Machine Learning in Bioinformatics

Probability Basics

Philipp Benner
philipp.benner@bam.de

VP.1 - eScience
Federal Institute of Materials Research and Testing (BAM)

February 8, 2026

Introduction to Probability Theory

Roulette Wheel

Roulette Wheel

Assume we have a fair roulette wheel with 37 segments, of which

- 18 are black
- 18 are red
- 1 is green and labeled with a zero

The red and black segments are labeled with numbers ranging from 1 to 36, where

	even	odd
red	8	10
black	10	8

Roulette Wheel - Simple probabilities

- What is the probability of black?

$$\mathrm{pr}(\text{black}) = \frac{\#\text{black segments}}{\#\text{segments}} = \frac{18}{37}$$

Roulette Wheel - Simple probabilities

- What is the probability of black?

$$\mathrm{pr}(\text{black}) = \frac{\#\text{black segments}}{\#\text{segments}} = \frac{18}{37}$$

- What is the probability of black or green?

$$\begin{aligned}\mathrm{pr}(\text{black } \textcolor{red}{\text{or}} \text{ green}) &= \frac{\#\text{black and green segments}}{\#\text{segments}} \\ &= \frac{\#\text{black segments}}{\#\text{segments}} + \frac{\#\text{green segments}}{\#\text{segments}} \\ &= \mathrm{pr}(\text{black}) + \mathrm{pr}(\text{green})\end{aligned}$$

This property is called additivity

Roulette Wheel - Simple probabilities

- What is the probability of observing first black and afterwards red?

$$\mathrm{pr}(\text{first black and then red}) = \mathrm{pr}(\text{black})\mathrm{pr}(\text{red})$$

Roulette Wheel - Simple probabilities

- What is the probability of observing first black and afterwards red?

$$\mathrm{pr}(\text{first black and then red}) = \mathrm{pr}(\text{black})\mathrm{pr}(\text{red})$$

- and ⇒ "multiplication"
- or ⇒ "addition"

Roulette Wheel - Simple probabilities

- What is the probability of a black segment with an even number?

$$\begin{aligned}\mathrm{pr}(\text{black and even number})\\ &= \mathrm{pr}(\text{black})\mathrm{pr}(\text{even number})\\ &= \frac{\#\text{black segments}}{\#\text{segments}}\frac{\#\text{even segments}}{\#\text{segments}}\end{aligned}$$

Roulette Wheel - Simple probabilities

- What is the probability of a black segment with an even number?

$$
\begin{aligned}
&\mathrm{pr}(\text{black and even number}) \\
&\qquad = \mathrm{pr}(\text{black})\mathrm{pr}(\text{even number}) \\
&\qquad = \frac{\#\text{black segments}}{\#\text{segments}} \frac{\#\text{even segments}}{\#\text{segments}}
\end{aligned}
$$

Wrong!

- Both events are not independent! Some black segments are even.

Formal definitions

Sample space

The set of all possible outcomes is called sample space and typically denoted Ω. The elements of the sample space are called outcomes or *samples*

- For our roulette wheel, if we only care about the color of segments, then the sample space is

$$\Omega = \{\text{red}, \text{black}, \text{green}\}$$

- If we consider both colors and numbers, then

$$\Omega = \{0 : \text{green}, 1 : \text{red}, 2 : \text{black}, \dots\}$$

- Colors and numbers are not independent:

$$\Omega \neq \{\text{red}, \text{black}, \text{green}\} \times \{0, 1, 2, \dots, 36\}$$

Formal definitions

Events

An event E is any subset of Ω, denoted $E \subseteq \Omega$

- We assign probabilities to events $E \subseteq \Omega$
- The probability of "black or green" is denoted

$$\mathrm{pr}(\{\text{black}, \text{green}\})$$

- More formally, we may write $\mathrm{pr}(E)$ for some $E \subseteq \Omega$

Probability Axioms – Axiom I

- What is the lowest possible probability?
- Assume that

$$\Omega = \{\text{yellow}, \text{red}, \text{black}, \text{green}\}$$

then $\mathrm{pr}(\{\text{yellow}\}) = 0$, since there is no yellow segment

- We could also write $\mathrm{pr}(\varnothing) = 0$
- First probability axiom:

$$\mathrm{pr}(E) \geq 0 \quad \text{for all } E \subseteq \Omega$$

Probability Axioms – Axiom II

- What is the largest possible probability?
- Assume that

$$\Omega = \{\text{red}, \text{black}, \text{green}\}$$

then $\mathrm{pr}(\{\text{red}, \text{black}, \text{green}\}) = 1$

- Second probability axiom:

$$\mathrm{pr}(\Omega) = 1$$

Probability Axioms – Axiom III

- The third axiom covers the additivity of independent events

$$\mathrm{pr}(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_i^n \mathrm{pr}(E_i)$$

  if all E_i are independent

- Independence is not given if for example

$$E_1 = \{\text{black}, \text{green}\}, E_2 = \{\text{red}, \text{green}\}.$$

- In this case we have

$$\mathrm{pr}(E_1 \cup E_2) \neq \mathrm{pr}(E_1) + \mathrm{pr}(E_2) > 1$$

Probability distribution (discrete case)

A probability distribution $\mathrm{pr} : \mathbb{P}(\Omega) \rightarrow [0, 1]$ is a function that assigns a probability to each element of the powerset of Ω. In addition, it fulfills the probability axioms I-III, i.e.

- $\mathrm{pr}(E) \geq 0$ for all $E \subseteq \Omega$
- $\mathrm{pr}(\Omega) = 1$
- $\mathrm{pr}(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_i^n \mathrm{pr}(E_i)$

where $E_1, E_2, \ldots, E_n$ are pairwise independent events.

Complement and sum rule

- There are several consequences of the probability axioms, one is the complement rule

$$\mathrm{pr}(E^c) = \mathrm{pr}(\Omega) - \mathrm{pr}(E) = 1 - \mathrm{pr}(E)$$

- For example, the probability of not observing *black* is given by

$$\mathrm{pr}(\{\text{black}\}^c) = 1 - \mathrm{pr}(\{\text{black}\})$$

- Another important consequence is the addition law or *sum rule*, given by

$$\mathrm{pr}(A \cup B) = \mathrm{pr}(A) + \mathrm{pr}(B) - \mathrm{pr}(A \cap B)$$

  where $\mathrm{pr}(A \cap B) = 0$ if A and B are independent

Random Variables

- Random variables (RVs) add another layer of formalism – Why do we need them?
- Assume we consider a more complex random experiment, i.e. we observe the roulette game for *n* rounds
- What is the probability of observing *black* in the *i*th round?
- To formalize this notion, we would associate a random variable X_i with the *i*th round and write

$$\mathrm{pr}(X_i = \{\text{black}\})$$

- Similarly, we write

$$\mathrm{pr}(X_i = \{\text{black}\}, X_j = \{\text{green}\})$$

for observing *black* in the *i*th round and *green* in the *j* round

Random Variables

- There exist two types of roulette wheels:
  - 1 green segment and 37 in total (what we considered)
  - 2 green segments and 38 in total
- Let X_1 correspond to the first type and X_2 the second
- We see that

$$\mathrm{pr}(X_1 = \{\text{black}\}) \neq \mathrm{pr}(X_2 = \{\text{black}\})$$

- Random variables correspond to different types of distributions (probability assigments)

Random Variables – Formal definition

Random Variable (RV)

A random variable $X : \Omega \to \mathbb{R}$ is a mapping from the sample space Ω to a measurable space, typically the real numbers $\mathbb{R}$. We denote by $\{X = x\}$ the *event* that X takes the value x and with $X \sim D$ that X has distribution D.

- Our previous notation, e.g. $X_1 = \{\text{black}\}$, is not correct
- We stick to this notation for simplicity
- If possible we avoid random variables, to simplify notation, i.e. we write

$$\mathrm{pr}(\{\text{black}\}) = \mathrm{pr}(X_1 = \{\text{black}\})$$

if unambiguous

Conditional probabilities

- A conditional probability is the probability of an event given that another event has happened or is known
- Assume we know that the ball has landed on a red segment. What is the probability that the segment has an even number?
- This is a conditional probability denoted as

$$\mathrm{pr}(\{\text{even}\}\,|\{\text{red}\}) = ?$$

Conditional probabilities - Derivation

- Consider the following ingredients:
  - $\mathrm{pr}(\{\text{even}\} \,|\{\text{red}\})$: The probability of an *even* segment, given that the ball has landed on a red segment
  - $\mathrm{pr}(\{\text{red}\})$: The probability that we observe *red*
- What is the probability of *red* and of an *even* segment, given that the ball has landed on a red segment?

$$\mathrm{pr}(\{\text{red}\})\mathrm{pr}(\{\text{even}\} \,|\{\text{red}\})$$

- Logically, this is equivalent to asking: What is the probability of observing a *red* segment with an *even* number

$$\mathrm{pr}(\{\text{red}\})\mathrm{pr}(\{\text{even}\} \,|\{\text{red}\}) = \mathrm{pr}(\{\text{even and red}\})$$

Conditional probabilities - Derivation

- Let *A* denote the set of all *red* segments
- Let *B* denote the set of all *even* segments
- The set of *red* segments with an *even* number is $A \cap B$
- Hence, we can rewrite

  $$\mathrm{pr}(\{\text{red}\})\mathrm{pr}(\{\text{even}\}\,|\{\text{red}\}) = \mathrm{pr}(\{\text{even and red}\})$$

  as follows:

  $$\begin{aligned} \mathrm{pr}(A)\mathrm{pr}(B\,|\,A) &= \mathrm{pr}(A \cap (A^c \cup B)) \\ &= \mathrm{pr}(A \cap B) \end{aligned}$$

- Note that this is equivalent to logic calculus:

  $$\begin{aligned} A \wedge (A \rightarrow B) &= A \wedge (\neg A \vee B) \\ &= A \wedge B \end{aligned}$$

Conditional probability and Bayes theorem

The *conditional probability* of A given B is defined through

$$\mathrm{pr}(A \mid B)\mathrm{pr}(B) = \mathrm{pr}(A \cap B) = \mathrm{pr}(B \mid A)\mathrm{pr}(A)$$

If $\mathrm{pr}(B) > 0$ it follows that

$$\mathrm{pr}(A \mid B) = \frac{\mathrm{pr}(A \cap B)}{\mathrm{pr}(B)} = \frac{\mathrm{pr}(B \mid A)\mathrm{pr}(A)}{\mathrm{pr}(B)}$$

This is called *Bayes theorem*, where we call

- $\mathrm{pr}(A \mid B)$: the posterior
- $\mathrm{pr}(B \mid A)$: the likelihood
- $\mathrm{pr}(A)$: the prior probability
- $\mathrm{pr}(B)$: the evindence or marginal likelihood

Independent events

Independence

Two events A and B are called *independent*, if

$$\mathrm{pr}(A \cap B) = \mathrm{pr}(A \,|\, B)\mathrm{pr}(B) = \mathrm{pr}(A)\mathrm{pr}(B)\,.$$

We also denote independence as $A \perp\!\!\!\perp B$.

- Example:
  - We observe two rounds of roulette, associated with random variables X_1, and X_2
  - The probability of observing first *black* and then *red* is

$$\begin{aligned}\mathrm{pr}(X_1 =&\{\text{black}\}, X_2 = \{\text{red}\}) \\ &= \mathrm{pr}(X_2 = \{\text{red}\} \,|\, X_1 = \{\text{black}\})\mathrm{pr}(X_1 = \{\text{black}\}) \\ &= \mathrm{pr}(X_2 = \{\text{red}\})\mathrm{pr}(X_1 = \{\text{black}\})\end{aligned}$$

The Gambler's Fallacy

The Gambler's Fallacy [Hacking, 2001]

Consider our roulette game. The Fallacious Gambler reasons as follows:

- The roulette wheel is fair
- I have just observed 12 black spins in a row
- Since the wheel is fair, black and red come up equally often
- Hence, red has to come up pretty soon, I'd better start betting red

The gambler thinks that a sequence of twelve blacks makes it more likely that the wheel will step at red next time. If so, a past sequence affects future outcomes and the wheel is not fair. So trials would not be independent and the gambler's premises are inconsistent.

Law of total probability

Law of total probability

Let $B_1, B_2, \ldots, B_n$ denote n mutually independent and exhaustive events, then

$$\mathrm{pr}(A) = \sum_{i=1}^{n} \mathrm{pr}(A \cap B_i) = \sum_{i=1}^{n} \mathrm{pr}(A \mid B_i)\mathrm{pr}(B_i)$$

- Example:

$$\mathrm{pr}(\{\text{black}\}) = \mathrm{pr}(\{\text{black and even}\}) + \mathrm{pr}(\{\text{black and odd}\})$$

- The set of even and odd segments does not overlap (independence) and covers the full sample space (exhaustive)

Examples

Inductive logic and learning from past events

Inductive logic [Hacking, 2001]

Inductive logic is about risky arguments. It analyses inductive arguments using probability. A risky argument can be a very good one, and yet its conclusion can be false.

- Probability assignments reflect beliefs about events
- They may come from simple distributions or complex models, such as neural networks
- Bayes theorem is used to derive probabilistic if-statements:
  - If B has happened, what is the probability of A?

Medical testing

Medical testing

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease?

Medical testing

- Let I denote a random variable indicating infection, i.e. $I = \text{true}$ if you are infected

- Let T denote a random variable associated with the test result, i.e. $T = \text{true}$ if the test is positive

- We know that:
  - $\mathrm{pr}(I = \text{true}\,) = 1/10,000$
  - $\mathrm{pr}(T = \text{true}\mid I = \text{true}\,) = 0.99$
  - $\mathrm{pr}(T = \text{false}| I = \text{false}) = 0.99$

Medical testing

$$
\begin{aligned}
\mathrm{pr}(I = \text{true} \mid T = \text{true}) \\
&= \frac{\mathrm{pr}(T = \text{true} \mid I = \text{true})\mathrm{pr}(I = \text{true})}{\sum_i \mathrm{pr}(T = \text{true} \mid I = i)\mathrm{pr}(I = i)} \\
&= \frac{0.99 \cdot 1/10,000}{0.99 \cdot 1/10,000 + (1 - 0.99) \cdot (1 - 1/10,000)} \\
&= \frac{0.000099}{0.000099 + 0.009999} \\
&\approx 0.0098
\end{aligned}
$$

Hence, the chance of actually having the disease is less than 1%.

The Monty Hall problem

The Monty Hall problem

On a game show, a contestant is told the rules as follows: There are three doors, labeled 1, 2, 3. A single prize has been hidden behind one of them with equal probability. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other two doors, and he will do so in such a way as not to reveal the prize. For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened, and you will receive whatever is behind your final choice of door.

The Monty Hall problem

The Monty Hall problem

Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant

- stick with door 1,
- switch to door 2,
- does it make a difference?

The Monty Hall problem

- Let $P \in \{1, 2, 3\}$ denote a random variable associated with the location of the price
- Let $D \in \{1, 2, 3\}$ denote the door that has been opened by the gameshow host
- A priori we have

$$\mathrm{pr}(P = i) = 1/3$$

- We are interested in the posterior probability

$$\mathrm{pr}(P = i \,|\, D = 3) = \frac{\mathrm{pr}(D = 3 \,|\, P = i)\mathrm{pr}(P = i)}{\mathrm{pr}(D = 3)}$$

The Monty Hall problem

- The case $i = 1$

$$\begin{aligned}
\mathrm{pr}(P = 1 \mid D = 3) &= \frac{\mathrm{pr}(D = 3 \mid P = 1)\mathrm{pr}(P = 1)}{\sum_i \mathrm{pr}(D = 3 \mid P = i)\mathrm{pr}(P = i)} \\
&= \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 1 \cdot 1/3 + 0 \cdot 1/3} \\
&= 1/3
\end{aligned}$$

- The case $i = 2$

$$\begin{aligned}
\mathrm{pr}(P = 2 \mid D = 3) &= \frac{\mathrm{pr}(D = 3 \mid P = 2)\mathrm{pr}(P = 2)}{\sum_i \mathrm{pr}(D = 3 \mid P = i)\mathrm{pr}(P = i)} \\
&= \frac{1 \cdot 1/3}{1/2 \cdot 1/3 + 1 \cdot 1/3 + 0 \cdot 1/3} \\
&= 2/3
\end{aligned}$$

The Monty Hall problem

- The case $i = 3$

$$\begin{aligned}\mathrm{pr}(P = 1 \mid D = 3) &= \frac{\mathrm{pr}(D = 3 \mid P = 1)\mathrm{pr}(P = 1)}{\sum_i \mathrm{pr}(D = 3 \mid P = i)\mathrm{pr}(P = i)} \\ &= \frac{0 \cdot 1/3}{1/2 \cdot 1/3 + 1 \cdot 1/3 + 0 \cdot 1/3} \\ &= 0\end{aligned}$$

- Hence, we have the posterior distribution

$$\mathrm{pr}(P = i \mid D = 3) = (1/3, 2/3, 0)$$

- Switching the door increases the probability of getting the price

Probability Distributions

Bernoulli distribution

Bernoulli distribution

Let X be a random variable taking values in $\{0, 1\}$. If X follows a Bernoulli distribution with parameter p, i.e.

$$X \sim \text{Bernoulli}(p)$$

then

$$\mathrm{pr}(X = 1) = p\,.$$

- Flipping a coin once can be modeled using a Bernoulli distribution
- The coin is fair if $p = 1/2$

Categorical distribution

Categorical or multinoulli distribution

Let X be a random variable taking values in $\{0, 1, \ldots, k\}$ for any integer $k > 0$. If X follows a categorical distribution with parameters $p = (p_1, \ldots, p_k)$ such that $\sum_i p_i = 1$, i.e.

$$X \sim \text{Categorical}(p)$$

then

$$\mathrm{pr}(X = i) = p_i \,.$$

- The categorical distribution is the extension of the Bernoulli distribution to k outcomes

Geometric distribution

Geometric distribution

Let X be a random variable taking values in $\{1, 2, 3, \ldots\}$. If X follows a geometric distribution with parameter $p \in [0, 1]$, i.e.

$$X \sim \text{Geometric}(p)$$

then

$$\mathrm{pr}(X = k) = (1 - p)^{k-1} p \,.$$

- The probability distribution of the number X of Bernoulli trials needed to get one success
- It gives the probability that the first occurrence of success requires k independent trials, each with success probability p.

Normal distribution

Normal or Gaussian distribution

Let X be a random variable taking values in $\mathbb{R}$. If X follows a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, i.e.

$$X \sim \text{Normal}(\mu, \sigma)$$

then

$$\mathrm{pr}(X \in A) = \int_A f_{\mu,\sigma}(x)\mathrm{d}x\,,$$

where $f_{\mu,\sigma}$ is the normal density function

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

Normal distribution

- The probability distribution of continuous random variables is usually defined through density functions
- For continuous distributions we have $\mathrm{pr}(X = x) = 0$ for all $x \in \mathbb{R}$, i.e. the probability that a single real value is observed is always zero

Parameter estimation

Parameter estimation

- Assume we observed n realisations $x = (x_1, \ldots, x_n)$ from a known distribution with unknown parameters θ
- How can we estimate the values of θ?
- Bayes theorem

$$\mathrm{pr}(\theta \,|\, x) = \frac{\mathrm{pr}(x \,|\, \theta)\mathrm{pr}(\theta)}{\mathrm{pr}(x)}$$

- We take the value with highest probability, i.e.

$$\hat{\theta} = \arg\max_{\theta} \mathrm{pr}(\theta \,|\, x)$$

this is called the *maximum a-posteriori (MAP) estimate*

Parameter estimation I

- Constants can be dropped when computing the MAP, i.e.

$$
\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} \mathrm{pr}(\theta \,|\, x) \\
&= \arg\max_{\theta} \frac{\mathrm{pr}(x \,|\, \theta)\mathrm{pr}(\theta)}{\mathrm{pr}(x)} \\
&= \arg\max_{\theta} \mathrm{pr}(x \,|\, \theta)\mathrm{pr}(\theta) \\
&= \arg\max_{\theta} \left[\log \mathrm{pr}(x \,|\, \theta) + \log \mathrm{pr}(\theta)\right]
\end{aligned}
$$

  since $\mathrm{pr}(x)$ does not depend on θ. We can apply the logarithm, because it is a monotonic (*order preserving*) function, which does not change the position of the maximum

Parameter estimation II

- If we assume that we have no prior information on θ, i.e. $\mathrm{pr}(\theta)$ is uniform (constant), then we obtain the *maximum likelihood (ML) estimate*

$$\hat{\theta} = \arg\max_{\theta} \mathrm{pr}(x \mid \theta)$$

Parameter estimation – Continuous variables

- For continuous variables we know that

$$\mathrm{pr}(x \mid \theta) = 0$$

and therefore also $\mathrm{pr}(x)$ is zero, which causes the posterior distribution to be undefined

- There exists a Bayes theorem for densities

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}$$

- Hence, the MAP for continuous variables is simply

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} f(x \mid \theta) f(\theta) \\ &= \arg\max_{\theta} \left[\log f(x \mid \theta) + \log f(\theta)\right]\end{aligned}$$

Normal distribution

- Assume we observed n realisations $x = (x_1, \ldots, x_n)$ from $X \sim \text{Normal}(\mu, \sigma)$ with unknown μ and σ
- Furthermore, let's assume we have no prior information about μ and σ, i.e. the prior probability $\mathrm{pr}(\mu, \sigma)$ is uniform
- We derive the ML estimate

$$\begin{aligned}
\hat{\mu}, \hat{\sigma} &= \arg\max_{\mu,\sigma} f_{\mu,\sigma}(x) \\
&= \arg\max_{\mu,\sigma} \prod_{i=1}^{n} f_{\mu,\sigma}(x_i) \\
&= \arg\max_{\mu,\sigma} \sum_{i=1}^{n} \log f_{\mu,\sigma}(x_i)
\end{aligned}$$

Normal distribution

- We derive the ML estimate

$$
\begin{aligned}
\hat{\mu}, \hat{\sigma} &= \operatorname*{arg\,max}_{\mu,\sigma} \sum_{i=1}^{n} \log f_{\mu,\sigma}(x_i) \\
&= \operatorname*{arg\,max}_{\mu,\sigma} + \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\
&= \operatorname*{arg\,max}_{\mu,\sigma} -\frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2
\end{aligned}
$$

Normal distribution

- We derive the ML estimate of μ

$$
\begin{aligned}
& & \frac{\partial}{\partial\mu}\left[-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] &= 0\\
&\Rightarrow & -\frac{\partial}{\partial\mu}\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 &= 0\\
&\Rightarrow & \frac{\partial}{\partial\mu}\sum_{i=1}^{n}(x_i-\mu)^2 &= 0\\
&\Rightarrow & \sum_{i=1}^{n}-2(x_i-\mu) &= 0\\
&\Rightarrow & \mu &= \frac{1}{n}\sum_{i=1}^{n}x_i
\end{aligned}
$$

Normal distribution

- We derive the ML estimate of σ^2

$$\frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]=0$$

$$\Rightarrow \qquad -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0$$

$$\Rightarrow \qquad \frac{1}{2\sigma^2}\left[-n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]=0$$

$$\Rightarrow \qquad \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2=n$$

$$\Rightarrow \qquad \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Random Variables
Formal Definition

Random Variables – Formal Definition I

Probability Space

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

- Ω is the sample space,
- $\mathcal{F}$ is a σ-algebra of subsets of Ω,
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure.

Random Variables – Formal Definition II

Random Variable (formal)

A random variable is a measurable function

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})),$$

where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$.

Measurability means:

$$X^{-1}(B) \in \mathcal{F}, \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Random Variables – Formal Definition III

Distribution of X

The *distribution* of X, written $X \sim D$ or $\mathbb{P}_X$, is the pushforward measure

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}).$$

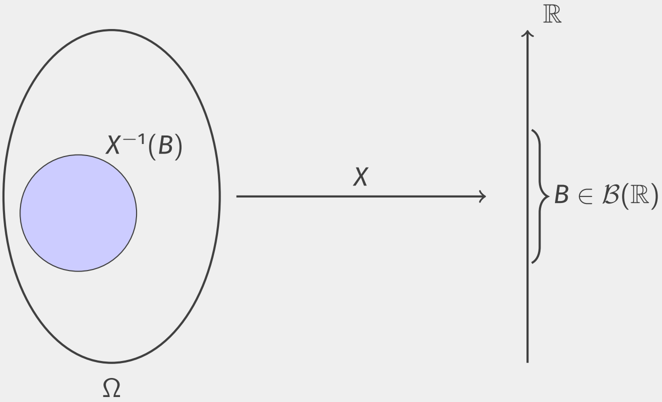
Random Variables – Formal Definition IV

Hint on granularity

The σ-algebra $\mathcal{F}$ determines the *level of detail* we can see.

- If $\mathcal{F}$ is coarse, its atoms are large sets, and the distribution cannot distinguish points inside them.
- Any measurable function (in particular X and thus $\mathbb{P}_X$) must be *constant on atoms of* $\mathcal{F}$.
- Intuition: the σ-algebra acts like the “resolution” of our probabilistic lens.

Random Variables – Formal Definition



Idea: A random variable X maps outcomes $\omega \in \Omega$ to values in $\mathbb{R}$. Measurability requires that the preimage $X^{-1}(B)$ of any Borel set B is an event in $\mathcal{F}$.

Conditional Probability Distribution I

Conditional Distribution

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. For a random variable $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the *conditional distribution of X given $\mathcal{G}$* is a probability kernel

$$\mathbb{P}(X \in B \mid \mathcal{G}) : \Omega \to [0, 1], \quad B \in \mathcal{B}(\mathbb{R}),$$

which is $\mathcal{G}$-measurable in ω, and satisfies

$$\int_{\omega \in A} \mathbb{P}(X \in B \mid \mathcal{G})(\omega)\, d\mathbb{P}(\omega) = \mathbb{P}\big(\{X \in B\} \cap A\big), \quad \forall A \in \mathcal{G}\,,$$

where

$$\{X \in B\} \;=\; \{\omega \in \Omega : X(\omega) \in B\}.$$

Conditional Probability Distribution II

Equivalent viewpoint for atoms of $\mathcal{G}$

If $A \in \mathcal{G}$ is an atom with $\mathbb{P}(A) > 0$, then the conditional distribution restricted to A can be written as a normalized conditional probability:

$$\mathbb{P}(X \in B \mid \mathcal{G})(\omega) = \frac{\mathbb{P}(\{X \in B\} \cap A)}{\mathbb{P}(A)}, \quad \forall \omega \in A.$$

This shows that the conditional distribution is a piecewise constant function on the atoms of $\mathcal{G}$.

Conditioning on a σ-algebra (intuition) I

Formally, for all $A \in \mathcal{G}$ and $B \in \mathcal{B}(\mathbb{R})$:

$$\int_A \mathbb{P}(X \in B \mid \mathcal{G})(\omega)\, d\mathbb{P}(\omega) \;=\; \mathbb{P}(\{X \in B\} \cap A).$$

How this creates coarse graining:

- The left-hand side forces $\mathbb{P}(X \in B \mid \mathcal{G})$ to be $\mathcal{G}$-measurable $\Rightarrow$ for fixed B it must be constant on atoms of $\mathcal{G}$.
- The right-hand side does not depend on ω! It involves the fine event $\{X \in B\} \in \mathcal{F}$, but restricted by $A \in \mathcal{G}$.
- Thus, within each atom A_i of $\mathcal{G}$, the conditional probability is the *same for all* $\omega \in A_i$ when B is fixed:

$$\mathbb{P}(X \in B \mid \mathcal{G})(\omega) = \frac{\mathbb{P}(\{X \in B\} \cap A_i)}{\mathbb{P}(A_i)} \quad (\omega \in A_i).$$

Conditioning on a σ-algebra (intuition) II

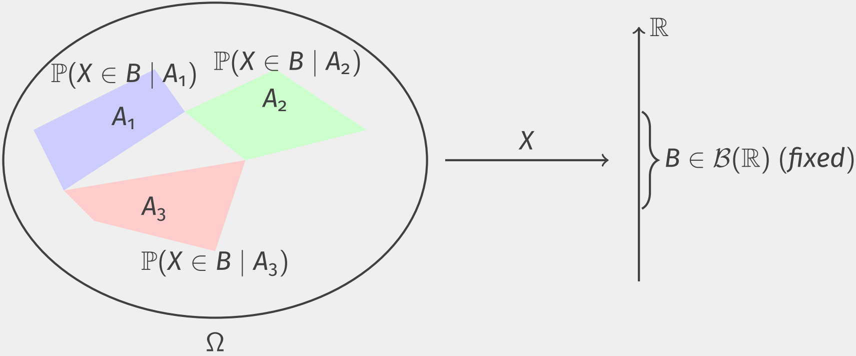
- Each atom $A_i \in \mathcal{G}$ is a "bucket of indistinguishable outcomes" according to the information in $\mathcal{G}$.
- The conditional distribution $\mathbb{P}(X \in B \mid \mathcal{G})$ is $\mathcal{G}$-measurable, so it cannot vary inside an atom.
- Intuition: "Inside the bucket, all outcomes look the same, so the conditional probability is constant."

Metaphor

Conditioning on $\mathcal{G}$ is like lowering the resolution of your camera:

- With $\mathcal{F}$ you see the exact outcome ω (maximum resolution).
- With $\mathcal{G}$ you only know "which atom A_i happened."
- Therefore the conditional distribution can only vary *between atoms*, never within them.

Conditioning on a σ-algebra (intuition) III



$\mathbb{P}(X \in B \mid A_1)$
$\mathbb{P}(X \in B \mid A_2)$
A_1
A_2
A_3
$\mathbb{P}(X \in B \mid A_3)$
Ω
X
$\mathbb{R}$
$B \in \mathcal{B}(\mathbb{R})$ (*fixed*)

Key idea: $\mathcal{G}$ partitions Ω into sets $A_1, A_2, A_3, \ldots$. The conditional distribution is a *function*:

$$\mathbb{P}(X \in B \mid A_i) = \mathbb{P}(\{X \in B\} \cap A_i)/\mathbb{P}(A_i) \quad \text{if } \omega \in A_i\,,$$

where $\{X \in B\} \in \mathbb{F}$.

Worked Example: Conditional Distribution I

Die Roll

Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ with uniform probability, and let $X(\omega) = \omega$.

Define the coarse σ-algebra $\mathcal{G}$ with atoms

$$A_1 = \{1, 2, 3\}, \quad A_2 = \{4, 5, 6\}.$$

Step 1: Fix $B \subseteq \mathbb{R}$. Define B such that

$$\{X \in B\} = \{1, 2\} \subseteq \Omega.$$

Step 2: Compute conditional probabilities.

$$\mathbb{P}(X \in B \mid A_1) = \frac{\mathbb{P}(\{1, 2\})}{\mathbb{P}(A_1)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3},$$

Worked Example: Conditional Distribution II

$$\mathbb{P}(X \in B \mid A_2) = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(A_2)} = 0.$$

Step 3: Result.

$$\mathbb{P}(X \in \{1, 2\} \mid \mathcal{G})(\omega) = \begin{cases} \frac{2}{3}, & \omega \in A_1, \\ 0, & \omega \in A_2. \end{cases}$$

Intuition

The conditional distribution is constant on each atom A_i of $\mathcal{G}$, but changes when B changes.

Two Perspectives on Random Variables

Probability Theory

- Random variable X is a measurable function

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

- σ-algebras model what events we can distinguish.
- Conditional probability defined via integrals and measurability.

Statistics (Applied View)

- Focus on the *distribution* of X only.
- Write directly: $X \sim \mathcal{N}(\mu, \sigma^2)$.
- Parameters μ, σ^2 are estimated from data.
- Rarely mention Ω or $\mathcal{F}$ explicitly.

Key idea: Kolmogorov's framework ensures rigor; statistics works with the *laws* of random variables.

Reading

- [Hacking, 2001]

References

Hacking, I. (2001).
An introduction to probability and inductive logic.
Cambridge university press.

Kolmogoroff, A. (1933).
Grundbegriffe der wahrscheinlichkeitsrechnung.