(Q_{n,p}) - F_n(Q_{n,p})| + |F_n(Q_{n,p}) - F(x_p)| \leq \|F_n - F\|_\infty + \left|\frac{\lfloor np \rfloor + 1}{n} - p\right| \xrightarrow[n\to+\infty]{\text{alm. surely}} 0$$

Where we have used Glivenko–Cantelli for $\|F_n - F\| \to 0$ almost surely (in the course, you have seen that $\mathbb{P}(\|F_n(t) - F(t)\|_\infty \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2}$, but more generaly, Glivenko–Cantelli states that $\|F_n - F\| \to 0$ almost surely).

Now, F is a continuous function, so if $F(Q_{n,p}) - F(x_p) \to 0$, then $Q_{n,p} \to x_p$ because they are both defined as the "first" (i.e. infimum) value to attain p (if F is strictly increasing at x_p, that's clear, otherwise write the details and it comes out directly).

You can go one step further (good luck, this is far harder), showing that if F admits a density f which is strictly positive on a neighborhood of x_p, then we have a central limit theorem:

$$\sqrt{n}(Q_{n,p} - x_p) \xrightarrow[n\to+\infty]{\text{almost surely}} \mathcal{N}\left(0, \left(\frac{p(1-p)}{f(x_p)}\right)^2\right)$$

Exercise 16 [⋆ Body temperature]

Solution.

Let $X_1, \ldots, X_{60}$ be the body temperatures of the selected individuals. We assume that they are independent and identically distributed with mean $\mu = 36.5$ and standard deviation $\sigma = 0.5$. Consider the sample mean $\overline{X}_{60} = \frac{1}{60}\sum_{i=1}^{60} X_i$. Then $\mathbb{E}[\overline{X}_{60}] = \mu = 36.5$, and $\mathbb{V}(\overline{X}_{60}) = \frac{\sigma^2}{60} = \frac{0.5^2}{60}$. Hence the standard deviation of $\overline{X}_{60}$ is $\frac{\sigma}{\sqrt{60}} = \frac{0.5}{\sqrt{60}}$.

By the Central Limit Theorem, $\overline{X}_{60}$ is approximately normally distributed: $\overline{X}_{60} \approx \mathcal{N}\Big(36.5,\ \frac{0.5^2}{60}\Big)$.

We want to compute $\mathbb{P}\big(\overline{X}_{60} \geq 37\big)$. Standardizing, we obtain $\mathbb{P}\big(\overline{X}_{60} \geq 37\big) = \mathbb{P}\Big(Z \geq \frac{37-36.5}{0.5/\sqrt{60}}\Big) \approx \mathbb{P}(Z \geq 7.75) \approx 4.6 \times 10^{-15}$, where $Z \sim \mathcal{N}(0,1)$.

This is extremely small! The benefit of sampling 60 persons is that the variance of the sampled mean get divided by $\sqrt{60}$, with respect to the true variance of the distribution.

Exercise 17 [• Light bulbs are dying]

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 3

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Recall that:

An *unbiased estimator* for a value θ is a random variable X such that $\mathbb{E}X = \theta$. The *bias* of an estimator is $\text{Bias}(X) = \mathbb{E}(X - \theta)$ (sometimes $\mathbb{E}(|X - \theta|)$ in some books). A sequence of random variables $(X_n)_n$ is unbiased if all its random variables are unbiased, i.e. $EX_n = \theta$ for all n (large enough); a sequence $(X_n)_n$ is *asymptotically unbiased* if $\mathbb{E}X_n \to \theta$ when $n \to +\infty$.

A *consistent estimator* for a value θ is a sequence of random variables X_n such that $X_n \xrightarrow[n\to+\infty]{\text{proba.}} \theta$.

The *Delta method* consist in the following: Let X_n be a sequence of (discrete of continuous) real valued random variables satisfying a central limit theorem, i.e. $\sqrt{n}(X_n - \theta) \xrightarrow{\text{distrib.}} \mathcal{N}(0, \sigma^2)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a function such that $g'(x)$ is defined on a neighborhood of θ, and $g'(\theta) \neq 0$. Then the following holds: $\sqrt{n}\big(g(X_n) - g(\theta)\big) \xrightarrow{\text{distrib.}} \mathcal{N}\Big(0, \big(\sigma \cdot g'(\theta)\big)^2\Big)$.

Exercise 1 [Biased versus consistent]

1. Let X be a random variable such that $\mathbb{E}X$ exists. For independent samples $\boldsymbol{x} = (x_1, \ldots, x_n)$ of X, we construct the estimator $T_n(\boldsymbol{x}) = x_n$. Show that T_n is a unbiased but inconsistent estimator of $\mathbb{E}X$.
2. Let θ be a number (say $\theta = 3$), and Y_n be a random variable satisfying $\mathbb{P}(Y_n = \theta) = 1 - \frac{1}{n}$ and $\mathbb{P}(Y_n = \theta + n) = \frac{1}{n}$. Show that Y_n is a consistent estimator for θ which is biased for every n.
3. Show that if an estimator is consistent, then it is asymptotically unbiased.

Exercise 2 [Expectation of the sampled covariance]

Sample points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ uniformly at random in a cube $[-a, a]^3$ for some $a > 0$. Let S be the covariance matrix associated to this sample.

1. What is the expectation of S?
2. What happens for the cube $[-a, a]^d$ in any dimension?
3. What happens the non-centered cube $[0, 1]^d$?
4. What happens the parallelotope $\prod_{k=1}^{d}[0, a_k]$ for some $a_k \geq 0$? (do not redo computations, but find a clean an efficient line of thoughts.)

Exercise 3 [⋆ Nonlinear transformations and bias]

Recall that an estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if $\mathbb{E}\hat{\theta} = \theta$.

1. Let X be a real-valued random variable with $\mathbb{E}X = \mu$ and $\text{Var}(X) = \sigma^2 > 0$. Show that X^2 is a biased estimator of μ^2.
2. Assume that $g : \mathbb{R} \to \mathbb{R}$ is a continuous and convex function. State Jensen's inequality, and show that $\mathbb{E}g(\hat{\theta}) \geq g(\mathbb{E}\hat{\theta}) = g(\theta)$.
3. Assume that $g : \mathbb{R} \to \mathbb{R}$ is an affine function. Show that if X is an unbiased estimator for θ, then $g(X)$ is an unbiased estimator for $g(\theta)$.
4. Reciprocally, assume that $g : \mathbb{R} \to \mathbb{R}$ is a continuous function such that for all random variable X, if X is an unbiased estimator for θ, then $g(X)$ is an unbiased estimator for $g(\theta)$. For all $\varepsilon > 0$ and $t \in \mathbb{R}$, consider the random variable Y such that $\mathbb{P}(Y = t + \varepsilon) = \mathbb{P}(Y = t - \varepsilon) = \frac{1}{2}$. Using Y, show that g is an affine function.

Exercise 4 [Eigenvalues of the sample covariance matrix]
You are allowed to use *Courant–Fischer Minimax Theorem*: Let $A \in \mathbb{R}^{d\times d}$ be a real symmetric matrix with eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A)$. Then, for each $k = 1, \ldots, d$, we have:

$$\lambda_k(A) = \min_{\substack{V\subset\mathbb{R}^d \\ \dim(V)=k}} \max_{\substack{x\in V \\ x\neq 0}} \frac{x^\top Ax}{x^\top x},$$

Let X be a random variable with a covariance matrix Σ. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be samples for X and let S_n be the associated covariance matrix.

1. Recall how does S_n converges towards Σ.
2. Let $\boldsymbol{\lambda}(M)$ be the ordered list of real eigenvalues of the matrix M (ordered decreasingly) Use Courant–Fischer theorem (why can you use it?) to prove that the map $M \mapsto \boldsymbol{\lambda}(M)$ is a continuous map on the set of symmetric matrices (it is even Lipschitz).
3. How does $\boldsymbol{\lambda}(S)$ converges? Is it a consistent/unbiased estimator of $\boldsymbol{\lambda}(\Sigma)$.

Exercise 5 [⋆ Mean square error]
The *mean square error* (MSE) of an estimator $\hat{\theta}$ of a parameter θ is defined by

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big].$$

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables such that $\mathbb{E}X_i = \theta$ and $\mathrm{Var}(X_i) = \sigma^2$, where θ is an unknown parameter.

1. Consider the estimator
$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
   (a) Compute $\mathbb{E}\hat{\theta}_1$.
   (b) Compute $\mathrm{Var}(\hat{\theta}_1)$.
   (c) Deduce the mean square error of $\hat{\theta}_1$.
2. Now consider the estimator
$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i + c,$$
   where $c \in \mathbb{R}$ is a constant.
   (a) Compute the bias of $\hat{\theta}_2$.
   (b) Compute its mean square error.
   (c) For which value of c is the MSE minimized?
3. Show that, for any estimator $\hat{\theta}$,
$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \big(\mathrm{Bias}(\hat{\theta})\big)^2.$$
4. (Asymptotic unbiasedness and MSE) Let $(\hat{\theta}_n)_{n\geq 1}$ be a sequence of estimators of θ.
   (a) Show that if $\mathrm{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is asymptotically unbiased, i.e.,
$$\mathbb{E}\hat{\theta}_n \to \theta.$$
   (b) Does the converse hold? Justify your answer.
5. (Consistency and MSE) Let $(\hat{\theta}_n)_{n\geq 1}$ be a sequence of estimators of θ.
   (a) Show that if $\mathrm{MSE}(\hat{\theta}_n) \to 0$, then $\hat{\theta}_n \xrightarrow{\text{proba.}} \theta$.
   (b) Is the converse true? Justify your answer or provide a counterexample.

Exercise 6 [⋆ Maximum likelihood estimator]
Consider a sample $x_1, \ldots, x_n$ of independent and identically distributed random variables with common density

$$f_\theta(x) = \theta x^{\theta-1}, \quad x \in (0,1),\ \theta > 0.$$

1. Write down the likelihood function and the log-likelihood function associated with the sample.
2. Show that the log-likelihood function is strictly concave in θ.
3. Compute the maximum likelihood estimator $\hat{\theta}_n$ of θ.
4. Show that $\hat{\theta}_n$ can be written as a function of the sample mean of $\log x_i$.
5. Prove that $\hat{\theta}_n$ is a consistent estimator of θ.
6. Determine the asymptotic distribution of $\hat{\theta}_n$ (you are allowed to use the "Delta method" seen in Exercise 14 of Sheet 2).
7. Is the estimator biased? If yes, discuss whether it is asymptotically unbiased.

Exercise 7 [Estimating the true range]

Let X be a random variable with uniform distribution on $[0, \theta]$, where $\theta > 0$ is an unknown parameter. Consider a sample $(x_1, x_2, \ldots, x_n)$ of size n drawn from X.

We consider the following estimators:

$$T_n = 2\overline{x}_n, \quad T'_n = \max(x_1, x_2, \ldots, x_n), \quad T''_n = \frac{n+1}{n} T'_n,$$

where $\overline{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ is the sample mean.

1. (a) Show that T_n is an unbiased estimator of θ.
   (b) Determine the mean square error of T_n.
   (c) Is T_n a consistent estimator of θ?
2. (a) Determine the bias and the mean square error $r_\theta(T'_n)$ of T'_n.
   (b) Give a simple equivalent as $n \to +\infty$ of $r_\theta(T'_n)$.
3. Answer the same questions for T''_n.
4. Which of the three estimators is the best?

Exercise 8 [⋆ Estimating the true range via moments]

Let $x_1, \ldots, x_n$ be a random sample from a uniform distribution on the interval $[a, b]$. Compute the two first moments of this sample and construct estimators $\hat{a}$ and $\hat{b}$ to estimate a and b from these sampled moments.

Exercise 9 [⋆ Who knows?]

A TV game show consists of asking a contestant a sequence of multiple-choice questions. The questions are asked in increasing order of difficulty and yield increasing amounts of money. The team designing the questions decides to test the difficulty of one of them in order to determine at which stage of the game it should be asked. To do so, a survey is conducted on the population.

Modeling of the problem

- Each surveyed person is offered 3 possible answers, the correct one being answer 1.
- The behavior of a surveyed person is as follows:
  - if they know the correct answer, they give it;
  - otherwise, they choose randomly one of the three proposed answers.

  This takes into account the possibility that a person gives the correct answer by chance.
- Let X be the random variable equal to the answer given by a surveyed person (i.e. $X \in \{1, 2, 3\}$).
- Let Y be the random variable equal to 1 if the person knows the correct answer; 0 otherwise.
- Finally, let θ (the parameter to be estimated) be the probability that a person in the population knows the correct answer.

Solving the problem

1. (a) What is the distribution of Y?
   (b) Determine, as a function of θ, the distribution of X. Let $p = \mathbb{P}(X = 1)$. Express θ as a function of p.
   (c) What is, as a function of θ, the probability that a person who chose answer 1 did so because they actually knew the answer?
2. To estimate θ, the population is divided into n groups of 30 people, each group being surveyed.

For $1 \leq i \leq n$, let V_i be the random variable equal to the number of answer 1 responses obtained in group i. The random variables V_i are assumed to be mutually independent. Finally, define

$$Z_n = \frac{V_1 + \cdots + V_n}{30n}.$$

(a) Determine the expectation and the variance of Z_n. From Z_n, construct an unbiased estimator T_n of θ and determine the mean square error of T_n.

(b) Show that for all $\varepsilon > 0$,

$$\mathbb{P}(|T_n - \theta| > \varepsilon) \leq \frac{1}{20\, n\varepsilon^2}.$$

What can be deduced about the estimator T_n?

3. In this question, we aim to obtain an estimate $\hat{p}$ of p.
To do so, we proceed as follows: starting from a realization $(v_1, v_2, \ldots, v_n)$ of the sample $(V_1, V_2, \ldots, V_n)$, we seek to obtain, from this data, the best estimator for p.

(a) Express likelihood $L(p)$ as a function of p.

(b) Show that the log-likelihood, and thus the likelihood, admits a maximum (the argument of this maximum will be the chosen $\hat{p}$).

Introduction to Statistics – Solutions Sheet 3

Exercise 1 [Biased versus consistent]

1. Let $T_n(\boldsymbol{x}) = x_n$. Since the X_i are i.i.d. with finite expectation, $\mathbb{E}[T_n] = \mathbb{E}[x_n] = \mathbb{E}[X]$, so T_n is unbiased for $\mathbb{E}[X]$.
   However, for any $\varepsilon > 0$, $\mathbb{P}(|T_n - \mathbb{E}[X]| > \varepsilon) = \mathbb{P}(|X_n - \mathbb{E}[X]| > \varepsilon)$, which does not depend on n. Hence it does not converge to 0 in general, so T_n is not consistent.
2. We compute the expectation: $\mathbb{E}[Y_n] = \theta\left(1 - \frac{1}{n}\right) + (\theta + n)\frac{1}{n} = \theta + 1$. Thus Y_n is biased for every n.
   For consistency, let $\varepsilon > 0$. Then $\mathbb{P}(|Y_n - \theta| > \varepsilon) = \mathbb{P}(Y_n = \theta + n) = \frac{1}{n} \to 0$. Hence $Y_n \xrightarrow{p} \theta$, so Y_n is a consistent estimator of θ.

Exercise 2 [Expectation of sampled covariance]

1. $S = \frac{1}{n-1}\sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$ where $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_i \boldsymbol{x}_i$. S is an unbiased estimator for the true covariance matrix, so let $\boldsymbol{X} \sim \mathrm{Unif}([-a, a]^3)$, we have $\mathbb{E}S = \mathrm{Cov}X$.
   Now, denoting $\boldsymbol{X} = (X_1, X_2, X_3)$, we get $\mathbb{E}X_i = \int_{-a}^{a} t\frac{\mathrm{d}t}{2a} = 0$, and $\mathbb{V}X_i = \int_{-a}^{a}(t^2 - 0)\frac{\mathrm{d}t}{2a} = \frac{a^2}{3}$, for each coordinate $i \in \{1, 2, 3\}$.
   Besides, by independence, $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$.
   Thus:
   $$\mathbb{E}S = \frac{a^2}{3} I_3$$
2. In higher dimensions, the computation is still exactly the same, so $\mathbb{E}S = \frac{a^2}{3} I_d$.
3. The cube $[0, 1]^d$ is the translate of the cube $[-\frac{1}{2}, +\frac{1}{2}]^d$. As translation does not impact covariance matrices, we get: $\mathbb{E}S = \frac{1}{12} I_d$.
4. Let $\boldsymbol{X}$ be distributed in the parallelotope. Let $\boldsymbol{U}$ be distributed uniformly in the cube $[0, 1]^d$. Let $A = \mathrm{diag}(a_1, \ldots, a_d)$ be a diagonal matrix, then $X = AU$. We get:
   $$\mathrm{Cov}X = \mathrm{Cov}AU = A(\mathrm{Cov}U)A^T = A\left(\frac{1}{12} I_d\right) A^T = \frac{1}{12} A^2 = \mathrm{diag}\left(\frac{a_1}{12}, \ldots, \frac{a_d}{12}\right)$$

Exercise 3 [⋆ Nonlinear transformations and bias]

1. Since $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$, we have the identity $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, hence $\mathbb{E}[X^2] = \mu^2 + \sigma^2$. Since $\sigma^2 > 0$, it follows that $\mathbb{E}[X^2] \neq \mu^2$, so X^2 is a biased estimator of μ^2.

2. **Jensen's inequality.** If $g : \mathbb{R} \to \mathbb{R}$ is convex and X is an integrable random variable, then $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$.
   Applying this to $\hat{\theta}$, we obtain $\mathbb{E}[g(\hat{\theta})] \geq g(\mathbb{E}[\hat{\theta}])$. If $\hat{\theta}$ is unbiased, i.e. $\mathbb{E}[\hat{\theta}] = \theta$, this yields $\mathbb{E}[g(\hat{\theta})] \geq g(\theta)$.

3. Assume that g is affine, i.e. $g(x) = ax + b$ for some $a, b \in \mathbb{R}$. Then $\mathbb{E}[g(X)] = \mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$. If X is an unbiased estimator of θ, then $\mathbb{E}[X] = \theta$, hence $\mathbb{E}[g(X)] = a\theta + b = g(\theta)$. Thus $g(X)$ is an unbiased estimator of $g(\theta)$.

4. Let g be continuous and assume that for every random variable X, if $\mathbb{E}[X] = \theta$, then $\mathbb{E}[g(X)] = g(\theta)$.
   Fix $t \in \mathbb{R}$ and $\varepsilon > 0$, and consider the random variable Y such that $\mathbb{P}(Y = t + \varepsilon) = \mathbb{P}(Y = t - \varepsilon) = \frac{1}{2}$. Then $\mathbb{E}[Y] = t$, so by assumption, $\mathbb{E}[g(Y)] = g(t)$. But $\mathbb{E}[g(Y)] = \frac{1}{2}\big(g(t+\varepsilon) + g(t-\varepsilon)\big)$, hence $g(t) = \frac{g(t+\varepsilon)+g(t-\varepsilon)}{2}$.
   Thus g satisfies the midpoint property: $g\left(\frac{x+y}{2}\right) = \frac{g(x)+g(y)}{2}$ for all $x, y \in \mathbb{R}$.
   By a classical result, any continuous function satisfying the midpoint property is affine. Therefore, $g(x) = ax + b$ for some $a, b \in \mathbb{R}$.

Exercise 4 [Eigenvalues of the sampled covariance]

1. We recall the classical result for covariance estimation: $S_n \xrightarrow{\text{proba.}} \Sigma$, and in fact, under finite second moments, $S_n \xrightarrow{\text{almost surely}} \Sigma$.
   This follows from the course.
2. We consider the map $M \mapsto \boldsymbol{\lambda}(M) = (\lambda_1(M), \dots, \lambda_d(M))$, where eigenvalues are ordered decreasingly.
   We may apply the Courant–Fischer theorem because every real symmetric matrix admits a complete orthonormal basis of eigenvectors and has real eigenvalues.
   Let $A, B \in \mathbb{R}^{d\times d}$ be symmetric. For each k, we write $\lambda_k(A) = \min_{\dim(V)=k} \max_{\substack{x\in V \\ \|x\|=1}} x^\top Ax$.
   Now observe that for any unit vector x, $|x^\top Ax - x^\top Bx| = |x^\top (A-B)x| \le \|A-B\|_{\text{op}}$.
   Taking the maximum over $x \in V$ and then the minimum over subspaces V, we obtain $|\lambda_k(A) - \lambda_k(B)| \le \|A-B\|_{\text{op}}$.
   Hence each eigenvalue is 1-Lipschitz with respect to the operator norm, and therefore $\|\boldsymbol{\lambda}(A) - \boldsymbol{\lambda}(B)\|_\infty \le \|A-B\|_{\text{op}}$.
   In particular, the map $M \mapsto \boldsymbol{\lambda}(M)$ is continuous (indeed Lipschitz) on the space of symmetric matrices.
3. Since $S_n \to \Sigma$ almost surely and the eigenvalue map is continuous, we deduce by the continuous mapping theorem that $\boldsymbol{\lambda}(S_n) \xrightarrow{\text{almost surely}} \boldsymbol{\lambda}(\Sigma)$.
   Thus, the ordered eigenvalues of the sample covariance matrix converge almost surely to those of the true covariance matrix.
   This implies that $\boldsymbol{\lambda}(S_n)$ is a consistent estimator of $\boldsymbol{\lambda}(\Sigma)$ (and hence an asymptotically).
   In general, $\boldsymbol{\lambda}(S_n)$ is *not* an unbiased estimator of $\boldsymbol{\lambda}(\Sigma)$, i.e. $\mathbb{E}[\boldsymbol{\lambda}(S_n)] \neq \boldsymbol{\lambda}(\Sigma)$. This is due to the non-linearity of the eigenvalue map: even though S_n is an unbiased estimator of Σ, the eigenvalues are nonlinear functions of the matrix entries.

Exercise 5 [⋆ Mean square error]

1. (a) By linearity of expectation, $\mathbb{E}[\hat\theta_1] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \theta$.
   (b) Using independence, $\mathrm{Var}(\hat\theta_1) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$.
   (c) Since $\hat\theta_1$ is unbiased, $\mathrm{MSE}(\hat\theta_1) = \mathrm{Var}(\hat\theta_1) = \frac{\sigma^2}{n}$.
2. (a) The bias is $\mathrm{Bias}(\hat\theta_2) = \mathbb{E}[\hat\theta_2] - \theta = (\theta + c) - \theta = c$.
   (b) Since adding a constant does not change the variance, $\mathrm{Var}(\hat\theta_2) = \frac{\sigma^2}{n}$. Hence, $\mathrm{MSE}(\hat\theta_2) = \mathrm{Var}(\hat\theta_2) + \mathrm{Bias}(\hat\theta_2)^2 = \frac{\sigma^2}{n} + c^2$.
   (c) The MSE is minimized when $c = 0$.
3. For any estimator $\hat\theta$, $\mathrm{MSE}(\hat\theta) = \mathbb{E}\big[(\hat\theta - \theta)^2\big] = \mathbb{E}\big[(\hat\theta - \mathbb{E}[\hat\theta] + \mathbb{E}[\hat\theta] - \theta)^2\big]$. Expanding and using linearity of expectation, $\mathrm{MSE}(\hat\theta) = \mathbb{E}\big[(\hat\theta - \mathbb{E}[\hat\theta])^2\big] + (\mathbb{E}[\hat\theta] - \theta)^2 + 2\,\mathbb{E}\big[(\hat\theta - \mathbb{E}[\hat\theta])\big](\mathbb{E}[\hat\theta] - \theta)$. The last term is zero, hence $\mathrm{MSE}(\hat\theta) = \mathrm{Var}(\hat\theta) + \big(\mathrm{Bias}(\hat\theta)\big)^2$.
4.
5. (a) Since $\mathrm{MSE}(\hat\theta_n) = \mathrm{Var}(\hat\theta_n) + \mathrm{Bias}(\hat\theta_n)^2 \to 0$, both terms must go to zero. In particular, $\mathrm{Bias}(\hat\theta_n)^2 \to 0 \quad \Rightarrow \quad \mathrm{Bias}(\hat\theta_n) \to 0$, so $\mathbb{E}[\hat\theta_n] \to \theta$. Thus the estimator is asymptotically unbiased.
   (b) The converse does not hold in general. An estimator may be asymptotically unbiased while its variance does not converge to zero, so the MSE does not converge to zero.
6. (a) If $\mathrm{MSE}(\hat\theta_n) \to 0$, then $\mathbb{E}\big[(\hat\theta_n - \theta)^2\big] \to 0$. By Chebyshev's inequality, for any $\varepsilon > 0$, $\mathbb{P}(|\hat\theta_n - \theta| > \varepsilon) \le \frac{\mathbb{E}[(\hat\theta_n - \theta)^2]}{\varepsilon^2} \to 0$. Hence $\hat\theta_n \xrightarrow{P} \theta$, i.e., the estimator is consistent.
   (b) The converse is not true in general. Consistency only requires convergence in probability, which does not necessarily imply convergence of the second moment. Hence, the MSE need not converge to zero without additional assumptions (e.g. uniform integrability or bounded second moments).

Exercise 6 [⋆ Maximum likelihood estimator]

The likelihood is $L(\theta) = \prod_{i=1}^n \theta X_i^{\theta-1} = \theta^n \prod_{i=1}^n X_i^{\theta-1}$. The log-likelihood is $\ell(\theta) = n\ln\theta + (\theta - 1)\sum_{i=1}^n \ln X_i$.

We compute $\ell''(\theta) = -\frac{n}{\theta^2} < 0$, so ℓ is strictly concave.

The score equation is $\ell'(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \ln x_i = 0$, hence $\hat\theta_n = -\frac{n}{\sum_{i=1}^n \ln x_i}$. Thus $\hat\theta_n = -\frac{1}{\frac{1}{n}\sum_{i=1}^n \ln x_i}$.

By the law of large numbers, $\frac{1}{n}\sum_{i=1}^n \ln x_i \xrightarrow{a.s.} \mathbb{E}[\ln X] = -\frac{1}{\theta}$, so $\hat\theta_n \to \theta$, hence consistency.

By the central limit theorem, $\sqrt{n}\left(\frac{1}{n}\sum \ln x_i + \frac{1}{\theta}\right) \Rightarrow \mathcal{N}(0, \mathbb{V}\ln X)$, where one can compute $\mathbb{V}\ln X = \frac{1}{\theta^2}$. Using the delta method with $g(x) = -1/x$, we obtain $\sqrt{n}(\hat\theta_n - \theta) = \sqrt{n}\big(g(\frac{1}{n}\sum_i \ln x_i) - g(\frac{-1}{\theta})\big) \xrightarrow{\text{distrib}} \mathcal{N}\big(0, (\mathbb{V}\ln X)\cdot g'(\frac{-1}{\theta})^2\big) = \mathcal{N}(0, \theta^2)$ (which implies convergence in probability because θ is a deterministic real number).

The estimator is biased for finite n (nonlinear transformation), but it is asymptotically unbiased since it is consistent.

Exercise 7 [Estimating the true range]

Let $X \sim \mathcal{U}([0,\theta])$. Then $\mathbb{E}[X] = \theta/2$ and $\mathbb{V}(X) = \theta^2/12$.

1. For $T_n = 2\overline{x}_n$:
   (a) $\mathbb{E}[T_n] = 2\mathbb{E}[\overline{x}_n] = 2\cdot\theta/2 = \theta$, so T_n is unbiased.
   (b) $\mathbb{V}(T_n) = 4\,\mathbb{V}(\overline{x}_n) = 4\cdot\frac{\theta^2}{12n} = \frac{\theta^2}{3n}$, hence $r_\theta(T_n) = \frac{\theta^2}{3n}$.
   (c) Since $\overline{x}_n \to \theta/2$ a.s., we have $T_n \to \theta$, so T_n is consistent.
2. For $T'_n = \max(x_1, \ldots, x_n)$:
   Recall $\mathbb{E}[T'_n] = \frac{n}{n+1}\theta$ and $\mathbb{V}(T'_n) = \frac{n}{(n+1)^2(n+2)}\theta^2$.
   (a) Bias: $\mathrm{Bias}(T'_n) = -\frac{\theta}{n+1}$. Thus $r_\theta(T'_n) = \frac{n}{(n+1)^2(n+2)}\theta^2 + \frac{\theta^2}{(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}$.
   (b) As $n\to\infty$, $r_\theta(T'_n) \sim \frac{2\theta^2}{n^2}$.
3. For $T''_n = \frac{n+1}{n}T'_n$:
   (a) $\mathbb{E}[T''_n] = \theta$, so it is unbiased. Also $\mathbb{V}(T''_n) = \left(\frac{n+1}{n}\right)^2\mathbb{V}(T'_n) = \frac{\theta^2}{n(n+2)}$. Hence $r_\theta(T''_n) = \frac{\theta^2}{n(n+2)}$.
   (b) As $n\to\infty$, $r_\theta(T''_n) \sim \frac{\theta^2}{n^2}$.
4. $r_\theta(T_n) \sim \frac{\theta^2}{3n}$, $\quad r_\theta(T'_n) \sim \frac{2\theta^2}{n^2}$, $\quad r_\theta(T''_n) \sim \frac{\theta^2}{n^2}$. Thus T''_n has the smallest mean square error and is the best estimator among the three.

Exercise 8 [⋆ Estimation the true range, via the moments]

Here, a and b are treated as parameters. That is, we only know that the sample comes from a uniform distribution on some interval, but we do not know from which interval. Our interest is to estimate this interval. The pdf of a uniform distribution is $f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & \text{otherwise} \end{cases}$.

Hence, the first two population moments are

$$\mu_1 = \mathbb{E}(X) = \int_a^b \frac{x}{b-a}\,dx = \frac{a+b}{2} \quad \text{and} \quad \mu_2 = \mathbb{E}(X^2) = \int_a^b \frac{x^2}{b-a}\,dx = \frac{a^2+ab+b^2}{3}.$$

The corresponding sample moments are

$$\hat\mu_1 = \overline{X} \quad \text{and} \quad \hat\mu_2 = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

Equating the first two sample moments to the corresponding population moments, we have

$$\hat\mu_1 = \frac{a+b}{2} \quad \text{and} \quad \hat\mu_2 = \frac{a^2+ab+b^2}{3}.$$

Solving for a and b, we obtain the method of moments estimators:

$$\hat{a} = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} \quad \text{and} \quad \hat{b} = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)}.$$

Exercise 9 [⋆ Who knows?]

1. (a) $Y \sim \mathcal{B}(\theta)$ Bernoulli variable.
   (b) $\mathbb{P}(X=1) = \theta + (1-\theta)\frac{1}{3} = \frac{1+2\theta}{3} =: p$, $\mathbb{P}(X=2) = \mathbb{P}(X=3) = \frac{1-\theta}{3}$. Hence $\theta = \frac{3p-1}{2}$.
   (c) By Bayes: $\mathbb{P}(Y=1 \mid X=1) = \frac{\theta}{p} = \frac{3\theta}{1+2\theta}$.
2. Each $V_i \sim \mathcal{B}(30, p)$ binomial variable, independent.
   (a) $\mathbb{E}[Z_n] = p$, $\mathbb{V}(Z_n) = \frac{p(1-p)}{30n}$.
   Since $\theta = \frac{3p-1}{2}$, define $T_n = \frac{3Z_n-1}{2}$. Then $\mathbb{E}[T_n] = \theta$.
   $\mathbb{V}(T_n) = \frac{9}{4}\mathbb{V}(Z_n) = \frac{3}{40n}p(1-p)$, $r_\theta(T_n) = \frac{3}{40n}p(1-p)$.
   (b) By Bienaymé–Chebyshev: $\mathbb{P}(|T_n - \theta| > \varepsilon) \leq \frac{\mathbb{V}(T_n)}{\varepsilon^2} = \frac{3}{40n}\frac{p(1-p)}{\varepsilon^2} \leq \frac{3}{160n\varepsilon^2} \leq \frac{1}{20n\varepsilon^2}$. Thus T_n is consistent.
3. (a) Since the V_i are independent: $L(p) = \prod_{i=1}^n \binom{30}{v_i} p^{v_i}(1-p)^{30-v_i}$.
   (b) $f(p) = \ln L(p) = C + \left(\sum v_i\right)\ln p + \left(30n - \sum v_i\right)\ln(1-p)$. Then $f'(p) = \frac{\sum v_i}{p} - \frac{30n - \sum v_i}{1-p}$, $f''(p) < 0$, so f is strictly concave.
   (c) The maximum is obtained at $f'(p) = 0$, giving $\hat{p} = \frac{1}{30n}\sum_{i=1}^n v_i = Z_n$.

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 4

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Recall that:

The *moment generating function* of a random variable X is the function $t \mapsto \mathbb{E}e^{tX}$. Denoting m_p the p^{th} moment of X, that is $m_p = \mathbb{E}X^p$, we have $\mathbb{E}e^{tX} = \sum_p \frac{m_p}{p!} t^p$.

The density function of the χ^2 distribution of parameter k is:

$$x \mapsto \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Exercise 1 [Moment generating function of normal distribution]

Compute the moment generating function of $X \sim \mathcal{N}(\mu, \sigma)$, and deduce its centered moments (i.e. the moments of $X - \mathbb{E}X$).

Let $X \sim \chi^2(k)$. Using the fact that $X = \sum_{i=1}^k Z_i^2$ where $Z_i \sim \mathcal{N}(0,1)$, write Markov, then Chebychev, and then Chernoff inequalities to control the tail of the distribution of X.

Exercise 2 [Mode of χ^2]

What is the mode of the χ^2 distribution?

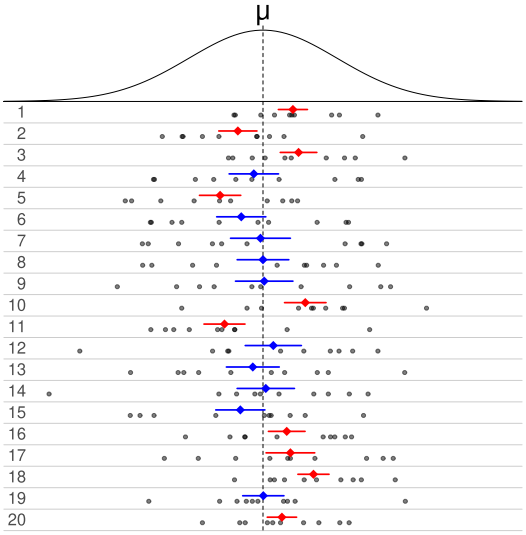
Exercise 3 [Sum of Gaussians]

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ independent. State and prove the distribution of $X + Y$.

Same question in higher dimensions: if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ are independent, then what is the distribution of $\boldsymbol{X} + \boldsymbol{Y}$ (prove it)?

Exercise 4 [⋆ Common misconceptions on confidence intervals]

In the following image (made, e.g., using R) the first horizontal line is made by sampling $m = 10$ times from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with fixed μ and σ, then computing the associated confidence interval for μ for some threshold α (the circle is at the middle of the interval, the endpoints are the extremities of the interval). This process is repeated $n = 20$ times, in order to obtain n examples shown in the picture.



1. Which value do you think was chosen for α? Give a confidence interval to support your claims!
2. Actually, $\alpha = 50\%$. Do you believe that this experiment was indeed done with the μ indicated on this figure?
3. Explain why the sizes of the confidence interval are not always the same.
4. Are the following true or false? Give a short justification for each, especially a counter-example for the false ones:
   (a) If I do a lot of trials, then the fraction of confidence intervals which do contain μ tends to $1 - \alpha$.
   (b) For each trial, there is a probability $1 - \alpha$ that the parameter μ lies within the confidence interval.
   (c) For each trial of 10 sampled points, a fraction $1 - \alpha$ (here 50%) of the sampled points lies within the confidence interval.
   (d) If I do new trials, then in a fraction $1 - \alpha$ of these, my confidence interval will intersect the one I have computed before (say, the confidence interval of the first trial).
   (e) The width of a confidence interval indicates the precision of our knowledge about the parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence errors correspond to imprecise knowledge.
   (f) A confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside.

If you want more fallacies, you should read the excellent paper "The fallacy of placing confidence in confidence intervals" by Richard D. Morey, Rink Hoekstra, Jeffrey N. Rouder, Michael D. Lee, and Eric-Jan Wagenmakers.

Exercise 5 [⋆ Confidence interval and uniform distribution]
Let $x_1, \dots, x_n$ be independent samples from a uniform distribution on $[0, b]$, where b is unknown. We want to estimate b using $M = \max(x_1, \dots, x_n)$, and establish a confidence interval.

1. Write the distribution that m is following, and deduce $\mathbb{P}(c < M < b)$ for all $c < b$.
2. Is M an unbiased and consistent estimator for b?
3. For all $\alpha \in [0, 1]$, write an interval I whose extremities depends on M, n and α (such that I is small as possible) such that $\mathbb{P}(b \in I) \geq 1 - \alpha$.

Exercise 6 [⋆ Confidence interval and exponential distribution]
Let $x_1, \dots, x_n$ be independent sample distributed according to an exponential distribution of parameter λ.

1. Recall the expectation and the variance of the exponential distribution of parameter λ.

2. Is $\overline{x}$ an unbiased and consistent estimator for $\frac{1}{\lambda}$. Is $\frac{1}{\overline{x}}$ an unbiased and consistent estimator for λ?
3. Write the confidence interval for $\frac{1}{\lambda}$ using the mean $\overline{x}$. Be careful, if the extremities of your interval depends on λ, then you can still simplify it.

Exercise 7 [⋆ • James–Stein estimator and normal distributions]

Let $\boldsymbol{x} = (x_1, \ldots, x_d)$ be a random vector such that $\boldsymbol{x} \sim \mathcal{N}(\theta, I_d)$, where $\theta \in \mathbb{R}^d$ is an unknown parameter and I_d is the $d \times d$ identity matrix. The goal is to estimate θ under quadratic loss $L(\theta, \delta) = \|\delta - \theta\|^2$.

1. Show that the usual estimator $\delta_0(\boldsymbol{x}) = \boldsymbol{x}$ is unbiased and compute its mean square error $r_\theta(\delta_0) = \mathbb{E}_\theta\left[\|\delta_0(\boldsymbol{x}) - \theta\|^2\right]$.
2. Consider estimators of the form (where $a > 0$ is a constant)

$$\delta_a(\boldsymbol{x}) = \left(1 - \frac{a}{\|\boldsymbol{x}\|^2}\right)\boldsymbol{x},$$

   Show that δ_a is biased.
3. Prove Stein's identity, that is if $Z \sim \mathcal{N}(0,1)$, and $f : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function, then:

$$\mathbb{E}[f'(Z)] = \mathbb{E}[Zf(Z)]$$

4. Using Stein's identity show that

$$r_\theta(\delta_a) = d + \mathbb{E}_\theta\left[\frac{a^2 - 2a(d-2)}{\|\boldsymbol{x}\|^2}\right].$$

5. Deduce that for $d \geq 3$, there exists a choice of a such that $r_\theta(\delta_a) < r_\theta(\delta_0)$ for all θ. Show that the optimal choice (in this class) is $a = d - 2$, leading to the James–Stein estimator

$$\delta_{\mathrm{JS}}(\boldsymbol{x}) = \left(1 - \frac{d-2}{\|\boldsymbol{x}\|^2}\right)\boldsymbol{x}.$$

6. (Optional) Explain briefly why this result is surprising from the point of view of classical estimation theory.

Exercise 8 [⋆ • Moments and median of χ^2 via Wilson–Hilferty transformation]

1. Compute the moment generating function of the χ^2 distribution with parameter k. Deduce its p^{th} moment (simplify the formula).

We want to approximate the median of the χ^2 distribution, for large k. Sadly, the exact formula is impossible to use in practice because it requires computing the reciprocal of the distribution function. Thus, we first approximate the distribution for large k.

2. Let $X \sim \chi^2(k)$. Justify that, when $k \to +\infty$, we have $\frac{X-k}{\sqrt{2k}} \xrightarrow{\text{distrib.}} \mathcal{N}(0,1)$. Computing the skewness and the kurtosis, explain why this convergence is slow.
3. Let $Z_k = \frac{X-k}{\sqrt{2k}}$. Compute, as a function of k, the three first moments of Z_k. Justify that $\mathbb{E}Z_k^p = O\left(k^{p/2-1}\right)$ for $p \geq 2$ when $k \to +\infty$.
4. Fix $\beta = \frac{1}{3}$, define $Y = \left(\frac{1}{k}X\right)^\beta$. Write the Taylor expansion of $\mathbb{E}Y$ and of $\mathbb{E}Y^2$ in term of $\mathbb{E}Z_k^p$, when $k \to +\infty$, with precision $O\left(\frac{1}{k^2}\right)$. *Hint*: three explicit terms are enough.
5. Deduce the distribution of Y is very close to a normal distribution (of which mean and variance?). Prove the following approximation of the median of X using this approximation of the distribution of Y (remember that $\beta = \frac{1}{3}$):

$$\text{median}(X) \approx k\left(1 - \frac{2}{9k}\right)^3 \qquad \text{when } k \to +\infty$$

Introduction to Statistics – Solutions Sheet 4

Exercise 1 [Moment generating function of normal distribution]
Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Write $X = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0,1)$. Then the moment generating function is $M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] = \mathrm{e}^{\mu t}\mathbb{E}[\mathrm{e}^{\sigma t Z}]$. Since $\mathbb{E}[\mathrm{e}^{sZ}] = \frac{1}{\sqrt{2\pi}} \int \mathrm{e}^{sx}\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x = \mathrm{e}^{s^2/2}$ (using a quick change of variables) for $Z \sim \mathcal{N}(0,1)$, we obtain:

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

For the centered variable $X - \mathbb{E}X = X - \mu$, we get $M_{X-\mu}(t) = \mathbb{E}[\mathrm{e}^{t(X-\mu)}] = \exp\left(\frac{\sigma^2 t^2}{2}\right)$.

Expanding the exponential, $\exp(\sigma^2 t^2/2) = \sum_{n\geq 0} \frac{1}{n!}\left(\frac{\sigma^2 t^2}{2}\right)^n$, and identifying coefficients gives the centered moments: $\mathbb{E}[(X-\mu)^{2n}] = \frac{(2n)!}{2^n n!}\sigma^{2n}$ and $\mathbb{E}[(X-\mu)^{2n+1}] = 0$.
In particular, $\mathbb{E}[(X-\mu)^2] = \sigma^2$ and $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$.

Let $X \sim \chi^2(k)$, so $X = \sum_{i=1}^k Z_i^2$ with $Z_i \sim \mathcal{N}(0,1)$ i.i.d. Hence $\mathbb{E}[X] = k$.

Markov inequality.
For any $t > 0$, $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} = \frac{k}{t}$. In particular, for $t = (1+\delta)k$, $\mathbb{P}(X \geq (1+\delta)k) \leq \frac{1}{1+\delta}$.

Chebyshev inequality.
Since $\mathrm{Var}(X) = 2k$, for any $a > 0$, $\mathbb{P}(|X-k| \geq a) \leq \frac{2k}{a^2}$. With $a = \delta k$, this becomes $\mathbb{P}(|X-k| \geq \delta k) \leq \frac{2}{\delta^2 k}$.

Chernoff bound.
For $Z \sim \mathcal{N}(0,1)$, we have $\mathbb{E}[\mathrm{e}^{tZ^2}] = (1-2t)^{-1/2}$ for $t < 1/2$. Hence for $X = \sum Z_i^2$, we get $\mathbb{E}[\mathrm{e}^{tX}] = (1-2t)^{-k/2}$.
By Chernoff's method, for any $t \in (0, 1/2)$, $\mathbb{P}(X \geq t) \leq \mathrm{e}^{-tt}(1-2t)^{-k/2}$.
Optimizing for $t = (1+\delta)k$ yields $\mathbb{P}(X \geq (1+\delta)k) \leq \exp\big(-\frac{k}{2}[\delta - \log(1+\delta)]\big)$.
Similarly, for $0 < \delta < 1$, $\mathbb{P}(X \leq (1-\delta)k) \leq \exp\big(-\frac{k}{2}[-\delta - \log(1-\delta)]\big)$.
For small δ, this implies Gaussian-type concentration: $\mathbb{P}(|X-k| \geq \delta k) \leq \exp(-ck\delta^2)$ for some universal $c > 0$.

Exercise 2 [Mode of χ^2]
Let $X \sim \chi^2(k)$. Its density is $f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1}e^{-x/2}$ for $x > 0$.

To find the mode, we maximize $f(x)$, equivalently we maximize $\log f(x)$: $\log f(x) = \big(\frac{k}{2} - 1\big)\log x - \frac{x}{2} + C$. Differentiate: $\frac{d}{dx}\log f(x) = \frac{k/2-1}{x} - \frac{1}{2}$.
Setting this derivative equal to zero gives $\frac{k/2-1}{x} = \frac{1}{2}$, hence $x = k-2$.
Since f "starts and finishes at 0", that is $f(x) = 0$ for $x < 0$, and $f(x) \to 0$ for $x \to +\infty$, and $f \geq 0$ this is indeed a maximum (not a minimum).
Therefore, $\mathrm{Mode}(X) = k-2$ if $k \geq 2$, and $\mathrm{Mode}(X) = 0$ if $0 < k < 2$.

Exercise 3 [Sum of Gaussians]
Let X, Y be independent variables with densities f and g. Then, then density of $X+Y$ can be deduced as follows:

$$\mathbb{P}(X+Y \leq t) = \mathbb{P}(X \leq t-y \text{ and } Y = y) = \int_{y\in\mathbb{R}} g(y) \int_{x\leq t-y} f(x)\mathrm{d}x\mathrm{d}y = \int_{x\leq t}\left(\int_{y\in\mathbb{R}} g(y)\cdot f(x-y)\mathrm{d}y\right)\mathrm{d}x$$

Thus the density of $X+Y$ is the so-called convolution function: $f * g : y \mapsto \int_{y\in\mathbb{R}} g(y)\cdot f(x-y)\mathrm{d}y$

Let $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ denote the density of $\mathcal{N}(0,1)$. Then its self-convolution is given by: $(\varphi * \varphi)(x) = \int_{\mathbb{R}} \varphi(y)\varphi(x-y)\,dy$. We compute: $(\varphi * \varphi)(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \exp\left(-\frac{y^2}{2} - \frac{(x-y)^2}{2}\right)dy$.

We can simplify the exponent: $y^2+(x-y)^2 = 2\left(y-\frac{x}{2}\right)^2+\frac{x^2}{2}$. Hence $(\varphi*\varphi)(x) = \frac{1}{2\pi}e^{-x^2/4}\int_{\mathbb{R}}\exp\left(-(y-x/2)^2\right)dy$.

Using $\int_{\mathbb{R}} e^{-u^2}du = \sqrt{\pi}$, we obtain $(\varphi*\varphi)(x) = \frac{1}{2\sqrt{\pi}}e^{-x^2/4}$. Therefore, $(\varphi*\varphi)(x) = \frac{1}{\sqrt{4\pi}}e^{-x^2/4}$, which is the density of $\mathcal{N}(0,2)$. That is to say: if $X\sim\mathcal{N}(0,1)$ and $Y\sim\mathcal{N}(0,1)$ are independent, then $X+Y\sim\mathcal{N}(0,2)$.

For the general case, if $X\sim\mathcal{N}(\mu_1,\sigma_1^2)$ and $Y\sim\mathcal{N}(\mu_2,\sigma_2^2)$, we can quick scribble the formula for the convolution of their density and obtain $X+Y\sim\mathcal{N}(\mu_1+\mu_2,\,\sigma_1^2+\sigma_2^2)$. **Be careful**: the *variances* are added, not the *standard deviations*!

In higher dimensions, you can continue with the annoyances of computing convolutions, or you can learn that two random variables have the same distribution if and only if they have the same characteristic functions defined as $t\mapsto\mathbb{E}e^{i\boldsymbol{t}^\top\boldsymbol{X}}$ where $\boldsymbol{t}^\top\boldsymbol{X}$ is the scalar product between the vector $\boldsymbol{t}$ and the vector $\boldsymbol{X}$. Then, write the following:

Let $\mathbf{X}\sim\mathcal{N}(\boldsymbol{\mu}_1,\Sigma_1)$ and $\mathbf{Y}\sim\mathcal{N}(\boldsymbol{\mu}_2,\Sigma_2)$ be independent random vectors in $\mathbb{R}^d$, and set $\mathbf{Z}=\mathbf{X}+\mathbf{Y}$. We compute the characteristic function. For any $t\in\mathbb{R}^d$, $\varphi_{\mathbf{Z}}(t) = \mathbb{E}[e^{i\boldsymbol{t}^\top(\mathbf{X}+\mathbf{Y})}] = \mathbb{E}[e^{i\boldsymbol{t}^\top\mathbf{X}}]\mathbb{E}[e^{i\boldsymbol{t}^\top\mathbf{Y}}]$ by independence. Then we use the characteristic function of a multivariate Gaussian, $\varphi_{\mathbf{X}}(t) = \exp\left(i\boldsymbol{t}^\top\boldsymbol{\mu}_1-\frac{1}{2}\boldsymbol{t}^\top\Sigma_1\boldsymbol{t}\right)$ and similarly for $\mathbf{Y}$. Hence, $\varphi_{\mathbf{Z}}(t) = \exp\left(i\boldsymbol{t}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)-\frac{1}{2}\boldsymbol{t}^\top(\Sigma_1+\Sigma_2)\boldsymbol{t}\right)$.

This is the characteristic function of a multivariate Gaussian, so $\mathbf{X}+\mathbf{Y}\sim\mathcal{N}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2,\Sigma_1+\Sigma_2)$.

Exercise 4 [⋆ Common misconceptions on confidence intervals]

1. I see 10 trials out of 20 where the confidence interval contains the true value, so a first guess is $\alpha=\frac{10}{20}=\frac{1}{2}$ (which is reasonable given that the data has been manually produced by a human, and humans tend to choose round numbers like 50%). We can compute a confidence interval (with threshold ρ) on α itself: containing the true value or not is a binary event, so the our 20 trials are 20 samples $x_1,\dots,x_{20}$ from a random variable $X\sim\text{Bernoulli}(\alpha)$. We have $\mathbb{E}X=\alpha$ and $\mathbb{V}X=\alpha(1-\alpha)$, so:

$$\mathbb{E}X=\alpha\in\left[\overline{x}-\frac{\sqrt{\mathbb{V}X}}{\sqrt{n\rho}},\,\overline{x}+\frac{\sqrt{\mathbb{V}X}}{\sqrt{n\rho}}\right] = \left[\frac{1}{2}-\frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{20\rho}},\,\frac{1}{2}+\frac{\sqrt{\alpha(1-\alpha)}}{\sqrt{20\rho}}\right]$$

   You can re-write the interval as two inequalities and inverse them to get more information on α, but it is more convenient to use $\alpha(1-\alpha)\leq\frac{1}{4}$ (especially since this inequality is an equality for $\alpha=\frac{1}{2}$, so this is very accurate):

$$\alpha\in\left[\frac{1}{2}-\frac{1}{2\sqrt{20\rho}},\,\frac{1}{2}+\frac{1}{2\sqrt{20\rho}}\right]$$

   Applying with $\rho=0.05$ yields: $\alpha\in[0,1]$, which is... not very informative.
2. Yes and no. Imagine μ was actually such that 0 confidence interval contains it. Then, according to the previous question, the confidence interval with threshold 95% would be $\alpha\in[-\frac{1}{2},\frac{1}{2}]$. As 0.5 lies in this interval, I would have been so surprised. To be more precise with this notion, wait for the course on testing and rejecting hypotheses.
3. The Chebychev method seen in the course give an confidence interval whose range depends only on the distribution, not on the samples, so it should be the same for each trial. However, here, the confidence intervals are not computed with the Chebychev method seen in the course, but exploit the properties of the normal distribution, which allows to use the variance of the sample to write the interval. It is very useful because we do not know the true variance *a priori*. See Student's t in the course for more details.
4. Everything is false except the first.
   (a) True: That is literally the definition.
   (b) False: This is a common confusion between knowledge *a priori* and knowledge *a posteriori*. Once a trial is done, it is done, the fact that the true value belong or not to the computed confidence interval is no longer a probabilistic question, but a deterministic one, so the probability that μ lies within the interval is either 0 or 1, not $1-\alpha$. It seems like an

annoying and overprecise play on words, but it reflect a profound (mis)understanding on what is the confidence interval.

(c) False: Take a balanced Bernoulli distribution $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$, and take 1 000 000 samples. Then the range of the confidence interval with threshold α is $2 \cdot \frac{\sqrt{\mathbb{V}X}}{\sqrt{n\alpha}} = \frac{1}{1000}\frac{1}{\sqrt{\alpha}}$ which is $\leq \frac{1}{100}$ for all $\alpha \geq 0.01$. Most likely, the sample mean would be roughly 0.5, which means that the confidence interval is approximately centered around 0.5 and is 0.01 large. However, all the values of the samples are either 0 or 1, so absolutely no sample is in the confidence interval!

(d) False: Imagine the confidence interval of first trial do not contain the true parameter. Worst, it is very far from the true parameter (which could happen out of bad luck). Then, almost no confidence interval of another trial will intersect the one of the first trial. Take trial n°1 in the figure and imagine it is slightly more to the right, then very few of the 19 other trial intersect it.

(e) False: I directly copy-paste the recommended article here ("CI" = confidence interval):
"There is no necessary connection between the precision of an estimate and the size of a confidence interval. One way to see this is to imagine that two researchers - a senior researcher and a PhD student - are analyzing the data of 50 participants from an experiment. As an exercise for the PhD student's benefit, the senior researcher decides to randomly divide the participants into two sets of 25 so that they can separately analyze half the data set. In a subsequent meeting, the two share with one another their Student's t confidence intervals for the mean. The PhD student's 95% CI is 52 ± 2, and the senior researcher's 95% CI is 53 ± 4. The senior researcher notes that their results are broadly consistent, and that they could use the equally-weighted mean of their two respective point estimates, 52.5, as an overall estimate of the true mean.
The PhD student, however, argues that their two means should not be evenly weighted: she notes that her CI is half as wide and argues that her estimate is more precise and should thus be weighted more heavily. Her advisor notes that this cannot be correct, because the estimate from unevenly weighting the two means would be different from the estimate from analyzing the complete data set, which must be 52.5. The PhD student's mistake is assuming that CIs directly indicate post-data precision. Later, we will provide several examples where the width of a CI and the uncertainty with which a parameter is estimated are in one case inversely related, and in another not related at all."

(f) False: This is a confusion between the value of one sample and the statistic over several samples. In *average*, or *over all*, repeating the experiment several times, the confidence intervals should contain the true parameter a lot of times, but this tells you nothing on whether your one experiment has yield a "nice" confidence interval or not. Computing the conditional probability on value of the parameter knowing the confidence interval is far much more complicated and heavily depends on the distribution at stake.

Exercise 5 [⋆ Confidence intervals and uniform distribution]

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{U}([0, b])$ and $M = \max(X_1, \dots, X_n)$.

1. For $t \in [0, b]$, $\mathbb{P}(M \leq t) = \mathbb{P}(X_1 \leq t \text{ and } \dots \text{ and } X_n \leq t) = \left(\frac{t}{b}\right)^n$ by independence. Hence M has density $f(t) = \frac{n}{b^n} t^{n-1} \mathbf{1}_{[0,b]}(t)$. For $0 \leq c < b$, we get:
$$\mathbb{P}(c < M < b) = 1 - \mathbb{P}(M \leq c) = 1 - \left(\frac{c}{b}\right)^n$$
2. We have $\mathbb{E}M = \int_0^b t\, f(t)\, \mathrm{d}t = \frac{n}{n+1} b$, so M is biased (it underestimates b). However, $\frac{M}{b} \xrightarrow[n\to\infty]{\mathbb{P}} 1$, since $\mathbb{P}(|M - b| > \varepsilon) = \mathbb{P}(M \leq b - \varepsilon) = \left(1 - \frac{\varepsilon}{b}\right)^n \to 0$, hence M is consistent.
3. For $\alpha \in (0, 1)$, use $\mathbb{P}\big(M \geq \alpha^{1/n} \cdot b\big) = 1 - \alpha$. This is equivalent to $\mathbb{P}\big(b \leq \frac{1}{\alpha^{1/n}} M\big) = 1 - \alpha$. Since always $M \leq b$, the smallest interval of the form $[M, cM]$ with this property is $I = \big[M,\ \frac{1}{\alpha^{1/n}} M\big]$, that is:
$$\mathbb{P}\left(b \in \left[M,\ \frac{1}{\alpha^{1/n}} M\right]\right) = 1 - \alpha.$$

Exercise 6 [⋆ Confidence interval and exponential distribution]

Let $X \sim \text{Exp}(\lambda)$, then:

1. $\mathbb{E}X = \frac{1}{\lambda}$, and $\mathbb{V}X = \frac{1}{\lambda^2}$.
2. Using the confidence interval of threshold α given by Chebychev inequality, we get that the following event occurs with a probability greater than $1-\alpha$:

$$\mathbb{E}X = \frac{1}{\lambda} \in \left[\overline{x} - \frac{\sqrt{1/\lambda^2}}{\sqrt{n\alpha}}, \overline{x} + \frac{\sqrt{1/\lambda^2}}{\sqrt{n\alpha}}\right]$$

This interval depends on $\frac{1}{\lambda}$ itself, so we can simplify further:

$$\overline{x} - \frac{1/\lambda}{\sqrt{n\alpha}} \leq \frac{1}{\lambda} \leq \overline{x} + \frac{1/\lambda}{\sqrt{n\alpha}} \iff \frac{1}{1+\frac{1}{\sqrt{n\alpha}}}\overline{x} \leq \frac{1}{\lambda} \leq \frac{1}{1-\frac{1}{\sqrt{n\alpha}}}\overline{x}$$

This gives a more relevant confidence interval (if n is big enough, then $\sqrt{n\alpha} > 1$ which ensures that both sides are well-defined):

$$\mathbb{P}\left(\frac{1}{\lambda} \in \left[\frac{1}{1+\frac{1}{\sqrt{n\alpha}}}\overline{x},\ \frac{1}{1-\frac{1}{\sqrt{n\alpha}}}\overline{x}\right]\right) \geq 1-\alpha$$

Exercise 7 [⋆ • James–Stein estimator]

1. Since $\boldsymbol{X} \sim \mathcal{N}(\theta, I_d)$, we have $\mathbb{E}_\theta[\boldsymbol{X}] = \theta$, hence $\delta_0(\boldsymbol{X}) = \boldsymbol{X}$ is unbiased. Moreover, $r_\theta(\delta_0) = \mathbb{E}_\theta[\|\boldsymbol{X}-\theta\|^2] = \text{tr}(I_d) = d$.
2. (a) The estimator $\delta_a(\boldsymbol{X}) = \left(1 - \frac{a}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}$ is nonlinear in $\boldsymbol{X}$, hence $\mathbb{E}_\theta[\delta_a(\boldsymbol{X})] \neq \theta$, so it is biased.
   (b) Write $\delta_a(\boldsymbol{X}) = \boldsymbol{X} - a\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|^2}$. Expanding the risk, $r_\theta(\delta_a) = \mathbb{E}_\theta\left[\|\boldsymbol{X}-\theta\|^2\right] + a^2\mathbb{E}_\theta\left[\frac{1}{\|\boldsymbol{X}\|^2}\right] - 2a\,\mathbb{E}_\theta\left[\frac{(\boldsymbol{X}-\theta)\cdot\boldsymbol{X}}{\|\boldsymbol{X}\|^2}\right]$. Using Stein's identity coordinatewise, $\mathbb{E}_\theta\left[\frac{(\boldsymbol{X}-\theta)\cdot\boldsymbol{X}}{\|\boldsymbol{X}\|^2}\right] = (d-2)\mathbb{E}_\theta\left[\frac{1}{\|\boldsymbol{X}\|^2}\right]$. Thus $r_\theta(\delta_a) = d + \mathbb{E}_\theta\left[\frac{a^2-2a(d-2)}{\|\boldsymbol{X}\|^2}\right]$.
3. The coefficient $a^2 - 2a(d-2)$ is negative for $0 < a < 2(d-2)$. Hence for $d \geq 3$, one can choose such an a so that $r_\theta(\delta_a) < d = r_\theta(\delta_0)$ for all θ.
   Minimizing $a^2-2a(d-2)$ gives $a = d-2$. The James–Stein estimator is $\delta_{\text{JS}}(\boldsymbol{X}) = \left(1 - \frac{d-2}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}$. Substituting, $r_\theta(\delta_{\text{JS}}) \leq d$, so it dominates δ_0, with strict inequality for $\theta \neq 0$ since $\mathbb{E}_\theta[1/\|\boldsymbol{X}\|^2] > 0$.
4. The result is surprising because it contradicts the classical intuition that the sample mean (here $\delta_0(\boldsymbol{X}) = \boldsymbol{X}$) is optimal under squared error loss. Indeed, for $d \geq 3$, there exist biased estimators (such as the James–Stein estimator) that have uniformly smaller risk than the unbiased estimator $\boldsymbol{X}$. This shows that unbiasedness is not always a desirable property in high dimensions. The improvement comes from a "shrinkage effect": the estimator pulls $\boldsymbol{X}$ toward the origin, reducing variance enough to compensate for the introduced bias. Thus, in dimensions $d \geq 3$, the estimator $\boldsymbol{X}$ is inadmissible, which is one of the most striking results in decision theory.
   Note that the James–Stein estimator is itself inadmissible (proof not given here)...

Exercise 8 [⋆ • Moments and median of χ^2 via Wilson–Hilferty transformation]

1. Let $X \sim \chi^2(k)$. Since $X = \sum_{i=1}^k Z_i^2$ with $Z_i \sim \mathcal{N}(0,1)$ independent, we first compute the mgf of Z^2. For $t < \frac{1}{2}$, $\mathbb{E}[e^{tZ^2}] = (1-2t)^{-1/2}$ (see first exercise, or just redo it writing the expectation via the density function formula).
   By independence, $M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}\prod_i \text{e}^{tZ_i} = \prod_i \mathbb{E}\text{e}^{tZ_i} = (1-2t)^{-k/2}$ for $t < \frac{1}{2}$.

To compute moments, we write $(1-2t)^{-k/2} = \sum_p \binom{-k/2}{p}(-2)^p t^p$. In particular, the coefficient on t^p is $(-2)^p \cdot \frac{-k/2(-k/2-1)(-k/2-2)\ldots(-k/2-p+1)}{p!} = \frac{1}{p!} \cdot k(k+2)(k+4)\ldots(k+2p-2)$. Hence, for integer p,

$$\mathbb{E}[X^p] = k(k+2)(k+4)\cdots(k+2(p-1))$$

2. Let $X \sim \chi^2(k)$. Then $\mathbb{E}[X] = k$ and $\mathrm{Var}(X) = 2k$. Write $Z_k = \frac{X-k}{\sqrt{2k}}$.
   By the central limit theorem applied to $X = \sum Z_i^2$, we obtain $Z_k \xrightarrow{d} \mathcal{N}(0,1)$ as $k \to \infty$.
   The skewness and kurtosis are $\gamma_1 = \frac{\mathbb{E}[(X-k)^3]}{(2k)^{3/2}} = \sqrt{\frac{8}{k}}$, $\gamma_2 = \frac{12}{k}$, using the computations for the moments done above.
   Since both decay slowly (only as $1/\sqrt{k}$), the convergence to normality is slow.
3. We compute moments using $X = k + \sqrt{2k}Z_k$.
   Mean: $\mathbb{E}[Z_k] = 0$.
   Variance: $\mathbb{E}[Z_k^2] = 1$.
   Third moment: $\mathbb{E}[Z_k^3] = \frac{\mathbb{E}[(X-k)^3]}{(2k)^{3/2}} = \sqrt{\frac{8}{k}}$.
   More generally, we look at $\mathbb{E}(X-k)^p = \sum_q \binom{q}{p}(-k)^{p-q}\mathbb{E}X^q$ where $\mathbb{E}X^q$ is a polynomial of degree q and leading coefficient 1 by the previous question. Hence, the coefficient on k^p in $\mathbb{E}(X-k)^p$ is $\sum_q \binom{q}{p}(-k)^{p-q}k^q = k^p \sum_q \binom{q}{p}(-1)^q = 0$. Thus $\mathbb{E}(X-k)^p$ is a polynomial in k of degree at most $p-1$, i.e. $\mathbb{E}(X-k)^p = O(k^{p-1})$. We get $\mathbb{E}Z_k^p = \frac{1}{\sqrt{2k^p}}\mathbb{E}(X-k)^p = O\left(k\frac{1}{2}^{p-1}\right)$ for $p \geq 2$
4. We write $\frac{1}{k}X = 1 + \sqrt{\frac{2}{k}}Z_k$, where $\sqrt{\frac{2}{k}}Z_k$ should be thought as a "small" number. Hence, we can perform a Taylor expansion of $(1+h)^\beta$ for $h \to 0$, that is:

$$(1+h)^\beta = 1 + \beta h + \frac{\beta(\beta-1)}{2}h^2 + O(h^3)$$

   This yields (this is a series expansion which converges, so we are justified to do this):

$$Y = \left(1 + \sqrt{\frac{2}{k}}Z_k\right)^\beta = 1 + \beta\sqrt{\frac{2}{k}}Z_k + \frac{\beta(\beta-1)}{2}\frac{2}{k}Z_k^2 + O\left(\frac{Z_k^3}{\sqrt{k^3}}\right)$$

   As the expectation is linear, we get, using the previous question:

$$\mathbb{E}Y = 1 + \beta\sqrt{\frac{2}{k}}\mathbb{E}Z_k + \beta(\beta-1)\frac{1}{k}\mathbb{E}Z_k^2 + O\left(\frac{\mathbb{E}Z_k^3}{\sqrt{k^3}}\right) = 1 + \frac{\beta(\beta-1)}{2}\frac{2}{k} + O\left(\frac{1}{k^2}\right)$$

   To compute $\mathbb{E}Y^2 = \mathbb{E}\left(\frac{1}{k}X\right)^{2\beta}$, we can just substitute 2β for β everywhere, yielding: $\mathbb{E}[Y^2] = 1 + 2\beta(2\beta-1)\frac{1}{k} + O(k^{-2})$. Going one step further: $\mathbb{V}Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \frac{2\beta^2}{k} + O(k^{-2})$.
5. From the previous step, applying with $\beta = \frac{1}{3}$, Y is approximately normal with $\mathbb{E}Y = 1 - \frac{2}{9k} + O(k^{-2})$, and $\mathbb{V}Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \frac{2}{9k} + O(k^{-2})$. Hence $Y \approx \mathcal{N}(1 - \frac{2}{9k}, \frac{2}{9k})$.
   Since the median of a normal distribution equals its mean, $\mathrm{median}(Y) \approx 1 - \frac{2}{9k}$.
   Undoing the transformation $Y = (X/k)^{1/3}$ gives $\left(\frac{\mathrm{median}(X)}{k}\right)^{1/3} \approx 1 - \frac{2}{9k}$.
   Cubing both sides yields $\mathrm{median}(X) \approx k \cdot \left(1 - \frac{2}{9k}\right)^3$.

(I still do not understand why taking $\beta = \frac{1}{3}$ works so well in practice, and not just another β. Especially, this $\frac{1}{3}$ does not minimize the variance, nor anything of this kind. There is something that I am missing here, but weirdly it works.)

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 5

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Recall that:
The *Cauchy distribution* as density proportional to $\frac{1}{1+x^2}$.
The density of *Student's t distribution* with parameter n is:

$$\frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

Exercise 1 [Linear transformations of normal distributions]
Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ in $\mathbb{R}^n$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathcal{M}_n(\mathbb{R})$ is symmetric positive semi-definite.

1. Let $A \in \mathcal{M}_{p,n}(\mathbb{R})$ and $b \in \mathbb{R}^p$. Show that the random vector $\boldsymbol{Y} = A\boldsymbol{X} + b$ is Gaussian.
2. Compute the expectation and covariance matrix of $\boldsymbol{Y}$.
3. Let $\boldsymbol{X} \sim \mathcal{N}\left(\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}4 & 1\\1 & 9\end{pmatrix}\right)$. Determine the law of $\boldsymbol{Y} = \begin{pmatrix}1 & -1\end{pmatrix} X$.

Exercise 2 [Covariance zero is independence for normal distribution]
Let $\boldsymbol{X} = (X_1, \dots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

1. Show that if X_i and X_j are independent, then $\mathrm{Cov}(X_i, X_j) = 0$.
2. Show that the converse is true for Gaussian vectors.
3. Give an example of two random variables that are uncorrelated but not independent.

Exercise 3 [⋆ Simulating non-standard normal distributions]
Let $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma \in \mathcal{M}_n(\mathbb{R})$ is symmetric positive definite.

1. Show that there exists an orthogonal matrix P and a diagonal matrix $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_i > 0$ such that $\Sigma = PDP^\top$.
2. Let $D^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$. Show that if $Z \sim \mathcal{N}(\mathbf{0}, I_n)$, then $\boldsymbol{X} \overset{\text{distrib.}}{=} PD^{1/2}Z$.
3. Explain how this decomposition can be used to simulate Gaussian vectors numerically.

Exercise 4 [Norm of normal distributed variables]
Let $X \sim \mathcal{N}(\mathbf{0}, I_n)$.

1. Compute $\mathbb{E}(\|X\|^2)$.
2. Compute $\mathrm{Var}(\|X\|^2)$.
3. Identify the distribution of $\|X\|^2$.
4. Show that $\frac{1}{n}\|X\|^2 \to 1$ in probability as $n \to \infty$.

Exercise 5 [Cochran-Type Orthogonal Decomposition]
Let $X_1, \dots, X_n$ be independent random variables with distribution $\mathcal{N}(0, 1)$.

1. Show that
$$\sum_{i=1}^n X_i^2 = n\overline{X}^2 + \sum_{i=1}^n (X_i - \overline{X})^2,$$
where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

2. Show that the random variables $n\overline{X}^2$ and $\sum_{i=1}^n (X_i - \overline{X})^2$ are independent.
3. Determine their distributions.

Exercise 6 [⋆ Student versus Cauchy versus normal]

Let t_n be the Student distribution with parameter $n \geq 1$.

1. Give the mean, median, mode of t_n.
2. Show that t_1 is the Cauchy distribution. (If you are motivated, re-do this question assuming you do not know the density of Student's distribution, but you only know that it is the distribution of $\frac{Z}{\sqrt{U/k}}$ where $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(k)$.)
3. When $n \to +\infty$, what is the limit of t_n? What is the mode of convergence?
4. Let F_n be the distribution function of t_n. For a fixed $x \geq 0$, prove that $F_n(x)$ increases with n. Prove that $F_n(0) = \frac{1}{2}$, and that $F_n - \frac{1}{2}$ is an odd function.
5. I want to compute the quantiles of the Student distribution, but I do not know how many parameters I should take for the Student distribution. Distinguishing between n small and n large, explain why I am justified to approximate the quantiles of the Student distribution by the one of either Cauchy or normal distribution, and comment on the error I am facing.

Exercise 7 [⋆ • χ^2, Student's t, etc, for other distributions]

Let $\mathcal{D}$ be a certain probability distribution with finite expectation $\mu = \mathbb{E}X$ and finite variance $\sigma^2 = \mathbb{V}X$. Let Δ be the distribution of the corresponding centered reduced variable, i.e. $X \sim \mathcal{D}$ if and only if $\frac{X-\mu}{\sigma} \sim \Delta$.

1. Let $X_1, \dots, X_k$ be independent distributed according to $\mathcal{D}$. Let $S_k = \sum_{i=1}^k \left(\frac{X_i - \mu}{\sigma}\right)^2$. Explain how to compute the density of S_k from the density of the $\mathcal{D}$ if it exists (do not do the computation, just explain how to do it). We denote $\psi^2(k)$ this distribution.
2. Write the centered reduced variable associated to S_k and show that its distribution approaches a standard normal distribution when $k \to +\infty$. Which assumption on $\mathcal{D}$ do you need to do so? We will make this assumption in the following questions.
3. Let $x_1, \dots, x_n$ be samples according to $\mathcal{D}$. Let s_x^2 be the corresponding sample variance, show that $\frac{n-1}{\sigma} s_x \sim \psi^2(k-1)$.
4. Let $Z \sim \Delta$, and let v_k be the distribution of $\frac{Z}{\sqrt{S_k}}$. Show that $\frac{\sqrt{n}}{s_x}(\overline{x} - \mu) \sim v_{k-1}$.

Exercise 8 [⋆ Marginal distributions]

1. Let $\boldsymbol{X} \sim \mathrm{Unif}([0,1]^d)$. Let $H \subset \mathbb{R}^d$ be an hyperplane (say $H = u^\perp$ for some unit vector $u \in \mathbb{R}^d$). Write a necessary and sufficient condition on H such that the distribution of the projection of X on H (i.e. the marginal distribution of $\mathrm{Unif}([0,1]^d)$ on H) is again a uniform distribution. For $d = 2$, draw the density of the marginal (distinguish case according to $u \in \mathbb{R}^2$) of $\mathrm{Unif}([0,1]^2)$.
2. Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \Sigma)$. State a necessary and sufficient condition on $u, v \in \mathbb{R}^d$ such that the marginals $u^\top \boldsymbol{X}$ and $v^\top \boldsymbol{X}$ are independent random variables.
3. (Cramér–Wold theorem) Let $(\boldsymbol{X}_n)_n$ be a sequence of random variables such that $\boldsymbol{X}_n \xrightarrow{\text{distrib.}} \boldsymbol{X}$ when $n \to +\infty$. Show that any marginal $u^\top \boldsymbol{X}_n \xrightarrow{\text{distrib.}} u^\top \boldsymbol{X}$ when $n \to +\infty$.

Exercise 9 [⋆ • Herschel–Maxwell's theorem]

Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random variable distributed according to a certain probability distribution. Suppose that the probability distribution is invariant under rotations, and that the coordinates of $\boldsymbol{X}$ are independent of each others (a priori, the the coordinate do not have the same distribution).

Show that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_n)$ for some real number σ. *Hint*: Start in dimension 2.

Introduction to Statistics – Solutions Sheet 5

Exercise 1 [Linear transformations of normal distributions]

1. Let $u \in \mathbb{R}^p$. Then $u^T \boldsymbol{Y} = u^T(A\boldsymbol{X} + b) = (A^T u)^T \boldsymbol{X} + u^T b$. Since $\boldsymbol{X}$ is Gaussian, every linear combination of its coordinates is a Gaussian random variable. Hence $(A^T u)^T \boldsymbol{X}$ is Gaussian, and therefore $u^T \boldsymbol{Y}$ is Gaussian. This proves that $\boldsymbol{Y}$ is a Gaussian vector because all its marginals are Gaussian.
2. By linearity of expectation, $\mathbb{E}\boldsymbol{Y} = \mathbb{E}(A\boldsymbol{X} + b) = A\mathbb{E}(\boldsymbol{X}) + b = A\boldsymbol{\mu} + b$. Moreover, $\mathbb{C}\mathrm{ov}(\boldsymbol{Y}) = \mathbb{E}\big[(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})^T\big]$. Since $\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y} = A(\boldsymbol{X} - \boldsymbol{\mu})$, we obtain
$$\mathbb{C}\mathrm{ov}(\boldsymbol{Y}) = \mathbb{E}\big[A(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T A^T\big] = A\Sigma A^T$$
3. Here,
$$A = \begin{pmatrix} 1 & -1 \end{pmatrix}, \qquad b = 0, \qquad \boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix}.$$
Therefore, $\mathbb{E}(\boldsymbol{Y}) = A\boldsymbol{\mu} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = -1$.
Its variance is $\mathbb{V}(\boldsymbol{Y}) = A\Sigma A^T$. We compute $A\Sigma = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 1 & 9 \end{pmatrix} = \begin{pmatrix} 3 & -8 \end{pmatrix}$, hence $A\Sigma A^T = \begin{pmatrix} 3 & -8 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 11$. Thus, $\boldsymbol{Y} \sim \mathcal{N}(-1, 11)$.

Exercise 2 [Covariance zero is independence for normal distribution]

1. If X_i and X_j are independent, then $\mathbb{E}(X_i X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$. Therefore, $\mathbb{C}\mathrm{ov}(X_i, X_j) = \mathbb{E}(X_i X_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = 0$.
2. Assume now that $\mathbb{C}\mathrm{ov}(X_i, X_j) = 0$. Since $\boldsymbol{X}$ is Gaussian, the vector (X_i, X_j) is a Gaussian vector in $\mathbb{R}^2$ with covariance matrix $\Sigma_{ij} = \begin{pmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_j^2 \end{pmatrix}$. Hence its density is
$$f(x, y) = \frac{1}{2\pi\sigma_i\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{(x-\mu_i)^2}{\sigma_i^2} + \frac{(y-\mu_j)^2}{\sigma_j^2}\right)\right).$$
This can be written as
$$f(x, y) = \left(\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}\right) \cdot \left(\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(y-\mu_j)^2}{2\sigma_j^2}}\right).$$
Thus $f(x, y) = f_{X_i}(x) \cdot f_{X_j}(y)$, which proves that X_i and X_j are independent.
3. Let X be uniformly distributed on $[-1, 1]$ and define $Y = X^2$. By symmetry, $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^3) = 0$. Hence $\mathbb{C}\mathrm{ov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X^3) = 0$. However, Y is a function of X, so X and Y are not independent.

Exercise 3 [⋆ Simulating non-standard normal distributions]

1. Since Σ is symmetric, the spectral theorem implies that there exists an orthogonal matrix P and a diagonal matrix $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ such that $\Sigma = PDP^\top$. Moreover, since Σ is positive definite, $x^\top \Sigma x > 0$ for all $x \neq 0$. Hence all eigenvalues of Σ are strictly positive, so $\lambda_i > 0$ for every i.
2. Let $Y = PD^{1/2}Z$. Since Y is an affine image of a Gaussian vector, it is Gaussian. Its expectation is $\mathbb{E}(Y) = PD^{1/2}\mathbb{E}(Z) = 0$. Its covariance matrix is $\mathbb{C}\mathrm{ov}(Y) = PD^{1/2}\mathbb{C}\mathrm{ov}(Z)(PD^{1/2})^\top$. Since $\mathbb{C}\mathrm{ov}(Z) = I_n$, $\mathbb{C}\mathrm{ov}(Y) = PD^{1/2}(D^{1/2})^\top P^\top$. Because $D^{1/2}$ is diagonal, $D^{1/2}(D^{1/2})^\top = D$, thus $\mathbb{C}\mathrm{ov}(Y) = PDP^\top = \Sigma$. Therefore Y and $\boldsymbol{X}$ are Gaussian vectors with the same mean and covariance matrix, hence
$$\boldsymbol{X} \overset{\text{distrib.}}{=} PD^{1/2}Z.$$

3. To simulate a Gaussian vector with distribution $\mathcal{N}(0, \Sigma)$, one may:
   (a) diagonalize $\Sigma = PDP^\top$;
   (b) generate independent standard Gaussian random variables $Z_1, \dots, Z_n$;
   (c) form the vector $Z = (Z_1, \dots, Z_n)$;
   (d) compute $X = PD^{1/2}Z$. Then X has distribution $\mathcal{N}(0, \Sigma)$.

Exercise 4 [Norm of normal distributed variables]

1. Since $\|X\|^2 = X_1^2 + \cdots + X_n^2$ and the coordinates X_i are independent standard Gaussian random variables, $\mathbb{E}(X_i^2) = 1$. Therefore, $\mathbb{E}(\|X\|^2) = \sum_{i=1}^n \mathbb{E}(X_i^2) = n$.
2. Since the variables X_i^2 are independent, $\mathbb{V}(\|X\|^2) = \sum_{i=1}^n \mathbb{V}(X_i^2)$. Now, for $X_i \sim \mathcal{N}(0,1)$, $\mathbb{E}(X_i^4) = 3$, hence $\mathbb{V}(X_i^2) = \mathbb{E}(X_i^4) - \mathbb{E}(X_i^2)^2 = 3 - 1 = 2$. Thus $\mathbb{V}(\|X\|^2) = 2n$.
3. Since the X_i are independent standard Gaussian random variables, $\|X\|^2 = X_1^2 + \cdots + X_n^2$ is the sum of n independent $\chi^2(1)$ random variables. Therefore, $\|X\|^2 \sim \chi^2(n)$.
4. We have $\mathbb{E}\big(\frac{1}{n}\|X\|^2\big) = 1$ and $\mathbb{V}\big(\frac{1}{n}\|X\|^2\big) = \frac{1}{n^2}\mathbb{V}(\|X\|^2) = \frac{2}{n}$. Since $\frac{2}{n} \to 0$, Chebyshev's inequality gives, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\|X\|^2 - 1\right| > \varepsilon\right) \le \frac{2}{n\varepsilon^2}.$$

   The right-hand side tends to 0 as $n \to \infty$. Hence $\frac{1}{n}\|X\|^2 \to 1$ in probability.

Exercise 5 [Cochran-Type Orthogonal Decomposition]

1. We write $X_i = (X_i - \overline{X}) + \overline{X}$. Hence $X_i^2 = (X_i - \overline{X})^2 + 2\overline{X}(X_i - \overline{X}) + \overline{X}^2$. Summing over i gives

$$\sum_{i=1}^n X_i^2 = \sum_{i=1}^n (X_i - \overline{X})^2 + 2\overline{X}\sum_{i=1}^n (X_i - \overline{X}) + n\overline{X}^2.$$

   Since $\sum_{i=1}^n (X_i - \overline{X}) = 0$, we obtain $\sum_{i=1}^n X_i^2 = n\overline{X}^2 + \sum_{i=1}^n (X_i - \overline{X})^2$.
2. The vector $(X_1, \dots, X_n)$ is Gaussian in $\mathbb{R}^n$. The random variable $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and the variables $X_i - \overline{X}$ are linear combinations of the X_i, hence jointly Gaussian.
   For every i, $\mathbb{C}\mathrm{ov}(\overline{X}, X_i - \overline{X}) = \mathbb{C}\mathrm{ov}(\overline{X}, X_i) - \mathbb{V}(\overline{X})$. Now, $\mathbb{C}\mathrm{ov}(\overline{X}, X_i) = \frac{1}{n}$ and $\mathbb{V}(\overline{X}) = \frac{1}{n}$. Therefore, $\mathbb{C}\mathrm{ov}(\overline{X}, X_i - \overline{X}) = 0$. Since the family is jointly Gaussian, this implies that $\overline{X}$ is independent of $(X_1 - \overline{X}, \dots, X_n - \overline{X})$. Consequently, $n\overline{X}^2$ and $\sum_{i=1}^n (X_i - \overline{X})^2$ are independent.
3. Since $\overline{X} \sim \mathcal{N}\big(0, \frac{1}{n}\big)$, we have $\sqrt{n}\,\overline{X} \sim \mathcal{N}(0,1)$. Thus $n\overline{X}^2 = (\sqrt{n}\,\overline{X})^2 \sim \chi^2(1)$.
   Moreover, $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$. Using the decomposition

$$\sum_{i=1}^n X_i^2 = n\overline{X}^2 + \sum_{i=1}^n (X_i - \overline{X})^2,$$

   together with independence and additivity of chi-square distributions, we obtain $\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(n-1)$.

Exercise 6 [⋆ Student versus Cauchy versus normal]

1. The density of the Student distribution with parameter n is $f_n(x) = c_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$, $x \in \mathbb{R}$, where $c_n = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}$. The density is even, hence the distribution is symmetric around 0. Therefore, the median is 0, the mode is 0, the mean is 0 when it exists, i.e. for $n > 1$. For $n = 1$, the expectation does not exist.
2. For $n = 1$, $c_1 = \frac{\Gamma(1)}{\sqrt{\pi}\,\Gamma(1/2)} = \frac{1}{\pi}$. Hence $f_1(x) = \frac{1}{\pi}\frac{1}{1+x^2}$, which is exactly the density of the Cauchy distribution.

3. For every fixed $x \in \mathbb{R}$, $\left(1+\frac{x^2}{n}\right)^{-n/2} \xrightarrow[n\to+\infty]{} e^{-x^2/2}$. Therefore, $f_n(x) \longrightarrow c\,e^{-x^2/2}$, for some constant c such that $\int_{\mathbb{R}} f = 1$. This is the density of the standard normal distribution (which directly proves that $c = \frac{1}{\sqrt{2\pi}}$).
   Hence $t_n \to \mathcal{N}(0,1)$, that is, convergence in distribution. Since the densities converge pointwise and are dominated, one also gets convergence of the associated distribution function $\int_{-\infty}^{t} f$, so it is a convergence in distribution.
4. Since the density is even, $F_n(0) = \frac{1}{2}$., and $F_n(-x) = 1 - F_n(x)$, thus $F_n(x) - \frac{1}{2} = -\left(F_n(-x) - \frac{1}{2}\right)$, so $F_n - \frac{1}{2}$ is odd.
   Fix $x \geq 0$. to be improved: As n increases, the Student distribution becomes more concentrated around 0 and converges to the Gaussian distribution. The tails become lighter. Therefore, $F_n(x) = \mathbb{P}(T_n \leq x)$ increases with n for every fixed $x \geq 0$. Equivalently, for $x \geq 0$, $\mathbb{P}(T_n > x)$ decreases with n.
5. In general, according to the previous questions, we have for all $\alpha \geq \frac{1}{2}$ and $i \leq j$ that:

$$q_\alpha(\text{standard normal}) \leq q_\alpha(t_j) \leq q_\alpha(t_i) \leq q_\alpha(\text{Cauchy})$$

   (for $\alpha \leq \frac{1}{2}$, the inequalities are reversed.)
   For small values of n, the Student distribution has heavy tails and resembles the Cauchy distribution. Therefore, using Cauchy quantiles gives a rough approximation. The α-quantile of the Cauchy distribution can be made explicit: $q_\alpha(\text{Cauchy}) = \tan\left(\pi(\alpha - \frac{1}{2})\right)$.
   For large values of n, the Student distribution is close to the standard normal distribution, so one may approximate its quantiles by Gaussian quantiles. These quantiles do not have an explicit formula but are widely studied. This last idea, together with some explicit computations, justifies that for $n \geq 30$, people uses the quantiles of the normal distribution in practice.

Exercise 7 [⋆ χ^2, Student's t, etc, for other distributions]

Let $\mathcal{D}$ be a certain probability distribution with finite expectation $\mu = \mathbb{E}X$ and finite variance $\sigma^2 = \mathbb{V}X$. Let Δ be the distribution of the corresponding centered reduced variable, i.e. $X \sim \mathcal{D}$ if and only if $\frac{X-\mu}{\sigma} \sim \Delta$.

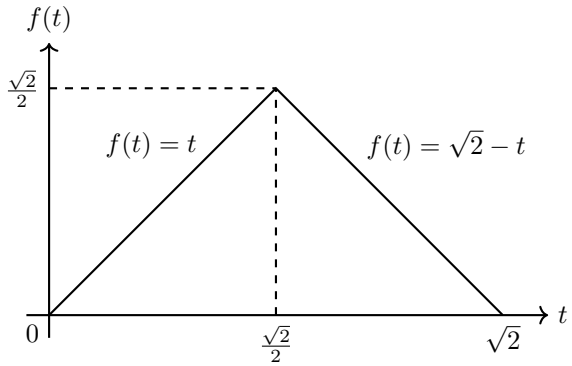
1. Let $X_1, \ldots, X_k$ be independent distributed according to $\mathcal{D}$. Let $S_k = \sum_{i=1}^{k} \left(\frac{X_i - \mu}{\sigma}\right)^2$. Explain how to compute the density of S_k from the density of the $\mathcal{D}$ if it exists (do not do the computation, just explain how to do it). We denote $\psi^2(k)$ this distribution.
2. Write the centered reduced variable associated to S_k and show that its distribution approaches a standard normal distribution when $k \to +\infty$. Which assumption on $\mathcal{D}$ do you need to do so? We will make this assumption in the following questions.
3. Let $x_1, \ldots, x_n$ be samples according to $\mathcal{D}$. Let s_x^2 be the corresponding sample variance, show that $\frac{n-1}{\sigma} s_x \sim \psi^2(k-1)$.
4. Let $Z \sim \Delta$, and let v_k be the distribution of $\frac{Z}{\sqrt{S_k}}$. Show that $\frac{\sqrt{n}}{s_x}(\overline{x} - \mu) \sim v_{k-1}$.

Exercise 8 [⋆ Marginal distributions]

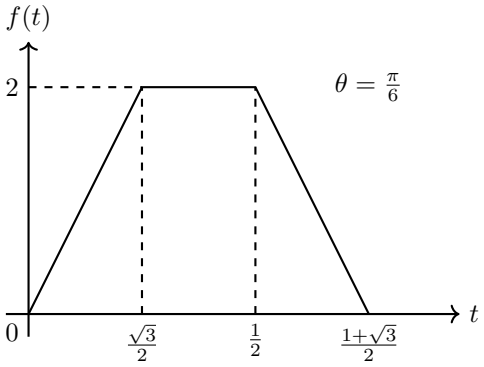
1. Let $X \sim \text{Unif}([0,1]^d)$ and let $H = \boldsymbol{u}^\perp$, where $\|\boldsymbol{u}\| = 1$. The marginal distribution on H is obtained by projecting the cube $[0,1]^d$ orthogonally onto H. Its density at a point $\boldsymbol{y} \in H$ is proportional to the length of the intersection $(\boldsymbol{y} + \mathbb{R}\boldsymbol{u}) \cap [0,1]^d$. Therefore, the marginal is uniform on the projection if and only if this intersection length is constant for all $\boldsymbol{y}$ in the projection. This happens if and only if the direction $\boldsymbol{u}$ is parallel to one coordinate axis, i.e. $u = \pm \boldsymbol{e}_i$ for some canonical basis vector $\boldsymbol{e}_i$.
   For $d = 2$, let $\boldsymbol{u} = (\cos\theta, \sin\theta)$. The projection of the square onto $H = u^\perp$ is an interval.
   If $\boldsymbol{u} = (1,0)$ or $(0,1)$, the marginal density is uniform on an interval. Otherwise, the density is trapezoidal or triangular: If $\theta = \pi/4$ (projection along a diagonal), the density is triangular; otherwise trapezoidal.

In particular, for $\boldsymbol{u} = (1,1)/\sqrt{2}$, the projected variable is proportional to $X_1 + X_2$, whose density is

$$f(t) = \begin{cases} t, & 0 \le t \le \frac{\sqrt{2}}{2}, \\ \sqrt{2} - t, & \frac{\sqrt{2}}{2} \le t \le \sqrt{2}, \\ 0, & \text{otherwise.} \end{cases}$$



Let's take $\theta = \frac{\pi}{6}$ to get a generic example, that is $\boldsymbol{u} = (\sqrt{3}, 1)/2$. To compute the height of our trapezoid, remember that the total area needs to be equal to 1.



2. Let $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$ and let $Y = u^\top X$, and $Z = v^\top X$. Since (Y, Z) is Gaussian, independence is equivalent to zero covariance. Now, $\mathbb{C}\mathrm{ov}(Y, Z) = \mathbb{C}\mathrm{ov}(u^\top X, v^\top X) = u^\top \Sigma v$. Therefore, ($u^\top X$ and $v^\top X$ are independent) if and only if $u^\top \Sigma v = 0$.
3. Let $X_n \xrightarrow{\text{distrib.}} X$. Fix $u \in \mathbb{R}^d$. The map $f(\boldsymbol{x}) = u^\top \boldsymbol{x}$ is continuous on $\mathbb{R}^d$. Hence, by the continuous mapping theorem, $u^\top X_n = f(X_n) \xrightarrow{\text{distrib.}} f(X) = u^\top X$. Therefore every one-dimensional marginal converges in distribution.

Exercise 9 [⋆ Herschel–Maxwell's theorem]

We first treat the case $n = 2$. Let $\boldsymbol{X} = (X_1, X_2)$.

Applying a rotation of 90 in the plane sends the first coordinate on the second coordinate, which proves that X_1 have the same distribution as X_2. The same works directly in higher dimension (rotating in each plane spanned by $\boldsymbol{e}_1$, and $\boldsymbol{e}_i$) shows that all the coordinated of $\boldsymbol{X}$ have the same distribution. Moreover, using a rotation of 180 in the plane proves that this common distribution is symmetric.

To be completed...

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 6

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Exercise 1 [• Rock-paper-scissors]

A computer plays rock-paper-scissors against every humans on the planet, in order to measure if humans have a preference between rock, paper, or scissors. To play against a given human, the computer will choose at random between rock (with probability p_r), paper (with probability p_p), or scissors (with probability p_s): these three numbers $\boldsymbol{p} = (p_r, p_p, p_s)$ shall be considered as fixed (and obviously $p_r + p_p + p_s = 1$). At the end of the experiment, the computer will record three numbers: the number of wins, of draws, and of loss. This result is given as a vector $\overline{\boldsymbol{x}} = (x_w, x_d, x_l)$ which are the number of win/draw/loss divided by the number of (tested) humans.

Consider that, each human chooses (independently) at random between rock, paper, and scissors according to the same (this is a lot to assume...) probability distribution $\boldsymbol{q} = (q_r, q_p, q_s)$. We write $\boldsymbol{X} = (X_w, X_d, X_l)$ a random variable counting the number of win/draw/loss of the computer, according to this distribution $\boldsymbol{q}$ (where the distribution $\boldsymbol{p}$ is a fixed parameter).

1. Write a matrix P such that $P\boldsymbol{q} = \overline{\boldsymbol{x}}$. Deduce that $\overline{\boldsymbol{x}}$ yields an unbiased and consistent estimator for $\boldsymbol{q}$, except if $\boldsymbol{p}$ has a certain value (use the indication).
2. This gives you a linear map $f : \mathbb{R}^3 \to \mathbb{R}^3$ such that $f(\overline{\boldsymbol{x}}) = \boldsymbol{q}$. Now, suppose that there is some error in the measurement, i.e. x_w, x_d, x_l are known up to some constant $\varepsilon > 0$. Using the 1-norm of the Jacobian matrix of f, estimate the error on $\boldsymbol{q}$ that this error on $\overline{\boldsymbol{x}}$ induces. How does this propagation of the error depend on the choice of $\boldsymbol{p}$?
3. The computer actually recorded the results of each game it played again each (tested) humans: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ where $\overline{\boldsymbol{x}}$ is the mean of these samples (and m the number of humans). Write the covariance matrix of these samples, and show that it is an estimator for the covariance matrix of $\boldsymbol{X}$.
4. Deduce that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ admits a central limit theorem. Is there a clever choice of $\boldsymbol{p}$ in order to minimize its standard deviation?

Indications:

- Let $M = \begin{pmatrix} a & b & 1-a-b \\ 1-a-b & a & b \\ b & 1-a-b & a \end{pmatrix}$. Then its possible eigenvalues are 1, and $\frac{3a-1}{2} \pm \frac{\sqrt{3}}{2}\, i\, |a + 2b - 1|$.
- The 1-norm of a vector is $\|\boldsymbol{x}\|_1 = \sum_k |x_k|$. The 1-norm of a matrix $\|A\|_1$ is the maximum of the 1-norms of its lines. We have that $\|A \cdot B\|_1 = \|A\|_1 \cdot \|B\|_1$ and $\|A\|_1 = \sup\{\|A\boldsymbol{x}\|_1 \ ; \ \|\boldsymbol{x}\|_1 \leq 1\}$.

Exercise 2 [Simpson paradox]

Exercise 3 [Defense attorney's fallacy]

Exercise 4 [G-test]

Exercise 5 [⋆ Gauss–Markov theorem]

Let $y \in \mathbb{R}^n$ be a random vector given by the linear model

$$y = X\beta + \varepsilon,$$

where $X \in \mathbb{R}^{n \times p}$ is a fixed matrix of full column rank $p \leq n$, $\beta \in \mathbb{R}^p$ is an unknown parameter, and ε is a random vector satisfying

$$\mathbb{E}[\varepsilon] = 0, \qquad \mathrm{Cov}(\varepsilon) = \sigma^2 I_n,$$

with unknown $\sigma^2 > 0$.

We consider linear estimators of the form

$$\hat{\beta} = Ay,$$

where $A \in \mathbb{R}^{p \times n}$ is a deterministic matrix.

1. Show that $\hat{\beta}$ is unbiased for β if and only if $AX = I_p$.
2. Compute the covariance matrix of any unbiased linear estimator $\hat{\beta} = Ay$.
3. Show that the ordinary least squares estimator

   $$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$

   is unbiased and identify its associated matrix A.
4. Let $\hat{\beta} = Ay$ be any unbiased linear estimator. Show that

   $$\mathrm{Cov}(\hat{\beta}) - \mathrm{Cov}(\hat{\beta}_{OLS})$$

   is positive semidefinite.
5. Deduce the Gauss–Markov theorem: among all linear unbiased estimators, $\hat{\beta}_{OLS}$ has minimal variance (in the sense of the Loewner order).
6. Interpret the result in terms of efficiency and discuss why no distributional assumption (such as normality) is needed.

Introduction to Statistics – Solutions Sheet 6

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 7

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Yes, you can use R, or any calculator/software. No, a LLM (e.g. ChatGPT) is not a calculator... Recall some useful lookup tables:

Student's t-distribution table:

For different values of n and $\alpha \in [0,1]$, the value $t_{n,\alpha}$ is defined by $P(T \leq t_{n,\alpha}) = \alpha$, when $T \sim t(n)$ (recall that the Student distribution is symmetric).

α	0.95	0.975
$t_{21,\alpha}$	1.721	2.08
$t_{22,\alpha}$	1.717	2.074
$t_{23,\alpha}$	1.714	2.069

Fisher distribution table:

For different values of (n_1, n_2) and $\alpha \in [0,1]$, the value $f_{n_1,n_2,\alpha}$ is defined by $P(F \leq f_{n_1,n_2,\alpha}) = \alpha$, when $F \sim \mathcal{F}(n_1, n_2)$.

α	0.025	0.05	0.95	0.975
$f_{22,20,\alpha}$	0.419	0.483	2.102	2.434
$f_{22,21,\alpha}$	0.421	0.486	2.073	2.394
$f_{22,22,\alpha}$	0.424	0.488	2.048	2.358
$f_{23,20,\alpha}$	0.424	0.488	2.092	2.420
$f_{23,21,\alpha}$	0.427	0.491	2.063	2.380
$f_{23,22,\alpha}$	0.430	0.494	2.038	2.344
$f_{24,20,\alpha}$	0.430	0.493	2.082	2.408
$f_{24,21,\alpha}$	0.433	0.496	2.054	2.368
$f_{24,22,\alpha}$	0.436	0.499	2.028	2.331

Standard normal distribution table:

For different values of q, the probability $\mathbb{P}(X \leq q)$ when $X \sim \mathcal{N}(0,1)$.

q	0.01	0.02	0.03	0.04	0.05	0.65	0.66	0.67	0.68	0.69	0.7
$\mathbb{P}(X \leq q)$	0.504	0.508	0.512	0.516	0.520	0.742	0.745	0.749	0.752	0.755	0.758

$$\mathbb{P}(X \leq 1.645) \approx 0.95 \qquad \mathbb{P}(X \leq 2.326) \approx 0.99$$

Binomial distribution table:

For different values of $\mathbb{P}(X \leq q)$ when $X \sim \text{Bin}(10, \frac{1}{2})$.

q	0	1	2	3	4	5	6	7	8	9
$\mathbb{P}(X \leq q)$	0.00098	0.01074	0.05469	0.1719	0.377	0.623	0.8281	0.9453	0.9893	0.999

Exercise 1 [⋆ Z-test]

In Germany, the reading skills of the whole population has been estimated. You can assume that there are enough people in Germany so that the actual distribution of the number of points at the reading exam is approximately normally distributed (the difference is negligible). After proper scaling, the mean and standard deviation at the reading exam is 100 points and 12 points respectively.

In Osnabrück, 55 students pass the reading exam, and their average score is 104. We want to test whether this score is abnormally high or not. Our null hypothesis is H_0: "The 55 exam takers are comparable to a simple random sample from the population of exam-takers". Design a test to accept/reject H_0 with significance level α using the quantity

$$z = \frac{\sqrt{\text{size of the sub-population}}}{\text{standard deviation in the whole population}}\big(\text{average in the sub-population} - \text{average in the whole population}\big).$$

Apply in the narrated setting and conclude.

Exercise 2 [⋆ Exponential test]

Does the samples

$$1.63,\quad 1.88,\quad 6.57,\quad 15.65,\quad 4,\quad 1.54,\quad 1.46,\quad 3.07,\quad 3.25,\quad 8.05$$

fit an exponential distribution of parameter $\lambda = 0.2$? You can assume that these numbers are indeed sampled from some exponential distribution $\mathrm{Exp}(\lambda)$ for some $\lambda \in \mathbb{R}_+^*$.

1. Write (and perform) a test with significance level $\alpha = 0.05$ for the null hypothesis $\lambda = 0.2$.
2. Same question for the null hypothesis $\lambda = 0.25$.
3. How many times should I sample if I want to be sure that (at least) one of these null hypotheses get rejected (with significance level $\alpha = 0.05$)?

Exercise 3 [⋆ Exact median test]

You have seen how to test if the median of a continuous random variable is in $[a, b]$. We now test whether the median is exactly some value a.

Let $x_1, \dots, x_n$ be i.i.d. samples of a continuous random variable X. We want accept/reject the null hypothesis $H_0 : \mathrm{med}(X) = a$ for some given number $a \in R$. Let $N = \#\{i \ ; \ x_i < a\}$.

1. Show, assuming H_0, that N follows a binomial distribution (precise its parameters).
2. We denote $b_{n,\theta}$ the θ quantile of $\mathrm{Bin}(n, \frac{1}{2})$, i.e. the smallest number such that $\mathbb{P}(Y \leq b_{n,\theta}) = \theta$ for $Y \sim \mathrm{Bin}(n, \frac{1}{2})$. Write a test for H_0 of significance level α (involving $b_{n,\theta}$ for some well-chosen θ).
3. *Application*: recall that the median of an exponential distribution $\mathrm{Exp}(\lambda)$ is $\frac{\ln 2}{\lambda}$ (where $0.693 \leq \ln 2 \leq 0.694$). Use an (exact) median test to accept/reject whether $\lambda = 0.2$ with following samples (same data as Exercise 2):

$$1.63,\quad 1.88,\quad 6.57,\quad 15.65,\quad 4,\quad 1.54,\quad 1.46,\quad 3.07,\quad 3.25,\quad 8.05$$

Exercise 4 [⋆ Insincerity in testing]

The sixth season of the American television series House has been broadcast in France since April 19 2011. In the United States, television magazines noted a drop in ratings for this sixth season, aired in 2009–2010, compared with the fifth season aired in 2008–2009. Season 5 contains 24 episodes and Season 6 contains 22. The audience figures (in millions of viewers) for these episodes were recorded.

The broadcaster wants to prove that the audience has not diminished, while the advertising agencies wants to prove that it diminished (because they are negotiating the prices).

For Season 5, the following results were obtained:

Episode	501	502	503	504	505	506	507	508	509	510	511	512
Audience	14.77	12.38	12.98	13.27	13.09	13.50	13.06	13.26	12.88	12.52	14.05	15.03

Episode	513	514	515	516	517	518	519	520	521	522	523	524
Audience	15.69	14.87	14.20	14.86	12.38	13.13	12.51	13.29	12.19	11.69	12.05	12.74

For Season 6, the following results were obtained:

Episode	601	602	603	604	605	606	607	608	609	610	611
Audience	16.50	16.50	14.44	13.74	13.50	11.65	13.31	12.67	11.95	13.25	12.24

Episode	612	613	614	615	616	617	618	619	620	621	622
Audience	14.21	13.38	13.60	12.82	11.37	10.80	10.82	10.85	9.98	9.48	11.06

Let Y be the random variable for the audience of an episode of season 5, i.e. the row "Audience" in the first tables are 24 samples $y_1, \dots, y_{24}$ taken from Y. We denote by Z and $z_1, \dots, z_{22}$ the corresponding data for season 6. We assume that $Y \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Z \sim \mathcal{N}(\mu_2, \sigma_2)$. We note that:

$$\sum_{i=1}^{24} y_i = 320.39, \qquad \sum_{i=1}^{24} y_i^2 = 4303.229, \qquad \sum_{i=1}^{22} z_i = 278.12, \qquad \sum_{i=1}^{22} z_i^2 = 3588.408.$$

The broadcaster want to accept the hypothesis $H_0 : \mu_1 = \mu_2$; while the advertising agency wants to reject $H_0' : \mu_1 \leq \mu_2$. All significance levels are taken to be $\alpha = 0.05$.

1. On the broadcaster side:
   (a) Use an F-test to reject the fact that Y and Z have the same standard deviation.
   (b) Let $y_1, \dots, y_n$ be samples for Y and $z_1, \dots, z_m$ be samples of Z. You have shown (Exercise sheet 6, exercise 5) that if n and m are big enough, then $T \sim \mathcal{N}(0,1)$ where:
   $$T = \frac{\overline{y} - \overline{z}}{\sqrt{\frac{s_y^2}{n} + \frac{s_z^2}{m}}}$$
   Use this knowledge to make a test that accepts $H_0 : m_1 = m_2$.
2. (**It is not mandatory to do these question.**) On the advertising agency side: For each episode $i = 1, \dots, 22$, the advertising agency computes $x_i = y_i - z_i$ (they toss out episodes 23 and 24 of season 5). We have $\sum_{i=1}^{22} x_i = 17.48$ and $\sum_{i=1}^{22} x_i^2 = 81.986$.
   (a) Let X be the random variable from which $x_1, \dots, x_{22}$ have been sampled. What is the distribution of X?
   (b) Show that we can accept the hypothesis that the variance of X is 3.
   (c) Remember that H_0' is $m \geq 0$. Perform a Student test to reject H_0'.
3. (**It is not mandatory to do this question.**) Both companies cheated a bit with the numbers, explain how.

Exercise 5 [⋆ • Friedman test]

We are going to design a new test, which is a prolongation of the median testing method.

$n \geq 1$ independent wine judges are testing $k \geq 2$ wines. You want to know if there exists a consistent ranking of the k wines or not. Your null hypothesis is H_0 : "all $k!$ ways of ordering the wines are equally likely".

Let R_{ij} denote the rank of the wine j according to the judge i, and define the rank sums $R_j = \sum_{i=1}^{n} R_{ij}$, for $j = 1, \dots, k$. The Friedman statistic is defined by $Q_n = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$.

1. Compute $\mathbb{E}R_{ij}$ and $\mathbb{E}R_j$. Compute $\mathbb{V}R_{ij}$ and $\mathbb{V}R_j$.
2. Show that for fixed i, $\mathrm{Cov}(R_{ij}, R_{i\ell}) = -\frac{k+1}{12}$, for $j \neq \ell$.
3. Define $Z_j = \frac{R_j - n(k+1)/2}{\sqrt{n(k^2-1)/12}}$.
   Show that the vector $(Z_1, \dots, Z_k)$ satisfies $\sum_{j=1}^{k} Z_j = 0$.
4. Using the multivariate central limit theorem, prove that under H_0, $(Z_1, \dots, Z_k)$ converges in distribution to a centered Gaussian vector supported on the hyperplane $H = \left\{ x \in \mathbb{R}^k : \sum_{j=1}^{k} x_j = 0 \right\}$. Determine its covariance matrix.
5. Show that $Q_n = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(R_j - \frac{n(k+1)}{2} \right)^2$.
6. Deduce that, assuming H_0, we get: $Q_n \overset{distrib.}{\longrightarrow} \chi^2_{k-1}$.
7. Explain why the Friedman test rejects H_0 for large values of Q_n.

Introduction to Statistics – Solutions Sheet 7

Exercise 1 [⋆ Z-test]

Let $X_1, \dots, X_{55}$ denote the reading scores of the 55 students from Osnabrück. Under the null hypothesis H_0, the variables $X_1, \dots, X_{55}$ are i.i.d. with distribution $\mathcal{N}(100, 12^2)$.

We denote by $\overline{X} = \frac{1}{55}\sum_{i=1}^{55} X_i$ the sample mean. Since the Gaussian distribution is stable under averaging, we have under H_0:

$$\overline{X} \sim \mathcal{N}\left(100, \frac{12^2}{55}\right).$$

We introduce the statistic $Z = \frac{\sqrt{55}}{12}(\overline{X} - 100)$.

If $Y \sim \mathcal{N}(m, \sigma^2)$, then $\frac{Y-m}{\sigma} \sim \mathcal{N}(0,1)$. Applying this to $\overline{X}$, we obtain under H_0: $Z \sim \mathcal{N}(0,1)$.

We want to test whether the average score in Osnabrück is abnormally high. For a significance level α, let z_α be such that $\mathbb{P}(N \leq z_\alpha) = 1 - \alpha$, where $N \sim \mathcal{N}(0,1)$.

The rejection region is therefore $Z > z_\alpha$.

In the present situation, $\overline{X} = 104$. Hence $z = \frac{\sqrt{55}}{12}(104 - 100) = \frac{\sqrt{55}}{3} \approx 2.47$.

For example (read from the lookup tables):

- if $\alpha = 5\%$, then $z_{0.05} \approx 1.645$,
- if $\alpha = 1\%$, then $z_{0.01} \approx 2.326$.

Since $2.47 > 1.645$ and $2.47 > 2.326$, we reject H_0 both at the 5% level and at the 1% level. Thus the observed average score of the students from Osnabrück appears to be significantly higher than the population average.

Exercise 2 [⋆ Exponential test]

Let $x_1, \dots, x_{10}$ be the observed data.

Remember that $\mathbb{E}X = \frac{1}{\lambda}$ and $\mathbb{V}X = \frac{1}{\lambda^2}$.

1. If we believe the null hypothesis H_0: $\lambda = 0.2$, then we should have $\overline{x} \in [\frac{1}{\lambda} - \frac{\sqrt{1/\lambda^2}}{\sqrt{n\alpha}}, \frac{1}{\lambda} - \frac{\sqrt{1/\lambda^2}}{\sqrt{n\alpha}}]$, using the Chebyshev inequality for building the confidence interval, as in the course. Here $\overline{x} = 4.71$, and $\frac{1}{\sqrt{n\alpha}} = \frac{1}{\sqrt{10 \cdot 1/20}} = \sqrt{2} \approx 1.41$. So $\frac{1}{\lambda} = 5$ yields the test $\overline{x} \in [5 - 5\sqrt{2},\ 5 + 5\sqrt{2}]$. Since $\overline{x} \in [5 - 5\sqrt{2},\ 5 + 5\sqrt{2}]$, we cannot reject that $\lambda = 0.2$.
2. The computations are similar, replacing $\frac{1}{\lambda} = 5$ by $\frac{1}{\lambda} = 4$. Since $\overline{x} = 4.17 \in [4 - 4\sqrt{2},\ 4 + 4\sqrt{2}]$, we cannot reject that $\lambda = 0.25$.
3. To be sure to reject one of the hypotheses $\lambda = 0.2$ or $\lambda = 0.25$, we need that the intervals of acceptance that they define do not intersect, i.e.

$$\left[5 - \frac{5}{\sqrt{n\alpha}},\ 5 + \frac{5}{\sqrt{n\alpha}}\right] \cap \left[4 - \frac{4}{\sqrt{n\alpha}},\ 4 + \frac{4}{\sqrt{n\alpha}}\right] = \emptyset\,.$$

This is equivalent to $4 + \frac{4}{\sqrt{n\alpha}} < 5 - \frac{5}{\sqrt{n\alpha}}$. This yields: $n\alpha > 81$ In the case of $\alpha = 0.05$, we get finally $n = 1620$.

Note: I made the data in R, using $\lambda = 0.2$.

Exercise 3 [⋆ Exact median test]

1. Assume that $H_0 : \mathrm{med}(X) = a$.
   By definition of the median, $\mathbb{P}(X < a) = \mathbb{P}(X > a) = \frac{1}{2}$.
   For each i, define $Y_i = \mathbf{1}_{\{x_i < a\}}$. Then the variables $Y_1, \dots, Y_n$ are i.i.d. Bernoulli random variables with parameter $\frac{1}{2}$.
   Since $N = \sum_{i=1}^{n} Y_i$, we conclude that under H_0, $N \sim \mathrm{Bin}\left(n, \frac{1}{2}\right)$.

2. Under H_0, the distribution of N is symmetric around $\frac{n}{2}$. Large or small values of N indicate that the median is not equal to a.
Let $Y \sim \text{Bin}\left(n, \frac{1}{2}\right)$, and let $b_{n,\theta}$ denote the θ-quantile of this distribution: $\mathbb{P}(Y \leq b_{n,\theta}) = \theta$.
We use a two-sided test of significance level α.
We reject H_0 whenever
$$N \leq b_{n,\alpha/2} \qquad \text{or} \qquad N \geq n - b_{n,\alpha/2}.$$
Indeed, by symmetry, $\mathbb{P}\big(N \leq b_{n,\alpha/2}\big) = \mathbb{P}\big(N \geq n - b_{n,\alpha/2}\big) = \frac{\alpha}{2}$. Hence the total probability of rejection under H_0 is α.
Equivalently, we accept H_0 when
$$b_{n,\alpha/2} < N < n - b_{n,\alpha/2}.$$

3. We now test whether the data are compatible with $\lambda = 0.2$. For an exponential distribution $\text{Exp}(\lambda)$, the median is $m = \frac{\ln 2}{\lambda}$. Under $H_0 : \lambda = 0.2$, the median is therefore $m = \frac{\ln 2}{0.2}$. Since $0.693 \leq \ln 2 \leq 0.694$, we obtain $3.465 \leq m \leq 3.47$.
The observations smaller than 3.47 are 1.63, 1.88, 1.54, 1.46, 3.07, 3.25. Hence $N = 6$. Under H_0, $N \sim \text{Bin}\left(10, \frac{1}{2}\right)$. For a significance level $\alpha = 0.05$, we have $\alpha/2 = 0.025$.
Thanks to the lookup table, we get $b_{10,\,0.025} = 1$ (we take). The rejection region is therefore ($N \leq 1$ or $N \geq 9$). Since $N = 6$, we do not reject the null hypothesis.
Thus the sample is compatible with an exponential distribution of parameter $\lambda = 0.2$.

Exercise 4 [⋆ Insincerity in testing]

1. (a) We first compute the empirical means: $\overline{y} = \frac{320.39}{24} \simeq 13.35$, $\overline{z} = \frac{278.12}{22} \simeq 12.64$. The empirical variances are $s_y^2 = \frac{1}{23}\left(4303.229 - \frac{320.39^2}{24}\right) \simeq 1.07$, and $s_z^2 = \frac{1}{21}\left(3588.408 - \frac{278.12^2}{22}\right) \simeq 3.56$.
We test $H_0 : \sigma_1^2 = \sigma_2^2$. The Fisher statistic is $F = \frac{s_z^2}{s_y^2} \simeq \frac{3.56}{1.07} \simeq 3.33$. Under H_0, $F \sim \mathcal{F}(23, 21)$.
From the table, $f_{23,21,0.975} = 2.380$. Since $3.33 > 2.380$, we reject H_0 at significance level 0.05. Hence the variances cannot be considered equal.
   (b) Since the sample sizes "are large enough" (in reality, they are not: 22 and 24 episodes is not large), we use $T = \frac{\overline{y} - \overline{z}}{\sqrt{\frac{s_y^2}{24} + \frac{s_z^2}{22}}}$. We obtain $T \simeq \frac{13.35 - 12.64}{\sqrt{\frac{1.07}{24} + \frac{3.56}{22}}} \simeq \frac{0.71}{0.455} \simeq 1.56$.
Under $H_0 : \mu_1 = \mu_2$, $T \sim \mathcal{N}(0, 1)$. At level $\alpha = 0.05$, the acceptance region for a two-sided test is $[-1.96, 1.96]$. Since $1.56 \in [-1.96, 1.96]$, we accept the hypothesis $H_0 : \mu_1 = \mu_2$.
2. (a) For each i, $x_i = y_i - z_i$. Assuming independence and normality, $X = Y - Z$ is again Gaussian because sums of Gaussians are Gaussian. More precisely, $X \sim \mathcal{N}(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
   (b) We compute $\overline{x} = \frac{17.48}{22} \simeq 0.795$. The empirical variance is $s_x^2 = \frac{1}{21}\left(81.986 - \frac{17.48^2}{22}\right) \simeq 3.26$.
We want to test $H_0 : \sigma^2 = 3$. The statistic $Q = \frac{(22-1)s_x^2}{3}$ follows approximately a chi-square distribution with 21 degrees of freedom under H_0. We obtain $Q \simeq \frac{21 \times 3.26}{3} \simeq 22.8$. Since this value is close to the expected value 21 of a χ_{21}^2 distribution, there is no reason to reject H_0 (one would need a lookup table for χ_{21}^2 and check that its α-quantile is smaller than 22.8, but you are a believer). Thus we accept the hypothesis that the variance of X is 3.
   (c) We test $H_0' : \mu \geq 0$ against $H_1 : \mu < 0$.
The Student statistic is $T = \frac{\overline{x}}{s_x/\sqrt{22}} \simeq \frac{0.795}{\sqrt{3.26}/\sqrt{22}} \simeq 2.06$. Under H_0', $T \sim t_{21}$.
For a one-sided test at level 0.05, the critical value is -1.721. Note that we only test $\mu \geq 0$, so we look for the 0.95-quantile of t_{21}, not for the 0.975-quantile: yes, Student's t distribution is symmetric, but our hypothesis is not.
We reject H_0' only if $T < -1.721$. Since $2.06 > -1.721$, we cannot reject H_0'.
Thus the advertising agency fails to prove that season 6 had a larger mean audience.

3. Both companies manipulated the data slightly.
   The broadcaster performed an unpaired comparison between the two seasons, treating the observations as independent samples. Not only are the episodes not independent (they literally follows each other's narratives), they are also not identically distributed since the hype on the first and last episodes is more important. Moreover, the broadcaster performed a test which is nonsensical when n and m (the number of episodes in the two seasons) are small: 22 and 24 episodes is not a large number.
   The advertising agency instead paired episode i of season 5 with episode i of season 6 and considered the differences $x_i = y_i - z_i$. This pairing artificially reduces the variance and gives a more favorable test statistic. Moreover, the agency discarded episodes 23 and 24 of season 5 in order to keep only 22 paired observations: these are the more hyped episodes, so you are sure that the test is messed up. One would need to try to discard other 2 episodes and see if the result changes, it is perhaps (I'm too lazy to test everything) possible to reject H_0' this just by selecting the correct 22 episodes, without changing the method.

Exercise 5 [⋆ Friedman test]

1. Under H_0, for a fixed judge i, the random variable R_{ij} is uniformly distributed on $\{1, \dots, k\}$. Hence $\mathbb{E}[R_{ij}] = \frac{1}{k}\sum_{r=1}^{k} r = \frac{k+1}{2}$, and $\mathbb{E}[R_{ij}^2] = \frac{1}{k}\sum_{r=1}^{k} r^2 = \frac{(k+1)(2k+1)}{6}$. Therefore $\mathrm{Var}(R_{ij}) = \frac{(k+1)(2k+1)}{6} - \left(\frac{k+1}{2}\right)^2 = \frac{k^2-1}{12}$.
   Since the judges are independent, $R_j = \sum_{i=1}^{n} R_{ij}$, so $\mathbb{E}[R_j] = n\frac{k+1}{2}$, and $\mathrm{Var}(R_j) = n\frac{k^2-1}{12}$.
2. For a fixed judge i, $\sum_{j=1}^{k} R_{ij} = \frac{k(k+1)}{2}$ is deterministic. Hence $0 = \mathrm{Var}\left(\sum_{j=1}^{k} R_{ij}\right)$. By symmetry, if $c = \mathrm{Cov}(R_{ij}, R_{i\ell}), \qquad j \neq \ell$, then $0 = k\frac{k^2-1}{12} + k(k-1)c$. Thus $c = -\frac{k^2-1}{12(k-1)} = -\frac{k+1}{12}$.
3. By definition, $Z_j = \frac{R_j - n(k+1)/2}{\sqrt{n(k^2-1)/12}}$. Now $\sum_{j=1}^{k} R_j = \sum_{i=1}^{n}\sum_{j=1}^{k} R_{ij} = n\frac{k(k+1)}{2}$. Therefore $\sum_{j=1}^{k}\left(R_j - \frac{n(k+1)}{2}\right) = 0$, and consequently $\sum_{j=1}^{k} Z_j = 0$.
4. Define the random vectors $X_i = (R_{i1}, \dots, R_{ik}), \qquad i = 1, \dots, n$. The vectors X_i are i.i.d. with finite second moments. Hence the multivariate central limit theorem gives $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where $\Sigma_{jj} = \frac{k^2-1}{12}, \qquad \Sigma_{j\ell} = -\frac{k+1}{12} \quad (j \neq \ell)$.
   After normalization, the vector $(Z_1, \dots, Z_k)$ converges to a centered Gaussian vector with covariance matrix $C_{jj} = 1, \qquad C_{j\ell} = -\frac{1}{k-1} \quad (j \neq \ell)$.
   Moreover, $\sum_{j=1}^{k} Z_j = 0$, so the limit distribution is supported on the hyperplane $H = \left\{x \in \mathbb{R}^k : \sum_{j=1}^{k} x_j = 0\right\}$, which has dimension $k-1$.
5. Expanding the definition of Q_n,

$$Q_n = \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_j^2 - 3n(k+1).$$

   Since $\sum_{j=1}^{k}\left(R_j - \frac{n(k+1)}{2}\right)^2 = \sum_{j=1}^{k} R_j^2 - n(k+1)\sum_{j=1}^{k} R_j + k\frac{n^2(k+1)^2}{4}$, and $\sum_{j=1}^{k} R_j = n\frac{k(k+1)}{2}$, we obtain $\sum_{j=1}^{k}\left(R_j - \frac{n(k+1)}{2}\right)^2 = \sum_{j=1}^{k} R_j^2 - \frac{nk(k+1)^2}{4}$. Multiplying by $\frac{12}{nk(k+1)}$ gives $Q_n = \frac{12}{nk(k+1)}\sum_{j=1}^{k}\left(R_j - \frac{n(k+1)}{2}\right)^2$.
6. Since $R_j - \frac{n(k+1)}{2} = \sqrt{\frac{n(k^2-1)}{12}}\, Z_j$, we obtain $Q_n = \frac{k-1}{k}\sum_{j=1}^{k} Z_j^2$.
   The limiting Gaussian vector lives in the $(k-1)$-dimensional hyperplane H. In an orthonormal basis of H, its coordinates are independent standard Gaussian random variables. Therefore the quadratic form $\frac{k-1}{k}\sum_{j=1}^{k} Z_j^2$ converges in distribution to χ^2_{k-1}. Hence $Q_n \xrightarrow{d} \chi^2_{k-1}$.
7. Under H_0, the rank sums R_j should remain close to their common mean $n(k+1)/2$. Large values of Q_n correspond to unusually large deviations between the rank sums and therefore

indicate that the wines are not equivalent.
Thus the Friedman test rejects H_0 for sufficiently large values of Q_n. Once a significance level α is given, one can test whether Q_n is bigger than the α-quantile of the distribution χ^2_{k-1} or not. Note that the distribution to test against (i.e. χ^2_{k-1}) does not depend on n, but Q_n does: the more judges there are, the more precise the value of Q_n gets, which makes it more likely to be rejected if we were wrong in assuming that H_0 holds.

Exercise 6 [Testing whether something is normally distributed]
We study, on a sample of 100 calls, the waiting time (in seconds) before a connection is established in a telephone information center.

Waiting time	2–6	6–10	10–12	12–14	14–18	18–26	26–34
Frequency	8	14	9	11	30	16	12

Can this distribution be considered normally distributed (perform a χ^2 test)?

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 8

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Table 1: **Mann–Whitney critical value for** $\alpha = 0.05$

$n1 \backslash n2$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2							0	0	0	0	1	1	1	1	1	2	2	2	2
3				0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4			0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
5		0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6		1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7		1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	69	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Table 2: **Wilks' Lambda critical value for** $\alpha = 0.05$ **and for** $p = 2$

n	m									
↓	1	2	3	4	5	6	7	8	9	10
11	.549	.404	.316	.255	.212	.179	.153	.133	.116	.102
12	.580	.437	.348	.286	.239	.204	.176	.154	.136	.120
13	.607	.467	.378	.314	.266	.229	.199	.175	.155	.138
14	.631	.495	.405	.340	.291	.252	.221	.195	.174	.156
15	.652	.519	.431	.365	.315	.275	.242	.215	.193	.174
16	.671	.542	.454	.389	.337	.296	.263	.235	.211	.191
17	.688	.562	.476	.410	.359	.317	.282	.254	.229	.208
18	.703	.581	.496	.431	.379	.337	.301	.272	.246	.225
19	.717	.598	.515	.450	.398	.355	.320	.289	.263	.241
20	.730	.614	.532	.468	.416	.373	.337	.306	.279	.256
30	.813	.725	.657	.601	.553	.512	.475	.443	.414	.388
40	.858	.786	.730	.682	.639	.602	.568	.537	.509	.484
60	.903	.853	.811	.774	.741	.710	.682	.656	.632	.609
80	.927	.887	.854	.825	.798	.772	.749	.727	.706	.686
100	.941	.909	.882	.857	.834	.813	.793	.774	.755	.738
120	.951	.924	.900	.879	.860	.841	.823	.807	.791	.775

Exercise 1 [⋆ Fisher–Snedecor distribution]

Let X be a random variable having a Fisher–Snedecor distribution with ν_1 and ν_2 degrees of freedom (the distribution involved in the F-test). Its probability density function is

$$f(x) = c \cdot x^{\nu_1/2-1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-(\nu_1+\nu_2)/2}, \qquad x > 0.$$

where $c > 0$ is a constant (with respect to x, not with respect to ν_1 and ν_2) such that $\int_0^{+\infty} f = 1$.

Assume throughout that $2 < \nu_1$ and $2 \leq \nu_2$.

1. Determine the median of the Fisher–Snedecor distribution, for $\nu_1 = \nu_2$.
2. Determine the mode of Fisher–Snedecor distribution.
3. (**not mandatory**) Determine the mean of the Fisher–Snedecor distribution.

If you need for question 3.:

$$c = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}$$

Exercise 2 [⋆ One-way MANOVA using Wilks' Λ]

Two groups of students are compared with respect to two response variables:

- Y_1: score on a mathematics test,
- Y_2: score on a statistics test.

The sample sizes are $n_1 = n_2 = 10$. The score of each student is assumed to be (approximately) normally distributed. A one-way MANOVA is performed to test whether the mean vectors of the two groups are equal: $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. The between-group and within-group sum-of-squares-and-cross-product matrices are

$$B = \begin{pmatrix} 8 & 4 \\ 4 & 2 \end{pmatrix}, \qquad W = \begin{pmatrix} 20 & 5 \\ 5 & 10 \end{pmatrix}.$$

1. Are the matrices B and W of the sizes you were expecting?
2. Compute Wilks' statistic[3], that is: $\Lambda = \frac{\det(W)}{\det(W+B)}$.
3. Using your value of Λ and the lookup table, decide whether there is evidence against H_0.
4. Interpret your conclusion in the context of the problem.

Exercise 3 [⋆ • Wilk's distribution]

Let $\Lambda(p, m, n)$ be Wilk's lambda distribution, that is the distribution[3] of $\frac{\det(W)}{\det(W+B)}$ where:

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}^{(i)}) \otimes (\boldsymbol{x}_j^{(i)} - \overline{\boldsymbol{x}}^{(i)}) \qquad B = \sum_{i=1}^{m} n_i (\overline{\boldsymbol{x}}^{(i)} - \overline{x}) \otimes (\overline{\boldsymbol{x}}^{(i)} - \overline{x})$$

Where the parameter p is the length of each vector $\boldsymbol{x}_j^{(i)}$, the parameter n is $\sum_{i=1} n_i$, and the parameter m is the number of different variables (i.e. the range of i in the above sums).

1. Let $\lambda_\alpha(p, m, n)$ be the α-quantile of $\Lambda(p, m, n)$. Justify (without rigorous mathematical proof) that $\lambda_\alpha(p, m, n)$ is: increasing when α increases; decreasing when p increases; decreasing when m increases; increasing when n increases.
2. (**very very hard, not mandatory**) Show that:

$$\Lambda(p, m, n) \sim \Lambda(n,\ m + n - p,\ p)$$

3. (**not mandatory**) Show that, for p, m, n sufficiently large, we have the following approximation:

$$\left(\frac{p - n + 1}{2} - m\right) \log \Lambda(p, m, n) \approx \chi^2(np)$$

[3] Be carefull, there might be a small misprint in the script of the course: the denominator should be $\det(W + B)$ not $\det(W) + \det(B)$.

Exercise 4 [⋆ Wilcoxon–Mann–Whitney test]

A researcher wants to compare the effectiveness of two teaching methods on student performance. Two independent groups of students are taught using different methods, and their test scores are recorded below (from 0 points to 100 points). Assume that the observations are independent.

Method A	**Method B**
72	85
68	90
75	78
80	88
77	82
70	91

1. State the null hypotheses for a two-sided Wilcoxon–Mann–Whitney test.
2. Combine all observations from both groups and rank them from smallest to largest. (In case of ties, say between the 4^{th} and the 5^{th} ranked values, then all involved values receive the average rank, here 4.5.)
3. Compute the sum of ranks for Method A and Method B, denoted R_A and R_B.
4. Calculate the Wilcoxon rank-sum statistics and determine which one is the smallest:
$$W_A = R_A - \frac{n_A(n_A+1)}{2}, \qquad W_B = R_B - \frac{n_B(n_B+1)}{2}.$$
5. Using a significance level of $\alpha = 0.05$, decide whether there is evidence of a difference between the two teaching methods (compare Method A "inside" Method B). Use either a critical-value table for the Wilcoxon–Mann–Whitney test, or the normal approximation. Interpret the result in the context of the study.
6. Calculate the effect size using
$$r = \frac{|Z|}{\sqrt{N}},$$
where N is the total sample size and Z is the centralized and standardize version of W_A. Interpret the magnitude of the effect using the following guidelines: small effect: $r \approx 0.1$; medium effect: $r \approx 0.3$; large effect: $r \approx 0.5$.

Exercise 5 [⋆ Henry's straight line and normal probability paper]

We want to know whether some samples are normally distributed or not (since a lot of tests are based on this hypothesis), and to determine the mean and the standard deviation if they are. To this end, we use the method of Henry's straight line (of course, this kind of test is now done far more efficiently using computers, but when it was invented, around 1880, it was simple and powerful).

During an exam (graded from 0 to 20):

- 10% of the students got less than 4 points;
- 30% of the students got less than 8 points;
- 60% of the students got less than 12 points;
- 80% of the students got less than 16 points;

We want to check whether the grades are normally distributed and what are the mean and the standard deviation.

You are given (next page) a normal probability paper: the right vertical axis is labeled regularly by t from -3 to $+3$, while the left axis is labeled by the $\mathbb{P}(X \leq t)$ for $X \sim \mathcal{N}(0,1)$ (written in %).

1. Add an horizontal axis to this paper, labeled regularly from 0 to 20, and place the points $(x, q(x))$ where $x \in \{4, 8, 12, 16\}$ and $q(x)$ is the proportion of students who got less than x points. Use the left vertical axis.
2. Justify that, under that assumption that the data is normally distributed, your data points should (approximately) be on a line, on the normal probability paper.
3. Are they on a line? If yes, draw this line by hand.

4. The mean μ of a normal distribution is also its 50%-quantile, and we have (approximately) that $\mu - \sigma$ is the 16%-quantile while $\mu + \sigma$ is the 68%-quantile. Using the normal probability paper, deduce μ and σ such that you cannot reject the hypothesis that $X \sim \mathcal{N}(\mu, \sigma^2)$.

Exercise 6 [Shuffle testing]

Shuffle testing is a way of creating tests rather than a single test. We illustrate it on an example.

Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be two series of independent samples. The null hypothesis is that both series come from the same distribution and are independent from each other. Let $\Delta_{obs} = \overline{x} - \overline{y}$ be the observed difference between the expectations of the two distributions.

A *shuffle* is a way of splitting $(x_1, \ldots, x_m, y_1, \ldots, y_n)$ into a group of m elements $(x'_1, \ldots, x'_m)$ and a group of n elements $(y'_1, \ldots, y'_n)$, i.e. $(x'_1, \ldots, x'_m, y'_1, \ldots, y'_n) = (x_1, \ldots, x_m, y_1, \ldots, y_n)$ but they are not necessarily in the same order.

1. Depending on m and n, how many shuffles are there?
2. For a shuffle σ, let $\overline{x}'$ and $\overline{y}'$ be the associated means of the shuffled groups. What is the distribution of $\Delta' = \overline{x}' - \overline{y}'$ under the null hypothesis?
3. Write a test for the null hypothesis, using $p = \frac{\text{number of shuffles such that } |\Delta'| \geq |\Delta_{obs}|}{\text{number of shuffles}}$, with significance level α.
4. In practice, it would be too long to test all shuffles. Hence, we randomly pick s shuffles and compute $p_s = \frac{1+\text{number of shuffles such that } |\Delta'| \geq |\Delta_{obs}|}{1+s}$. Write a test for the null hypothesis, using p_s instead.

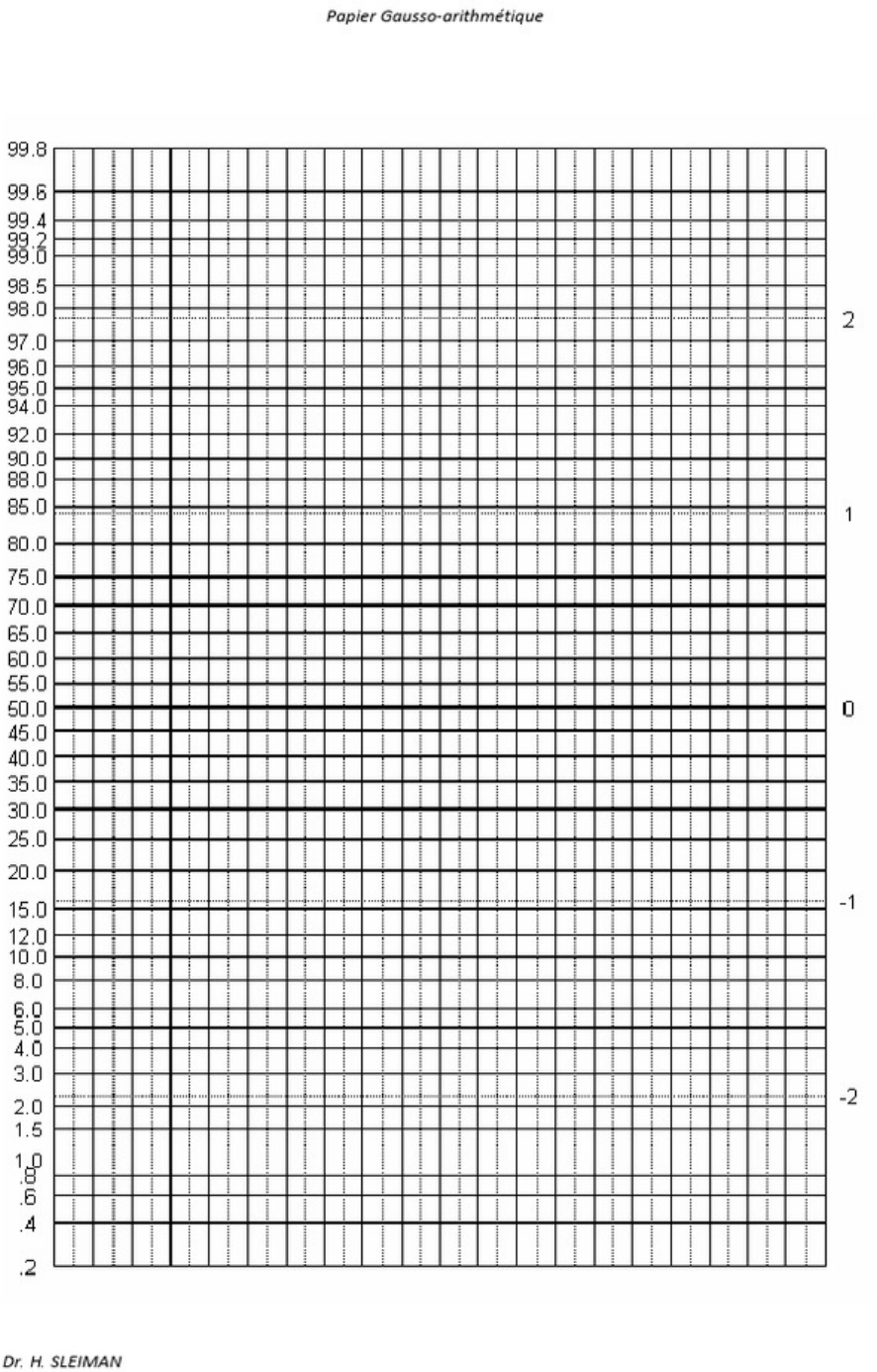


Figure 2: Normal probability paper

Introduction to Statistics – Solutions Sheet 8

Exercise 1 [⋆ Fisher–Snedecor distribution]

Let

$$f(x) = c\, x^{\nu_1/2-1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-(\nu_1+\nu_2)/2}, \qquad x > 0,$$

with

$$c = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}.$$

1. Thanks to the course, we know that $X = \frac{U_1/\nu_1}{U_2/\nu_2}$ follows a Fisher–Snedecor distribution when $U_1 \sim \chi^2(d_1)$ and $U_2 \sim \chi^2(\nu_2)$. In particular, with $\nu = \nu_1 = \nu_2$, we get:

$$\mathbb{P}(X \le t) = \frac{1}{2} \quad \Leftrightarrow \quad \mathbb{P}\left(\frac{U_1}{\nu} \le t \cdot \frac{U_2}{\nu}\right) = \mathbb{P}(U_1 \le t \cdot U_2) = \frac{1}{2}$$

   where both U_1 and U_2 are $\chi^2(\nu)$ distributed. In particular, since they follow the same distribution, it is as likely that $U_1 \le U_2$ as it is that $U_1 \ge U_2$, so for $t = 1$ we get $\mathbb{P}(U_1 \le t \cdot U_2) = \frac{1}{2}$. Said differently, the median of X is 1.
2. Since $f(x) > 0$, maximizing f is equivalent to maximizing

$$\ell(x) := \log f(x) = \left(\frac{\nu_1}{2} - 1\right) \log x - \frac{\nu_1 + \nu_2}{2} \log\left(1 + \frac{\nu_1}{\nu_2} x\right) + \log c.$$

   Differentiating,

$$\ell'(x) = \frac{\frac{\nu_1}{2} - 1}{x} - \frac{\nu_1 + \nu_2}{2} \frac{\nu_1/\nu_2}{1 + (\nu_1/\nu_2)x}.$$

   Setting $\ell'(x) = 0$ gives

$$\frac{\nu_1 - 2}{x} = \frac{\nu_1(\nu_1 + \nu_2)}{\nu_2 + \nu_1 x}.$$

   Hence

$$(\nu_1 - 2)(\nu_2 + \nu_1 x) = \nu_1(\nu_1 + \nu_2)x,$$

   and therefore

$$(\nu_1 - 2)\nu_2 = \nu_1(\nu_2 + 2)x.$$

   Thus the unique critical point is

$$x_{\text{mode}} = \frac{\nu_1 - 2}{\nu_1} \frac{\nu_2}{\nu_2 + 2}.$$

   Since $f(x) \to 0$ as $x \to 0$ (because $\nu_1 > 2$) and $f(x) \to 0$ as $x \to \infty$, this critical point is necessarily the global maximum, i.e. the mode.
3. We compute

$$\mathbb{E}[X] = \int_0^\infty x\, f(x)\, dx.$$

   Using the change of variable given by:

$$t = \frac{\nu_1 x}{\nu_2 + \nu_1 x}, \qquad x = \frac{\nu_2}{\nu_1} \frac{t}{1 - t},$$

   we get that

$$f(x)\, dx = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, t^{a-1}(1-t)^{b-1}\, dt, \qquad \text{where} \quad a = \frac{\nu_1}{2}, \quad b = \frac{\nu_2}{2}.$$

Hence

$$\mathbb{E}[X] = \frac{\nu_2}{\nu_1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 t^a (1-t)^{b-2}\, dt.$$

The integral is a beta integral:

$$\int_0^1 t^a (1-t)^{b-2}\, dt = B(a+1, b-1) = \frac{\Gamma(a+1)\Gamma(b-1)}{\Gamma(a+b)}.$$

Therefore

$$\mathbb{E}[X] = \frac{\nu_2}{\nu_1} \frac{\Gamma(a+1)}{\Gamma(a)} \frac{\Gamma(b-1)}{\Gamma(b)}.$$

Using

$$\Gamma(a+1) = a\,\Gamma(a), \qquad \Gamma(b) = (b-1)\Gamma(b-1),$$

we obtain

$$\mathbb{E}[X] = \frac{\nu_2}{\nu_1} \frac{a}{b-1}.$$

Substituting

$$a = \frac{\nu_1}{2}, \qquad b = \frac{\nu_2}{2},$$

gives

$$\mathbb{E}[X] = \frac{\nu_2}{\nu_1} \frac{\nu_1/2}{(\nu_2-2)/2} = \frac{\nu_2}{\nu_2-2}.$$

Thus, for $\nu_2 > 2$,

$$\mathbb{E}[X] = \frac{\nu_2}{\nu_2-2}.$$

If $\nu_2 = 2$, the integral diverges and the mean does not exist.

Exercise 2 [⋆ One-way MANOVA using Wilks' Λ]

There are $p = 2$ response variables (Y_1 and Y_2), and two groups with sample sizes $n_1 = n_2 = 10$.

1. Since there are $p = 2$ response variables, both the between-group and within-group sum-of-squares-and-cross-product matrices must be 2×2 matrices. Therefore the sizes of

$$B = \begin{pmatrix} 8 & 4 \\ 4 & 2 \end{pmatrix}, \qquad W = \begin{pmatrix} 20 & 5 \\ 5 & 10 \end{pmatrix}$$

   are exactly what we expect.
2. We first compute

$$\det(W) = \begin{vmatrix} 20 & 5 \\ 5 & 10 \end{vmatrix} = 20 \cdot 10 - 5 \cdot 5 = 200 - 25 = 175.$$

   Next, $W + B = \begin{pmatrix} 28 & 9 \\ 9 & 12 \end{pmatrix}$, so

$$\det(W+B) = \begin{vmatrix} 28 & 9 \\ 9 & 12 \end{vmatrix} = 28 \cdot 12 - 9 \cdot 9 = 336 - 81 = 255.$$

   Hence Wilks' statistic is

$$\Lambda = \frac{\det(W)}{\det(W+B)} = \frac{175}{255} = \frac{35}{51} \approx 0.686.$$

3. Wilks' statistic takes values between 0 and 1. Small values of Λ indicate stronger evidence against H_0 (if the value is 1, then B is "negligible" against W, i.e. all the variance is explained by the variance inside the groups, not by the variance between the groups).
   We obtained $\Lambda \approx 0.686$. Using the Wilks' Λ lookup table (with $p = 2$ variables, $m = 2$ groups, and $N = 10 + 10 = 20$ observations), the critical value is 0.614. Since $\Lambda > 0.614$, we cannot reject H_0.
4. We fail to reject the null hypothesis that the two groups have the same mean vector. In the context of the problem, that mean we should think that the score on a math test and on a statistic test are equality distributed.

Exercise 3 [⋆ Wilk's distribution]

1. When α increases, then $\lambda_\alpha(p, m, n)$ increases because we are talking about quantiles: the more you want, the more you need. When n increases (i.e. we take more and more samples), then B (the between-group) should become statistically insignificant, so $\det(W + B) \approx \det(W)$, that is $\Lambda \approx 1$, and in particular $\lambda_\alpha(p, m, n)$ increases. On the opposite, if m increases, that is if we have a fixed number of samples but we split them into more and more groups: the between-group variance shall increase, thus $\det(W + B)$ becomes bigger and bigger, so Λ decreases and $\lambda_\alpha(p, m, n)$ decreases. When p (i.e. the dimension of the problem) increases, then the reason of the decrease is slightly different: the higher the dimension, the higher the values of the determinant; we always have that $\det(W + B)$ is bigger than $\det(W)$ whatever the dimension, but this "bigger-ness" is worsen in higher dimensions, so Λ decreases, so $\lambda_\alpha(p, m, n)$ decreases.
2. That is very complicated...
   Let $\in \mathbb{R}^{m\times p}$, $Y \in \mathbb{R}^{n\times p}$ have independent standard Gaussian entries, and set $W = X^T X$, $B = Y^T Y$. Then
   $$\Lambda(p, m, n) \overset{d}{=} \frac{\det(X^T X)}{\det(X^T X + Y^T Y)}.$$
   Let
   $$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \in \mathbb{R}^{(m+n)\times p},$$
   so that $X^T X + Y^T Y = Z^T Z$. Write the QR decomposition $Z = UR$, where U has orthonormal columns and R is invertible. If P denotes the orthogonal projection onto the first m coordinates of $\mathbb{R}^{m+n}$, then $X = PZ$, hence
   $$\Lambda = \frac{\det(R^T U^T P U R)}{\det(R^T R)} = \det(U^T P U).$$
   The columns of U form an orthonormal basis of the random p-dimensional subspace $S = \mathrm{span}(Z) \subset \mathbb{R}^{m+n}$. Since Z is Gaussian, S is uniformly distributed among all p-dimensional subspaces of $\mathbb{R}^{m+n}$. Therefore Λ depends only on this random subspace.
   Let V be an orthonormal basis of $S^\perp$, and let $Q = I_n - P$, the projection onto the last n coordinates. A standard identity for complementary subspaces gives $\det(U^T P U) = \det(V^T Q V)$. (Indeed, the nonzero eigenvalues of $U^T P U$ and $V^T Q V$ coincide, being the squared cosines of the principal angles between S and $\mathrm{Im}(P)$.)
   Now $S^\perp$ is a uniformly distributed $(m+n-p)$-dimensional subspace. Hence the right-hand side has the same distribution as the Wilks statistic obtained from a random $(m+n-p)$-dimensional subspace and a projection of rank n, namely $\Lambda(m + n - p, n, m)$. Thus
   $$\Lambda(p, m, n) \overset{d}{=} \Lambda(m + n - p, n, m).$$
   Finally, exchanging the first m and last n coordinates leaves the Gaussian model invariant, so
   $$\Lambda(m + n - p, n, m) \overset{d}{=} \Lambda(n, m + n - p, p).$$
3. This is Bartlett χ^2 approximation, see the literature.

Exercise 4 [⋆ Wilcoxon–Mann–Whitney test]

Let $n_A = n_B = 6$, so the total sample size is $N = n_A + n_B = 12$.

1. For a two-sided Wilcoxon–Mann–Whitney test, the null hypothesis is that the two populations have the same distribution.
2. Combining the observations from both groups and ordering them from smallest to largest gives

$$68,\ 70,\ 72,\ 75,\ 77,\ 78,\ 80,\ 82,\ 85,\ 88,\ 90,\ 91.$$

There are no ties. The ranks are therefore

Score	Method	Rank
68	A	1
70	A	2
72	A	3
75	A	4
77	A	5
78	B	6
80	A	7
82	B	8
85	B	9
88	B	10
90	B	11
91	B	12

3. For Method A, $R_A = 1 + 2 + 3 + 4 + 5 + 7 = 22$.
   For Method B, $R_B = 6 + 8 + 9 + 10 + 11 + 12 = 56$.
   As a check, $R_A + R_B = 22 + 56 = 78 = \frac{12(12+1)}{2}$.
4. Using $n_A = n_B = 6$, we get $W_A = 22 - 21 = 1$, and $W_B = 56 - 21 = 35$.
   Therefore $W_A < W_B$ and $\min(W_A, W_B) = 1$.
5. Using the Mann–Whitney lookup table provided, we get for $n_A = n_B = 6$ the critical value 5. Since $W_A = 1 < 5$, we reject H_0.
   Using the normal approximation, we need to check whether R_A is distributed according to $\mathcal{N}(\mu, \sigma^2)$ where $\mu = \frac{1}{2} n_A(n_A + n_B + 1) = 39$ and $\sigma^2 = \frac{1}{12} n_A n_B (n_A + n_B + 1) = 39$. We have $\frac{22-39}{\sqrt{39}} \approx 2.722$. For a two-sided test, we have that $\mathbb{P}(|N| \leq 2.722) \approx 0.0065$ where $N \sim \mathcal{N}(0, 1)$. Since $0.0065 < 0.05$, we reject H_0.
   Consequently, there is statistically significant evidence that the two teaching methods produce different score distributions. Since the observations from Method B receive substantially larger ranks, Method B appears to lead to higher test scores than Method A.
6. The effect size is

$$r = \frac{|Z|}{\sqrt{N}} = \frac{2.722}{\sqrt{12}} \approx \frac{2.722}{3.464} \approx 0.786.$$

According to the guidelines, $r \approx 0.76 > 0.5$, which corresponds to a large effect.
Thus, the difference between the two teaching methods is not only statistically significant, but also practically large. Method B appears to have a strong positive impact on student performance compared with Method A.

Exercise 5 [⋆ Henry's straight line and normal probability paper]

We denote by X the exam grade.

1. We first convert the empirical proportions into points on the normal probability paper:

$$q(4) = 0.10, \quad q(8) = 0.30, \quad q(12) = 0.60, \quad q(16) = 0.80.$$

We therefore plot the points

$$(4, 10\%),\ (8, 30\%),\ (12, 60\%),\ (16, 80\%),$$

on the normal probability paper, using the left vertical axis (percent scale).
We then add a horizontal axis from 0 to 20 and mark $4, 8, 12, 16$ on it.

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

$$\mathbb{P}(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

where Φ is the distribution function of $\mathcal{N}(0,1)$.
Let $t = \Phi^{-1}(q(x))$. Then

$$t = \frac{x-\mu}{\sigma} \quad \Longleftrightarrow \quad x = \mu + \sigma t.$$

Thus the points (x, t) lie on a straight line in the (t, x)-plane. Since the normal probability paper is designed so that the left axis corresponds to $q(x)$ and the right axis to $t = \Phi^{-1}(q(x))$, the transformed points must lie approximately on a line if the data are normal.

3. You can do it visually, or read $\Phi^{-1}(0.1)$, $\Phi^{-1}(0.3)$, $\Phi^{-1}(0.6)$ and $\Phi^{-1}(0.8)$ on the right vertical axis.
We compute the corresponding t-values:

$$\Phi^{-1}(0.10) \approx -1.28, \quad \Phi^{-1}(0.30) \approx -0.52, \quad \Phi^{-1}(0.60) \approx 0.25, \quad \Phi^{-1}(0.80) \approx 0.84.$$

We obtain the points:

$$(4, -1.28),\ (8, -0.52),\ (12, 0.25),\ (16, 0.84).$$

These points are approximately aligned (small deviations are expected due to sampling variability), so we do not reject the hypothesis of normality. A straight line can be drawn through them (for instance by least-squares or by eye).

4. μ is the value on the horizontal axis where our straight line crosses the horizontal axis at 50%. Here $\mu \approx 11$.
Similarly, 2σ is the range between the values on the horizontal axis where our straight line crossed the horizontal axis at 16% and at 68%. Here we get $2\sigma \approx 16,7 - 5,2$, so $\sigma \approx 5.7$.
Thus we cannot reject the model $X \sim \mathcal{N}(11,\ 5.7^2)$ based on Henry's straight line method.

Exercise 6 [Shuffle testing]

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 9

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Exercise 1 [⋆ t-test for uncorrelatedness]
Let $(X_1, Y_1), \ldots, (X_{12}, Y_{12})$ be independent observations from a bivariate normal distribution with correlation coefficient ρ. The sample correlation coefficient is observed to be $r = 0.58$. (Recall that the correlation is the covariance divided by the product of the standard deviations.) We wish to test $H_0 : \rho = 0$ at significance level $\alpha = 0.05$.

1. Compute the test statistic $T = r\sqrt{\frac{n-2}{1-r^2}}$.
2. Under the null hypothesis, which distribution does this statistic follow (precise its parameters)?
3. Using the critical value $t_{0.975,10} = 2.228$, decide whether H_0 should be rejected.
4. State your conclusion regarding the correlation between X and Y.

Exercise 2 [⋆ Least-squares fitting of a line]
A researcher collects the four data points $(1, 6), (2, 5), (3, 7), (4, 10)$. The researcher wishes to approximate the data by an affine function $y = \beta_1 + \beta_2 x$.

1. Write the corresponding system of equations for the unknown parameters β_1 and β_2. Explain why this system is overdetermined.
2. Introduce the residuals $r_i = y_i - (\beta_1 + \beta_2 x_i)$ for $i = 1, \ldots, 4$, and compute the sum of the squared residuals, named S.
3. Compute the partial derivatives of S with respect to β_1 and β_2, and explicit the normal equations
$$\frac{\partial S}{\partial \beta_1} = 0, \qquad \frac{\partial S}{\partial \beta_2} = 0.$$
4. Solve these normal equations and determine the least-squares line. Compute the corresponding residuals and the minimum value of S.
5. Let
$$y = \begin{pmatrix} 6 \\ 5 \\ 7 \\ 10 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$
Show that the least-squares solution satisfies $X^\top X\beta = X^\top y$, and verify that $\beta = (X^\top X)^{-1} X^\top y$. Interpret geometrically the vector $X\beta$ as the orthogonal projection of y onto the column space of X.

Exercise 3 [⋆ Least-squares fitting of a parabola]
We use the same data as in Exercise 2, that is $(1, 6), (2, 5), (3, 7), (4, 10)$, but we want to fit a parabola of the form $y = \beta_1 x^2$ through these points.

We define:
$$y = \begin{pmatrix} 6 \\ 5 \\ 7 \\ 10 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 \\ 4 \\ 9 \\ 16 \end{pmatrix}, \qquad \beta = (\beta_1)$$

1. Write the corresponding system of equations for the unknown β_1, using the provided matrices.
2. Write the minimization problem that β should satisfied, and deduce that $X\beta$ is the orthogonal projection of y onto a certain vector space.
3. Deduce that $\beta = (X^\top X)^{-1} X^\top y$, and apply numerically in order to express β.

Exercise 4 [⋆ Heteroscedasticity and weighted linear regression?]
Let $y \in \mathbb{R}^n$ be a vector of observations and let $X \in \mathbb{R}^{n\times p}$ be a design matrix with full column rank, that is we want to find the "best" $\beta \in \mathbb{R}^p$ such that $X\beta$ approximates y. Suppose that each observation is assigned a positive weight $w_i > 0$, and define the diagonal weight matrix

$$W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}.$$

The weighted least-squares estimator is defined as the vector $\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p} (y - X\beta)^\top W (y - X\beta)$.

1. Show that the objective function $f(\beta) = (y - X\beta)^\top W (y - X\beta)$ is a quadratic function of β. Expand $f(\beta)$ explicitly.
2. Compute the gradient $\nabla f(\beta)$ and show that any minimizer satisfies the *weighted normal equations* $X^\top W X\, \beta = X^\top W y$.
3. Assuming that X has full column rank and that all weights are strictly positive, prove that $X^\top W X$ is invertible.
4. Deduce that the weighted least-squares estimator is unique and is given by $\hat{\beta} = (X^\top W X)^{-1} X^\top W y$.
5. Define the fitted vector $\hat{y} = X\hat{\beta}$. Show that the residual vector $r = y - \hat{y}$ satisfies $X^\top W r = 0$. Interpret this relation geometrically.
6. Consider the weighted scalar product $\langle u, v\rangle_W = u^\top W v$. Show that $\hat{y}$ is the orthogonal projection of y onto $\mathrm{Col}(X)$ with respect to this weighted scalar product.
7. Define the matrix $P_W = X(X^\top W X)^{-1} X^\top W$. Show that $\hat{y} = P_W y$. Verify that $P_W^2 = P_W$, and explain why P_W is called a projection matrix.

Exercise 5 [⋆ Heteroscedasticity and generalized least squares]
Consider the linear model $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $\beta \in \mathbb{R}^p$, and the error vector ε satisfies $\mathbb{E}\varepsilon = 0$, $\quad \mathrm{Cov}(\varepsilon) = \Sigma$. Assume that Σ is a diagonal matrix of the form

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix},$$

where the variances σ_i^2 are not necessarily equal.

Recall that the ordinary least-squares estimator is $\hat{\beta}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top y$.

The case $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$ is called *homoscedasticity*, whereas the covariance matrix Σ above corresponds to a *heteroscedastic* model.

1. Show that $\mathbb{E}[\hat{\beta}_{\mathrm{OLS}}] = \beta$. Deduce that OLS remains unbiased even under heteroscedasticity.
2. Starting from $\hat{\beta}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top y$, show that $\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) = (X^\top X)^{-1} X^\top \Sigma X (X^\top X)^{-1}$.
3. Verify that when $\Sigma = \sigma^2 I_n$, the previous formula reduces to the classical expression $\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}) = \sigma^2 (X^\top X)^{-1}$.
4. Let $W = \Sigma^{-1}$. Consider the estimator $\hat{\beta}_{\mathrm{WLS}} = (X^\top W X)^{-1} X^\top W y$. Show that $\hat{\beta}_{\mathrm{WLS}}$ is the solution of $\min_{\beta\in\mathbb{R}^p} (y - X\beta)^\top W (y - X\beta)$. (Yes, you can summon knowledge from other exercises.)
5. Define $\tilde{y} = \Sigma^{-1/2} y$, and $\tilde{X} = \Sigma^{-1/2} X$. Show that the weighted least-squares problem can be rewritten as $\min_\beta \|\tilde{y} - \tilde{X}\beta\|^2$.
6. Compute the covariance matrix of the transformed error $\tilde{\varepsilon} = \Sigma^{-1/2}\varepsilon$. What important property does this transformation restore?

Exercise 6 [Anscombe's quartet]

Dataset I		Dataset II		Dataset III		Dataset IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Table 3: Anscombe's quartet

For the four series of data points detailed in the table, we have (each quantity is not exact, but precise up to the number of indicated digits):

$$\overline{x} = 9, \quad s_x^2 = 11, \quad \overline{y} = 7.50, \quad s_y^2 = 4.125$$

$$y \approx 3.00 \cdot x + 0.500 \qquad R^2 = 0.67$$

where the second row comes from a linear regression on a linear fit.

Draw each of the four series of data points in the plane, with their fitted line, and comment the result.

Introduction to Statistics – Solutions Sheet 9

Exercise 1 [⋆ t-test for uncorrelatedness]

We have $n = 12$ and $r = 0.58$.

1. The test statistic is
$$T = r\sqrt{\frac{n-2}{1-r^2}} = 0.58\sqrt{\frac{12-2}{1-0.58^2}} \approx 2.25.$$
2. Under the null hypothesis $H_0 : \rho = 0$, the statistic
$$T = r\sqrt{\frac{n-2}{1-r^2}}$$
follows a Student t distribution with $n-2$ degrees of freedom. Since $n = 12$, $T \sim t_{10}$.
3. This is a two-sided test at significance level $\alpha = 0.05$. The critical values are
$$\pm t_{0.975,10} = \pm 2.228.$$
Since $|T_{\text{obs}}| \approx 2.25 > 2.228$, the observed statistic falls in the rejection region. Therefore, we reject H_0.
4. At the 5% significance level, there is sufficient evidence to conclude that the correlation between X and Y is nonzero. Since the sample correlation is positive ($r = 0.58$), the data indicate a statistically significant positive correlation between X and Y.

Exercise 2 [⋆ Least-squares fitting of a line]

1. The model $y = \beta_1 + \beta_2 x$ must satisfy
$$\beta_1 + \beta_2 = 6,\ \beta_1 + 2\beta_2 = 5,\ \beta_1 + 3\beta_2 = 7,\ \beta_1 + 4\beta_2 = 10.$$
There are four equations and only two unknowns. Hence the system is overdetermined. Since the four points do not lie on a common line, the system has no exact solution.
2. The residuals are
$$r_1 = 6 - (\beta_1 + \beta_2),\ r_2 = 5 - (\beta_1 + 2\beta_2),\ r_3 = 7 - (\beta_1 + 3\beta_2),\ r_4 = 10 - (\beta_1 + 4\beta_2).$$
Therefore $S(\beta_1, \beta_2) = r_1^2 + r_2^2 + r_3^2 + r_4^2$, that is,
$$S(\beta_1, \beta_2) = [6 - (\beta_1 + \beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 + [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2.$$
After expansion, $S(\beta_1, \beta_2) = 4\beta_1^2 + 20\beta_1\beta_2 + 30\beta_2^2 - 56\beta_1 - 154\beta_2 + 210$.
3. Differentiating with respect to β_1 and β_2 gives $\frac{\partial S}{\partial \beta_1} = 8\beta_1 + 20\beta_2 - 56$, and $\frac{\partial S}{\partial \beta_2} = 20\beta_1 + 60\beta_2 - 154$. At a minimum these derivatives vanish:
$$\begin{cases} 8\beta_1 + 20\beta_2 = 56, \\ 20\beta_1 + 60\beta_2 = 154. \end{cases}$$
These are the normal equations.
4. Solving the system, $\beta_1 = 3.5$ and $\beta_2 = 1.4$. Hence the least-squares line is $y = 3.5 + 1.4x$. The fitted values are 4.9, 6.3, 7.7, 9.1, so the residuals are
$$\begin{aligned} r_1 &= 6 - 4.9 = 1.1, \\ r_2 &= 5 - 6.3 = -1.3, \\ r_3 &= 7 - 7.7 = -0.7, \\ r_4 &= 10 - 9.1 = 0.9. \end{aligned}$$
The minimum value of the sum of squared residuals is
$$\begin{aligned} S(3.5, 1.4) &= 1.1^2 + (-1.3)^2 + (-0.7)^2 + 0.9^2 \\ &= 1.21 + 1.69 + 0.49 + 0.81 \\ &= 4.2. \end{aligned}$$

5. We write the problem in matrix form: $y = \begin{pmatrix} 6 \\ 5 \\ 7 \\ 10 \end{pmatrix}$, $X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$.
The normal equations are $X^T X, \beta = X^T y$. Now $X^T X = \begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix}$, $X^T y = \begin{pmatrix} 28 \\ 77 \end{pmatrix}$.
Thus $\begin{pmatrix} 4 & 10 \\ 10 & 30 \end{pmatrix} (\beta_1 \ \beta_2) = \begin{pmatrix} 28 \\ 77 \end{pmatrix}$, which is exactly the system obtained in Question 3.
Since $(X^T X)^{-1} = \frac{1}{20} \begin{pmatrix} 30 & -10 \\ -10 & 4 \end{pmatrix}$, we obtain $\beta = (X^T X)^{-1} X^T y = \frac{1}{20} \begin{pmatrix} 30 & -10 \\ -10 & 4 \end{pmatrix} \begin{pmatrix} 28 \\ 77 \end{pmatrix} = \begin{pmatrix} 3.5 \\ 1.4 \end{pmatrix}$. Therefore $\beta = \begin{pmatrix} 3.5 \\ 1.4 \end{pmatrix}$. Geometrically, the column space of X is the plane $\mathrm{Im}(X) = \mathrm{span}\left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \right\} \subset \mathbb{R}^4$. The vector $\widehat{y} = X\beta$ is the orthogonal projection of y onto $\mathrm{Im}(X)$.
The residual vector $r = y - \widehat{y}$ is orthogonal to every vector in $\mathrm{Im}(X)$, which is equivalent to $X^T r = 0$. This orthogonality condition is precisely the normal equation $X^T(X\beta - y) = 0$.

Exercise 3 [⋆ Least-squares fitting a parabola]

1. The model is $y \approx \beta_1 x^2 = X\beta$, we obtain
$$X\beta = \begin{pmatrix} 1 \\ 4 \\ 9 \\ 16 \end{pmatrix} \beta_1 = \begin{pmatrix} \beta_1 \\ 4\beta_1 \\ 9\beta_1 \\ 16\beta_1 \end{pmatrix}.$$
Hence the corresponding system is
$$\begin{cases} \beta_1 \approx 6, \\ 4\beta_1 \approx 5, \\ 9\beta_1 \approx 7, \\ 16\beta_1 \approx 10. \end{cases}$$
Since there are more equations than unknowns, the system is overdetermined and is solved in the least-squares sense.
2. The least-squares estimate β minimizes the residual norm: $\min_{\beta \in \mathbb{R}} \|y - X\beta\|^2$. In particular, $\{X\beta \ ; \ \beta \in \mathbb{R}\}$ is the vector space spanned by the column(s) of X, and the point of this space that minimizes the distance to y is the orthonormal projection of y onto $\mathrm{Span}(X) = \mathrm{Span}\left\{ \begin{pmatrix} 1 \\ 4 \\ 9 \\ 16 \end{pmatrix} \right\}$. The least-squares solution is **characterized** by the fact that $X\beta$ is the orthogonal projection of y onto $\mathrm{Col}(X)$.
3. The matrix for the orthogonal projection onto the span of the columns of X is simply $X^\top$, hence we get the normal equation:
$$X^\top X \, \beta = X^\top y.$$
Since $X^\top X$ is a 1×1 non-zero matrix, it is invertible and we get: $\beta = (X^\top X)^{-1} X^\top y$.
Let us compute the quantities:
$$X^\top X = 1^2 + 4^2 + 9^2 + 16^2 = 354, \qquad X^\top y = 1 \cdot 6 + 4 \cdot 5 + 9 \cdot 7 + 16 \cdot 10 = 249.$$
Therefore, $\beta = \frac{249}{354} = \frac{83}{118} \approx 0.703$. Thus the least-squares parabola is
$$y \approx \frac{83}{118} x^2 \approx 0.703 \cdot x^2.$$

Exercise 4 [⋆ Heteroscedasticity and weighted linear regression]

1. We first expand $f(\beta) = (y - X\beta)^\top W(y - X\beta)$. Using matrix distributivity,
$$f(\beta) = y^\top Wy - y^\top WX\beta - \beta^\top X^\top Wy + \beta^\top X^\top WX\beta.$$
Since $y^\top WX\beta$ is a scalar, $y^\top WX\beta = (y^\top WX\beta)^\top = \beta^\top X^\top Wy$. Therefore, $f(\beta) = y^\top Wy - 2\beta^\top X^\top Wy + \beta^\top X^\top WX\beta$. Hence f is a quadratic function of β.
2. Differentiating with respect to β, we obtain $\nabla f(\beta) = -2X^\top Wy + 2X^\top WX\,\beta$. A minimizer must satisfy $\nabla f(\beta) = 0$, which gives $-2X^\top Wy + 2X^\top WX\,\beta = 0$. After dividing by 2, $X^\top WX\,\beta = X^\top Wy$. These are called the *weighted normal equations*.
3. Let $z \in \mathbb{R}^p$ be nonzero. Then $z^\top(X^\top WX)z = (Xz)^\top W(Xz)$. Since W is diagonal with strictly positive entries, $(Xz)^\top W(Xz) = \sum_{i=1}^n w_i(Xz)_i^2$. Each term is nonnegative and all weights satisfy $w_i > 0$.
Because X has full column rank, $Xz \neq 0$ whenever $z \neq 0$. Hence $\sum_{i=1}^n w_i(Xz)_i^2 > 0$. Therefore $z^\top(X^\top WX)z > 0 \qquad (z \neq 0)$. Thus $X^\top WX$ is positive definite and consequently invertible.
4. Since $X^\top WX$ is invertible, the normal equations have the unique solution $\hat\beta = (X^\top WX)^{-1}X^\top Wy$.
5. Let $\hat y = X\hat\beta, \qquad r = y - \hat y$. From the normal equations, $X^\top WX\hat\beta = X^\top Wy$. Substituting $\hat y = X\hat\beta$, $X^\top W\hat y = X^\top Wy$. Hence $X^\top W(y - \hat y) = 0$, that is, $X^\top Wr = 0$.
Geometrically, this means that the residual vector r is orthogonal to every column of X with respect to the weighted inner product determined by W.
6. The weighted inner product is $\langle u, v\rangle_W = u^\top Wv$. Let $x \in \mathrm{Col}(X)$. Then there exists $a \in \mathbb{R}^p$ such that $x = Xa$.
Using $X^\top Wr = 0$, $\langle r, x\rangle_W = r^\top W(Xa) = (X^\top Wr)^\top a = 0$. Thus r is orthogonal to every vector of $\mathrm{Col}(X)$.
Since $y = \hat y + r, \qquad \hat y \in \mathrm{Col}(X)$, and r is W-orthogonal to $\mathrm{Col}(X)$, we conclude that $\hat y$ is the orthogonal projection of y onto $\mathrm{Col}(X)$ with respect to $\langle\cdot,\cdot\rangle_W$.
7. Using the expression of $\hat\beta$,
$$\begin{aligned}\hat y &= X\hat\beta \\ &= X(X^\top WX)^{-1}X^\top Wy.\end{aligned}$$
Hence $\hat y = P_W y, \qquad P_W = X(X^\top WX)^{-1}X^\top W$.
To show that P_W is idempotent,
$$\begin{aligned}P_W^2 &= X(X^\top WX)^{-1}X^\top WX(X^\top WX)^{-1}X^\top W \\ &= X(X^\top WX)^{-1}(X^\top WX)(X^\top WX)^{-1}X^\top W \\ &= X(X^\top WX)^{-1}X^\top W \\ &= P_W.\end{aligned}$$
Therefore $P_W^2 = P_W$.
A matrix satisfying $P^2 = P$ is called a projection matrix because applying it twice produces the same result as applying it once. In this case, P_W projects vectors onto $\mathrm{Col}(X)$ with respect to the weighted inner product induced by W.

Exercise 5 [⋆ Heteroscedasticity and generalized least squares]

1. We compute the expectation of the OLS estimator: $\hat\beta_{\mathrm{OLS}} = (X^\top X)^{-1}X^\top y$. Using the model $y = X\beta + \varepsilon$,
$$\begin{aligned}\mathbb{E}[\hat\beta_{\mathrm{OLS}}] &= (X^\top X)^{-1}X^\top \mathbb{E}[y] \\ &= (X^\top X)^{-1}X^\top \mathbb{E}[X\beta + \varepsilon] \\ &= (X^\top X)^{-1}X^\top (X\beta + \mathbb{E}[\varepsilon]) \\ &= (X^\top X)^{-1}X^\top X\beta \\ &= \beta.\end{aligned}$$

Thus OLS is unbiased even when the errors are heteroscedastic.

2. Write $\hat{\beta}_{\text{OLS}} = Ay$ with $A = (X^\top X)^{-1}X^\top$. Then $\text{Cov}(\hat{\beta}_{\text{OLS}}) = A\,\text{Cov}(y)\,A^\top$. Since $y = X\beta + \varepsilon$ and β is deterministic, $\text{Cov}(y) = \text{Cov}(\varepsilon) = \Sigma$. Hence

$$\text{Cov}(\hat{\beta}_{\text{OLS}}) = (X^\top X)^{-1}X^\top \Sigma X (X^\top X)^{-1}.$$

3. If $\Sigma = \sigma^2 I_n$, then

$$\begin{aligned}\text{Cov}(\hat{\beta}_{\text{OLS}}) &= (X^\top X)^{-1}X^\top(\sigma^2 I_n)X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}.\end{aligned}$$

4. Consider the objective function $f(\beta) = (y - X\beta)^\top W(y - X\beta)$. Expanding,

$$f(\beta) = y^\top W y - 2\beta^\top X^\top W y + \beta^\top X^\top W X \beta.$$

Differentiating, $\nabla f(\beta) = -2X^\top W y + 2X^\top W X\beta$. Setting the gradient to zero gives the normal equations $X^\top W X\beta = X^\top W y$. Since $X^\top W X$ is invertible (under standard rank assumptions), $\hat{\beta}_{\text{WLS}} = (X^\top W X)^{-1}X^\top W y$. Thus WLS is the minimizer.

5. Since $W = \Sigma^{-1}$, define $\tilde{y} = \Sigma^{-1/2}y, \qquad \tilde{X} = \Sigma^{-1/2}X$. Then

$$\begin{aligned}\|\tilde{y} - \tilde{X}\beta\|^2 &= (\tilde{y} - \tilde{X}\beta)^\top(\tilde{y} - \tilde{X}\beta) \\ &= (y - X\beta)^\top \Sigma^{-1}(y - X\beta) \\ &= (y - X\beta)^\top W(y - X\beta).\end{aligned}$$

Thus the weighted least-squares problem is equivalent to an ordinary least-squares problem on transformed variables.

6. Define $\tilde{\varepsilon} = \Sigma^{-1/2}\varepsilon$. Then

$$\begin{aligned}\text{Cov}(\tilde{\varepsilon}) &= \Sigma^{-1/2}\,\text{Cov}(\varepsilon)\Sigma^{-1/2} \\ &= \Sigma^{-1/2}\Sigma\Sigma^{-1/2} \\ &= I_n.\end{aligned}$$

Hence the transformation $\Sigma^{-1/2}$ *whitens* the noise: it restores homoscedasticity (unit variance and no correlation between components).

Exercise 6 [Anscombe's quadret]

See Figure 3. Even though the linear fit obtained by linear regression is the same and is "as precise" (i.e. R^2 is the same) for all four series of data points, each series is drastically different. In each case (except Dataset I), another regression (either using another model, or using another type of regression) would be more adapted. This emphasizes the fact that: 1) plotting is essential; 2) statistics is a far more subtle subject than just "apply the recipes for regression".

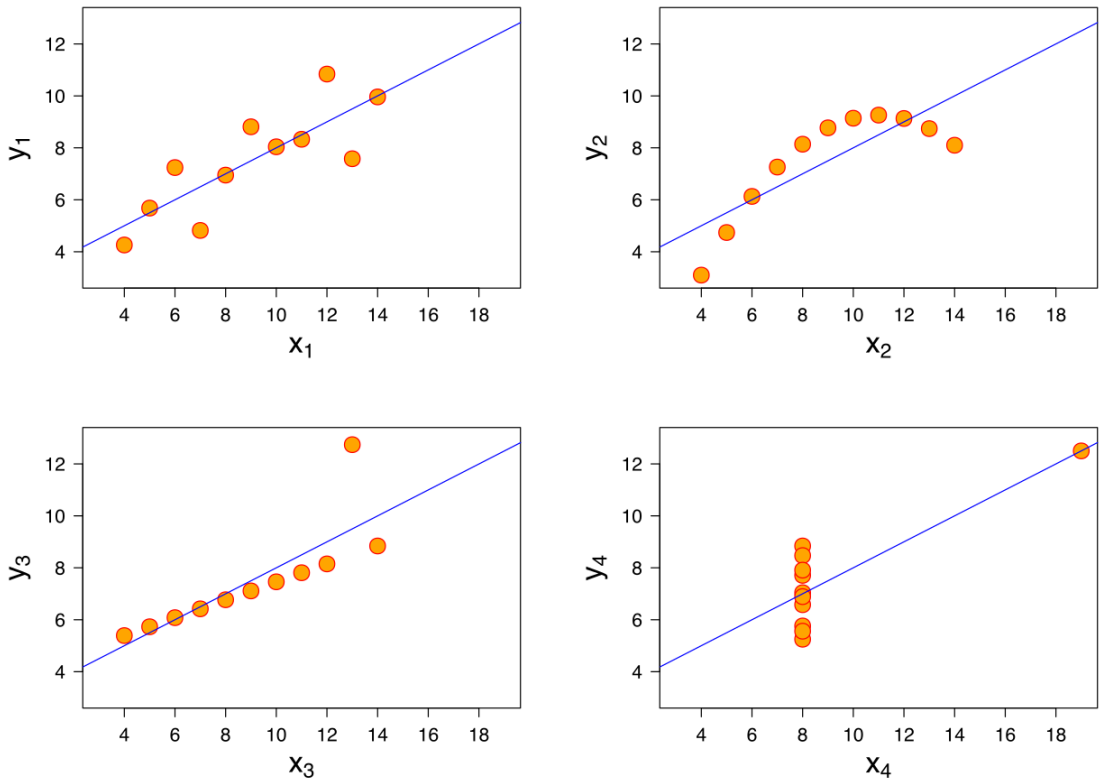


Figure 3: Plot of Anscombe's quadret

University of Osnabrück
Institute of Mathematics

Matthias Reitzner
Germain Poullot
Bernhard Hafer

Introduction to Statistics – Summer Semester 2026

Exercise Sheet 10

⋆ = be prepared to be asked questions on this exercise
• = a bit more difficult exercise

Recall: A computer with R is a good tool for processing data...

Exercise 1 [⋆ Understanding the coefficient of determination R^2 in linear regression]
You are given data on $n = 120$ students. For each student i, you observe:

$$y_i = \text{Final exam score}, \quad x_{i1} = \text{Study hours per week}, \quad x_{i2} = \text{Attendance rate}.$$

You consider the following linear models:

$$\mathcal{M}_1 : \quad y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i,$$
$$\mathcal{M}_2 : \quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

Let $\hat{y}_i$ denote fitted values from a model, and define:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad R^2 = 1 - \frac{SSE}{SST}.$$

1. Show that $R^2 = \frac{SST - SSE}{SST}$. Explain in words what the numerator represents, and interpret R^2 as a proportion.
2. Show that when moving from model $\mathcal{M}_1$ to $\mathcal{M}_2$, the residual sum of squares cannot increase: $SSE_{\mathcal{M}_2} \leq SSE_{\mathcal{M}_1}$. Deduce that $R^2_{\mathcal{M}_2} \geq R^2_{\mathcal{M}_1}$. Explain why this does *not* necessarily mean that $\mathcal{M}_2$ is a better predictive model.
3. Assume that for model $\mathcal{M}_2$ you obtain $R^2 = 0.72$.
   Explain precisely what the statement "$R^2 = 0.72$" means in terms of variability of y. Then discuss why it is incorrect to interpret this as:
   - "72% of individual exam scores are correctly explained," or
   - "study hours and attendance explain 72% of each student's grade."

Exercise 2 [⋆ t-test on almost constant data]
Consider the simple linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$, where the errors ε_i are assumed to be independent and identically distributed as $\mathcal{N}(0, \sigma^2)$.

The following observations are available:

i	1	2	3	4	5
x_i	1	2	3	4	5
y_i	5	4	6	4	6

1. Compute the least-squares estimators $\hat{\alpha}$ and $\hat{\beta}$, and write the fitted regression line.
2. Compute the residuals $e_i = y_i - \hat{y}_i$ and deduce the residual sum of squares $\text{RSS} = \sum_{i=1}^{5} e_i^2$.
3. Using the estimator $\hat{\sigma}^2 = \frac{\text{RSS}}{n-2}$, compute the estimated variance of the errors and the standard error of $\hat{\beta}$.

4. Consider the hypotheses $H_0 : \beta = 0$. Compute the corresponding t-statistic and specify its distribution under H_0. At the significance level $\alpha = 5\%$ and using a lookup table, perform the test and interpret the result in terms of the existence of a linear relationship between x and y.

Exercise 3 [⋆ Stochastic Trigonometric Regression Model]

A coastal engineering team is studying daily tide height variations at a harbor. The tide is approximately periodic but affected by random environmental noise (wind, pressure systems, measurement error). The observed data y_t for $t \in \{0, ..., 23\}$ are assumed to follow the stochastic model

$$y_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{24}\right) + \beta_2 \cos\left(\frac{2\pi t}{24}\right) + \varepsilon_t,$$

where $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ are independent random variables with unknown variance $\sigma^2 > 0$.

The observed values are given in the table below:

t	0	1	2	3	4	5	6	7	8	9	10	11
y_t	1.2	0.5	−0.3	−1.1	−1.4	−1.0	−0.2	0.6	1.5	2.0	1.7	1.0

t	12	13	14	15	16	17	18	19	20	21	22	23
y_t	0.2	−0.6	−1.3	−1.6	−1.2	−0.4	0.5	1.3	1.9	1.8	1.1	0.3

1. Show that the model can be written in matrix form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$. Explicitly construct the design matrix X.
2. Using the method of least squares, derive an estimator for $\boldsymbol{\beta}$ and compute the explicit form $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}$.
3. Define the amplitude-phase representation $\beta_1 \sin(\omega t) + \beta_2 \cos(\omega t) = A \sin(\omega t + \varphi)$, with $\omega = \frac{2\pi}{24}$. Express A and φ in terms of β_1 and β_2 and interpret their physical meaning in the context of periodic phenomena.
4. The error variance is not constant but satisfies $\mathrm{Var}(\varepsilon_t) = 0.5 + 0.03t$. Discuss whether ordinary least squares remains efficient in this setting and propose an alternative estimation method. Describe the corresponding weighting scheme.

Exercise 4 [⋆ When sampled regressors and errors terms are correlated]

Let $(X_i, Y_i)_{1 \le i \le n}$ be independent copies of a random vector (X, Y) satisfying the linear model

$$Y = \beta X + \varepsilon,$$

where $\beta \in \mathbb{R}$ is an unknown parameter and ε is a random error. Assume that $\mathbb{E}[X] = 0$, $\mathbb{E}[\varepsilon] = 0$, $\mathbb{E}[X^2] < \infty$, $\mathbb{E}[\varepsilon^2] < \infty$. In contrast with the classical linear regression model, do *not* assume that X and ε are independent. Denote $\sigma_X^2 = \mathrm{Var}(X) > 0$, $\tau = \mathrm{Cov}(X, \varepsilon)$. The ordinary least-squares estimator of β is $\widehat{\beta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$.

1. Show that $\widehat{\beta}_n = \beta + \frac{\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i}{\frac{1}{n}\sum_{i=1}^n X_i^2}$.
2. Using the law of large numbers, prove that $\widehat{\beta}_n \xrightarrow{\text{a.s.}} \beta + \frac{\tau}{\sigma_X^2}$.
3. Deduce that the ordinary least square estimator is consistent if and only if $\tau = 0$.
4. The quantity $\frac{\tau}{\sigma_X^2}$ is called the *asymptotic bias* of the estimator. Give an interpretation of its sign and magnitude.
5. Assume now that (X, ε) is jointly Gaussian and that $\varepsilon = \rho X + \eta$, where η is centered and independent of X. Compute the probability limit of $\widehat{\beta}_n$ and discuss its dependence on ρ.

Exercise 5 [⋆ Variance decomposition in quadratic stochastic regression]

Let $(X_i, U_i)_{i=1}^n$ be i.i.d. random variables such that $\mathbb{E}[U_i \mid X_i] = 0$, $\mathbb{V}(U_i \mid X_i) = \sigma^2 < \infty$, and consider the quadratic stochastic regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + U_i.$$

Assume $\mathbb{E}[X^4] < \infty$ and define the regressors $Z_{1,i} = X_i$, $Z_{2,i} = X_i^2$. Let $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ be the ordinary least square estimators from regressing Y_i on $(1, X_i, X_i^2)$, and define $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$, $\hat{U}_i = Y_i - \hat{Y}_i$.

1. Show that the population variance of Y satisfies $\mathbb{V}(Y) = \mathbb{V}(\beta_0 + \beta_1 X + \beta_2 X^2) + \sigma^2$, and interpret the two components as explained and unexplained variance.
2. Show that $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2$, and deduce that $\mathbb{V}(\mathbb{E}[Y \mid X]) = \mathbb{V}(\beta_0 + \beta_1 X + \beta_2 X^2)$.
3. Let $\mathcal{Z} = \text{span}(1, X, X^2)$ and interpret $(\hat{Y}_1, \dots, \hat{Y}_n)$ as the orthogonal projection of $(Y_1, \dots, Y_n)$ onto $\mathcal{Z}$. Show that the variance decomposition $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ holds.
4. Show that the ordinary least square estimator can be written as $\hat{\beta} = \beta + (X^\top X)^{-1} X^\top U$, where X is the design matrix with columns $(1, X, X^2)$, and deduce that $\mathbb{E}\hat{\beta} = \beta$.
5. Define $s^2 = \frac{1}{n-3} \sum_{i=1}^n \hat{U}_i^2$. Show that s^2 is an unbiased estimator of σ^2 and explain why the denominator is $n - 3$ in terms of the dimension of the projection space.

Introduction to Statistics – Solutions Sheet 10

Exercise 1 [⋆ Understanding the coefficient of determination R^2 in linear regression]

We work with the definitions $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and $R^2 = 1 - \frac{SSE}{SST}$.

1. Starting from the definition, $R^2 = 1 - \frac{SSE}{SST}$.
   Putting the terms over a common denominator gives $R^2 = \frac{SST}{SST} - \frac{SSE}{SST} = \frac{SST - SSE}{SST}$.
   The numerator $SST - SSE$ represents the reduction in squared variation achieved by the regression model compared to the baseline model that predicts $\bar{y}$ for all observations. It is therefore the amount of variation in y explained by the regressors.
   Hence R^2 is the proportion of total variation in y explained by the model.
2. Since $\mathcal{M}_2$ includes all regressors of $\mathcal{M}_1$ plus an additional regressor, the parameter space of $\mathcal{M}_1$ is contained in that of $\mathcal{M}_2$.
   Ordinary least squares minimizes the residual sum of squares over the available parameter space. Expanding this space cannot increase the minimum achievable value, hence $SSE_{\mathcal{M}_2} \leq SSE_{\mathcal{M}_1}$.
   Since SST does not depend on the model, we obtain $R^2_{\mathcal{M}_2} = 1 - \frac{SSE_{\mathcal{M}_2}}{SST} \geq 1 - \frac{SSE_{\mathcal{M}_1}}{SST} = R^2_{\mathcal{M}_1}$.
   This does not imply that $\mathcal{M}_2$ is a better predictive model, since adding regressors always weakly improves in-sample fit, even when the additional variable is irrelevant, potentially leading to overfitting and worse out-of-sample performance.
3. The statement $R^2 = 0.72$ means that $\frac{SST - SSE}{SST} = 0.72$, so 72% of the total variation in exam scores around their mean $\bar{y}$ is explained by the linear regression model using study hours and attendance, while 28% remains unexplained in the residuals.
   It is incorrect to interpret this as:
   - "72% of individual exam scores are correctly explained," because R^2 measures explained variance, not correctness at the level of individual predictions.
   - "study hours and attendance explain 72% of each student's grade," because R^2 is a global variance decomposition measure and does not apply to individual observations.

Exercise 2 [⋆ t-test on almost constant data]

1. Compute sample means: $\bar{x} = \frac{1+2+3+4+5}{5} = 3$, $\bar{y} = \frac{5+4+6+4+6}{5} = 5$. Compute: $S_{xx} = \sum_{i=1}^{5}(x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$, $S_{xy} = \sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = (-2)(0) + (-1)(-1) + 0 \cdot 1 + 1 \cdot (-1) + 2 \cdot 1 = 2$.
   Hence: $\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{2}{10} = 0.2$, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 5 - 0.2 \cdot 3 = 4.4$.
   Fitted regression line: $\hat{y} = 4.4 + 0.2x$.
2. Fitted values:
   $$\hat{y}_1 = 4.6,\ \hat{y}_2 = 4.8,\ \hat{y}_3 = 5.0,\ \hat{y}_4 = 5.2,\ \hat{y}_5 = 5.4.$$
   Residuals $e_i = y_i - \hat{y}_i$:
   $$e_1 = 0.4, \quad e_2 = -0.8, \quad e_3 = 1.0, \quad e_4 = -1.2, \quad e_5 = 0.6.$$
   Residual sum of squares: $\text{RSS} = 0.4^2 + (-0.8)^2 + 1^2 + (-1.2)^2 + 0.6^2 = 3.6$.
3. Error variance and standard error of $\hat{\beta}$ $\hat{\sigma}^2 = \frac{\text{RSS}}{n-2} = \frac{3.6}{3} = 1.2$.
   $\text{SE}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{1.2}{10}} = \sqrt{0.12} \approx 0.3464$.
4. t-statistic $t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} = \frac{0.2}{0.3464} \approx 0.577$.
   Under $H_0 : \beta = 0$, we have: $t \sim t_{n-2} = t_3$.
5. Critical value: $t_{0.975,3} \approx 3.182$.
   Since $|0.577| < 3.182$, we do not reject H_0.
   Conclusion: there is no statistically significant evidence at the 5% level of a linear relationship between x and y.

Exercise 3 [⋆ Stochastic Trigonometric Regression Model]

1. Define
$$\mathbf{y} = (y_0, \dots, y_{23})^\top, \qquad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top, \qquad \boldsymbol{\varepsilon} = (\varepsilon_0, \dots, \varepsilon_{23})^\top.$$
The model
$$y_t = \beta_0 + \beta_1 \sin\left(\frac{2\pi t}{24}\right) + \beta_2 \cos\left(\frac{2\pi t}{24}\right) + \varepsilon_t$$
can be written as
$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$
where the design matrix $X \in \mathbb{R}^{24\times 3}$ has rows
$$X_{t+1,\cdot} = \left(1,\ \sin\left(\frac{2\pi t}{24}\right),\ \cos\left(\frac{2\pi t}{24}\right)\right), \qquad t = 0, \dots, 23.$$

2. The least squares estimator minimizes $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$ and satisfies the normal equations
$$X^\top X\hat{\boldsymbol{\beta}} = X^\top \mathbf{y}.$$
Thus,
$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}.$$
Using the full 24-point equally spaced sampling over one period, we use orthogonality relations:
$$\sum_{t=0}^{23} \sin\left(\frac{2\pi t}{24}\right) = 0, \quad \sum_{t=0}^{23} \cos\left(\frac{2\pi t}{24}\right) = 0,$$
$$\sum_{t=0}^{23} \sin\left(\frac{2\pi t}{24}\right)\cos\left(\frac{2\pi t}{24}\right) = 0,$$
$$\sum_{t=0}^{23} \sin^2\left(\frac{2\pi t}{24}\right) = \sum_{t=0}^{23} \cos^2\left(\frac{2\pi t}{24}\right) = 12.$$
Hence,
$$X^\top X = \begin{pmatrix} 24 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 12 \end{pmatrix}, \qquad (X^\top X)^{-1} = \begin{pmatrix} \frac{1}{24} & 0 & 0 \\ 0 & \frac{1}{12} & 0 \\ 0 & 0 & \frac{1}{12} \end{pmatrix}.$$
Therefore,
$$\hat{\beta}_0 = \frac{1}{24}\sum_{t=0}^{23} y_t, \quad \hat{\beta}_1 = \frac{1}{12}\sum_{t=0}^{23} y_t \sin\left(\frac{2\pi t}{24}\right), \quad \hat{\beta}_2 = \frac{1}{12}\sum_{t=0}^{23} y_t \cos\left(\frac{2\pi t}{24}\right).$$

3. We write
$$\beta_1 \sin(\omega t) + \beta_2 \cos(\omega t) = A\sin(\omega t + \varphi), \qquad \omega = \frac{2\pi}{24}.$$
Using the identity
$$A\sin(\omega t + \varphi) = A\sin(\omega t)\cos\varphi + A\cos(\omega t)\sin\varphi,$$
we match coefficients:
$$\beta_1 = A\cos\varphi, \qquad \beta_2 = A\sin\varphi.$$
Thus,
$$A = \sqrt{\beta_1^2 + \beta_2^2}, \qquad \varphi = \operatorname{atan2}(\beta_2, \beta_1).$$
Interpretation: A is the amplitude of the periodic tide oscillation (strength of the daily cycle), while φ is the phase shift, determining the timing of high and low tides within the 24-hour cycle.

4. Since
$$\mathrm{Var}(\varepsilon_t) = 0.5 + 0.03t,$$
the errors are heteroskedastic. In this case, ordinary least squares is still unbiased but no longer efficient (it does not achieve minimum variance among linear unbiased estimators).
An appropriate alternative is *weighted least squares* (WLS), where observations are weighted by the inverse of their variances:
$$w_t = \frac{1}{0.5 + 0.03t}.$$
Let
$$W = \mathrm{diag}(w_0, \ldots, w_{23}).$$
Then the WLS estimator is
$$\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (X^\top W X)^{-1} X^\top W \mathbf{y}.$$
This weighting scheme gives less influence to observations with larger variance (larger t), improving efficiency.

Exercise 4 [⋆ When sampled regressors and errors terms are correlated]

1. We start from
$$\widehat{\beta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$
Using the model $Y_i = \beta X_i + \varepsilon_i$, we obtain
$$\sum_{i=1}^n X_i Y_i = \sum_{i=1}^n X_i(\beta X_i + \varepsilon_i) = \beta \sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i \varepsilon_i.$$
Hence
$$\widehat{\beta}_n = \frac{\beta \sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2} = \beta + \frac{\sum_{i=1}^n X_i \varepsilon_i}{\sum_{i=1}^n X_i^2}.$$
Dividing numerator and denominator by n gives
$$\widehat{\beta}_n = \beta + \frac{\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i}{\frac{1}{n}\sum_{i=1}^n X_i^2}.$$

2. Since (X_i, ε_i) are i.i.d. with finite second moments, the strong law of large numbers yields
$$\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i \xrightarrow{\text{a.s.}} \mathbb{E}[X\varepsilon] = \mathrm{Cov}(X, \varepsilon) = \tau,$$
because $\mathbb{E}[X] = \mathbb{E}[\varepsilon] = 0$.
Similarly,
$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{\text{a.s.}} \mathbb{E}[X^2] = \sigma_X^2.$$
Since $\sigma_X^2 > 0$, by the continuous mapping theorem,
$$\widehat{\beta}_n \xrightarrow{\text{a.s.}} \beta + \frac{\tau}{\sigma_X^2}.$$

3. The estimator is consistent if and only if its probability limit equals β, i.e.
$$\beta + \frac{\tau}{\sigma_X^2} = \beta,$$
which holds if and only if $\tau = 0$.
Thus, OLS is consistent if and only if X is uncorrelated with the error term ε.

4. The asymptotic bias is
$$\frac{\tau}{\sigma_X^2}.$$
Its sign is determined by $\tau = \mathrm{Cov}(X, \varepsilon)$:
   - If $\tau > 0$, large values of X tend to be associated with positive errors, so OLS overestimates β asymptotically.
   - If $\tau < 0$, large values of X tend to be associated with negative errors, so OLS underestimates β asymptotically.

   Its magnitude increases with the strength of the dependence between X and ε, and decreases with the dispersion of X through σ_X^2.
5. Assume
$$\varepsilon = \rho X + \eta,$$
where η is centered and independent of X. Then
$$\mathbb{E}[X\varepsilon] = \mathbb{E}[X(\rho X + \eta)] = \rho\mathbb{E}[X^2] + \mathbb{E}[X\eta] = \rho\sigma_X^2,$$
since independence and centering imply $\mathbb{E}[X\eta] = 0$.
Thus,
$$\tau = \rho\sigma_X^2,$$
and the probability limit becomes
$$\widehat{\beta}_n \xrightarrow{p} \beta + \rho.$$
Discussion: the endogeneity parameter ρ shifts the limit of the estimator directly. A positive ρ inflates the estimated slope, while a negative ρ deflates it. The OLS estimator therefore cannot distinguish between the structural effect β and the spurious correlation induced by the dependence between X and ε.

Exercise 5 [⋆ Variance decomposition in quadratic stochastic regression]

1. Write
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + U.$$
Since $\beta_0 + \beta_1 X + \beta_2 X^2$ is a function of X,
$$\mathrm{Cov}\left(\beta_0 + \beta_1 X + \beta_2 X^2,\, U\right) = \mathbb{E}\big[(\beta_0 + \beta_1 X + \beta_2 X^2)U\big] = \mathbb{E}\big[\mathbb{E}[(\beta_0 + \beta_1 X + \beta_2 X^2)U \mid X]\big].$$
Using $\mathbb{E}[U \mid X] = 0$, this equals 0. Hence
$$\mathrm{Var}(Y) = \mathrm{Var}(\beta_0 + \beta_1 X + \beta_2 X^2 + U) = \mathrm{Var}(\beta_0 + \beta_1 X + \beta_2 X^2) + \mathrm{Var}(U),$$
and since $\mathbb{E}[\mathrm{Var}(U \mid X)] = \sigma^2$, we get
$$\mathrm{Var}(Y) = \mathrm{Var}(\beta_0 + \beta_1 X + \beta_2 X^2) + \sigma^2.$$
Thus the first term is the explained variance (systematic variation due to X), and σ^2 is the unexplained noise.
2. Taking conditional expectation,
$$\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \mathbb{E}[U \mid X] = \beta_0 + \beta_1 X + \beta_2 X^2.$$
Therefore,
$$\mathrm{Var}(\mathbb{E}[Y \mid X]) = \mathrm{Var}(\beta_0 + \beta_1 X + \beta_2 X^2).$$

3. Let $\mathcal{Z} = \text{span}(1, X, X^2)$. The fitted values $(\hat{Y}_1, \ldots, \hat{Y}_n)$ are the orthogonal projection of $(Y_1, \ldots, Y_n)$ onto $\mathcal{Z}$ in $\mathbb{R}^n$.
   By the Pythagorean theorem for orthogonal projections,

$$\|Y - \bar{Y}\mathbf{1}\|^2 = \|\hat{Y} - \bar{Y}\mathbf{1}\|^2 + \|Y - \hat{Y}\|^2.$$

   In scalar form,

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

4. Let X be the design matrix with columns $(1, X, X^2)$. Then

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

   Using $Y = X\beta + U$, we obtain

$$\hat{\beta} = (X^\top X)^{-1} X^\top (X\beta + U) = \beta + (X^\top X)^{-1} X^\top U.$$

   Taking expectations and using $\mathbb{E}[U \mid X] = 0$,

$$\mathbb{E}[\hat{\beta} \mid X] = \beta + (X^\top X)^{-1} X^\top \mathbb{E}[U \mid X] = \beta.$$

   Hence $\mathbb{E}[\hat{\beta}] = \beta$.
5. Define

$$s^2 = \frac{1}{n-3}\sum_{i=1}^{n}\hat{U}_i^2.$$

   In OLS, the residual sum of squares satisfies

$$\sum_{i=1}^{n}\hat{U}_i^2 = U^\top (I - H)U,$$

   where H is the projection (hat) matrix onto a 3-dimensional space. Hence $\text{rank}(H) = 3$ and $\text{tr}(I - H) = n - 3$.
   Using $\mathbb{E}[UU^\top \mid X] = \sigma^2 I$,

$$\mathbb{E}\left[\sum_{i=1}^{n}\hat{U}_i^2 \mid X\right] = \sigma^2 \,\text{tr}(I - H) = \sigma^2(n-3).$$

   Thus

$$\mathbb{E}[s^2 \mid X] = \sigma^2, \quad \text{so } \mathbb{E}[s^2] = \sigma^2.$$

   The denominator $n-3$ appears because three degrees of freedom are used to estimate $(\beta_0, \beta_1, \beta_2)$, leaving $n - 3$ independent residual degrees of freedom.