Statistics for Data Science

Wintersemester 2023/24

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

First lecture, October 16, 2023

Practical matters

- Lectures on Mondays at 10:15-12:00 in A6/032 by Vesa Kaarnioja.
- Exercises on Tuesdays at 8:30-10:00 in A6/008 by Vesa Kaarnioja starting next week.
- Weekly exercises published after each lecture. Please complete the tasks before the exercise session. In the exercise sessions, please be prepared to present your solutions to any problems that you have completed successfully.
- The conditions for completing this course are
  (1) *successfully completing at least 60% of the course's exercises*,
  (2) *presenting at least 2 solutions to exercise problems in the exercise sessions*, and
  (3) *successfully passing the course exam*.
- The course exam will be held **February 12, 2024,** starting at 10:00 in room A6/032.

Course contents

This course will consist of three main parts:

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

Introduction

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

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

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

Using probabilistic notations,

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

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

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

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

Probability foundations

Sample space

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

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

Example

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

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

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

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

Example

We throw a coin three times. To model this experiment, we consider $\Omega = \{H, T\}^3$, i.e., the set of all vectors with 3 entries, with each entry taking value H or T. Here, H stands for "Heads" and T for "Tails" (choosing, e.g., $\Omega = \{0, 1\}^2$, with 0 and 1 representing Heads and Tails, respectively, would be equally valid). An event is any subset of $\{H, T\}^3$. For instance, we may consider the events

$$A = \{(H, H, T), (H, T, H), (T, H, H)\} \qquad \text{(get Heads exactly twice)}$$
$$B = \{(H, H, H), (T, T, T)\}. \qquad \text{(get 3 times the same result)}$$

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

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

Then the joint occurrence of B and C is

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

Probability measure

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

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

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

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

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

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

Definition

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

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

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

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

Example

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

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

Example

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

The event "both dice > 2", is

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

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

Example

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

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

Corollaries

Proposition

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

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

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

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

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

Proof.

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

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

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

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

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

4. Apply part 2 with $B = \Omega$ to get $\mathbb{P}(\Omega \setminus A) = \mathbb{P}(\Omega) - \mathbb{P}(A) = 1 - \mathbb{P}(A)$.

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

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

Recalling that $\mathbb{P}(E_2) = \mathbb{P}(A \cap B)$ yields the assertion.
6. We define $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1})$ for $n > 1$. Then the B_n are pairwise disjoint and $\bigcup_n B_n = \bigcup_n A_n$. Therefore

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

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

Definition (Conditional probability)

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

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

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

Remarks.

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

Example

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

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

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

Independence of events

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

Definition

Two events A and B are said to be independent if

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

This notion can be expressed in terms of conditional probability.

Lemma

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

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

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

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

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

Example

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

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

Then

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

while

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

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

Theorem (Law of total probability)

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

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

Proof. We have

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

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

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

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

Theorem (Bayes' theorem)

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

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

Proof. By definition,

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

On the other hand,

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

which yields the assertion. □