Albert-Ludwigs-Universität Freiburg

Lecture Notes

Basics in Applied Mathematics

Winter term 2024/2025

Part I: Stochastics

Contents

1 Probability

Motivation 1.1. *(i) How large is the probability to throw (with one single throw) a "6" with a dice? Result:*

$$\frac{1}{6} = \frac{1}{\textit{Number of possible outcomes}}.$$

(ii) Probability to get "10" as the sum of two dice rolls?

- *Favorable outcomes:* $(6,4)$, $(5,5)$, $(4,6)$
- *Possible outcomes:* $6^2 = 36$

$$\mathbb{P}(\textit{Sum } = 10) = \frac{\textit{Number of favorable outcomes}}{\textit{Number of possible outcomes}} = \frac{1}{12}.$$

The set $\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$ *is called sample space. The event of getting* 10 *as the sum can be considered as subset of* Ω*, namely as* $A = \{(4,6),(5,5),(6,4)\}$*. Other events:*

- *Sum uneven* $B = \{(1,2),(1,4),\ldots,(5,6)\}$
- *In the first (out of two) throws a "6":* $C = \{(6,1),(6,2),\ldots,(6,6)\}$
- *Doubles* $D = \{(1,1),\ldots,(6,6)\}$

For all events, we can formally specify the probability as

$$\frac{\textit{Number of favorable outcomes}}{\textit{Number of possible outcomes}},$$

for instance

$$\mathbb{P}(D) = \frac{|D|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

That is, $\mathbb{P}$ *assigns a probability to every subset of* Ω*. One can "calculate" with these probabilities:*

- $B \cap C =$ *Sum uneven and first throw a "6" or*
- $C \cup D =$ *first throw a "6" or doubles*

$$\mathbb{P}(C \cup D) = \frac{|C \cup D|}{|\Omega|} = \frac{11}{36} = \mathbb{P}(C) + \mathbb{P}(D) - \mathbb{P}(C \cap D).$$

If two sets E *and* F *are disjoint, i.e.* $E \cap F = \emptyset$*, then*

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F).$$

This property is the additivity of $\mathbb{P}$*.*

Definition 1.2 (Axioms of Kolmogorov)**.** *Let Ω be a non-empty countable set with power set $\mathcal{P}(\Omega)$. A map $\mathbb{P}: \mathcal{P}(\Omega) \to [0,1]$ is called (discrete) probability distribution on $\mathcal{P}(\Omega)$, if*

(i) $\mathbb{P}(\Omega) = 1$,

(ii) $\mathbb{P}(A) \geq 0$ *for all* $A \in \mathcal{P}(\Omega)$,

(iii) $\mathbb{P}\Big(\cup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *for all pairwise disjoint sets* $A_i \in \mathcal{P}(\Omega)$ *(σ-additivity).*

The triple $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ is called discrete probability space.

[One also writes $\sum_{i=1}^{\infty} A_i$ for $\cup_{i=1}^{\infty} A_i$ if the sets A_i are pairwise disjoint.]

Example 1.3 (Laplace distribution)**.** *Ω finite, $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$.*

Example 1.4 (Coin flip)**.** *Throw a coin to the ground. Let H be the event that the head is on top, T the event trails, i.e. that the number is on top. Then $\Omega = \{H, T\}$. If $p = \mathbb{P}(H)$, then $\mathbb{P}(T) = 1 - p$. We have $p = 1/2$ for a fair coin. If $p \neq 1/2$, then $\mathbb{P}$ is not a Laplace distribution.*

Proposition 1.5 (Properties of a probability distribution)**.** *Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a discrete probability space, $A, A_i \in \mathcal{P}(\Omega)$ for $i \in \mathbb{N}$. Then*

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

(ii) $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$

(iii) $\mathbb{P}\Big(\cup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

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

Proof. (i)–(iii): Homework. (iv): $A \cup B = (A \setminus B) + (B \setminus A) + A \cap B$.

$$\overset{\text{Kolm.}(iii)}{\Longrightarrow} \quad \begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B) \\ &= \big(\mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)\big) + \big(\mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)\big) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B). \end{aligned}$$

□

Remark 1.6. *We had required that Ω is countable in Definition 1.2. Slightly more general, one can request the discretization of a probability measure in the following sense: There exists a countable set $\Omega_0 \subset \Omega$ such that $\mathbb{P}(\Omega_0^c) = 0$. If such Ω_0 does not exist either, we leave the special case of discrete probability spaces. Then one restricts attention to specific σ-algebras $\mathcal{A}$ instead of $\mathcal{P}(\Omega)$ → measure theory.*

Example 1.7. *Bet: In the audience are at least two students which have the same birthday. Idea: The bet is profitable if*

$$\mathbb{P}(A) > \mathbb{P}(A^c) \overset{\text{Prop.1.5}(i)}{\iff} \mathbb{P}(A) > 1/2.$$

Model: Everyone has chosen its birthday at random out of the 365 days in the year. Assumption: No leap-year, all birthdays are equally probable.

$r =$ *Number of students*

$A = \{(\omega_1, \dots, \omega_r) |\, \exists\, i \neq j \text{ with } \omega_i = \omega_j\}$

$A^c = \{(\omega_1, \dots, \omega_r) |\, \omega_i \neq \omega_j \ \forall\, i \neq j\}$

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\
&= 1 - \frac{|A^c|}{|\Omega|} \\
&= 1 - \frac{365 \cdot 364 \cdot \dots \cdot (365 - r + 1)}{365^r} \\
&= 1 - \Big[1\Big(1 - \frac{1}{365}\Big)\Big(1 - \frac{2}{365}\Big) \cdot \dots \cdot \Big(1 - \frac{r-1}{365}\Big)\Big] \\
&\approx 1 - \Big[\exp(0) \exp\Big(-\frac{1}{356}\Big) \dots \exp\Big(-\frac{r-1}{365}\Big)\Big] = 1 - \exp\Big(-\underbrace{\sum_{k=1}^{r-1} \frac{k}{365}}_{= \frac{r(r-1)}{730}}\Big),
\end{aligned}$$

where $e^x = 1 + x + \mathcal{O}(x^2) \approx 1 + x$ *for small* x *has been used.*

Discrete measures can be easily constructed from their probability mass function.

Definition 1.8 (Probability mass function)**.** *Let* $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ *be a discrete probability space. The map* $f : \Omega \to [0,1],\ \omega \mapsto \mathbb{P}(\{\omega\})$ *is called probability mass function of* $\mathbb{P}$*.*

Vice versa, for a map $\pi : \Omega \to [0,1]$ with $\sum_{\omega \in \Omega} \pi(\omega) = 1$,

$$\nu(A) := \sum_{\omega \in \Omega \cap A} \pi(\omega) \quad \text{for all } A \in \mathcal{P}(\Omega) \text{ (where the empty sum is set to zero)},$$

defines obviously a discrete probability measure on $(\Omega, \mathcal{P}(\Omega))$. That is, for any countable state space Ω, there is a bijection between probability measures on $(\Omega, \mathcal{P}(\Omega))$ and probability mass functions on Ω.

Example 1.9 (Once again the Laplace distribution)**.** $\Omega \neq \emptyset$ *finite,* $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ *for any* $A \in \mathcal{P}(\Omega)$*. Then the probability mass function is given by*

$$f(\omega) = \mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}.$$

Definition 1.10. *For any $0 \leq k \leq n$ and $n \in \mathbb{N}$, the binomial coefficient is defined as*

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}.$$

Example 1.11 (Drawing without order and without putting back)**.** *How many possibilities are there to choose 6 numbers out of 49?*

$$\frac{49 \cdot 48 \cdot \cdots \cdot 44}{6 \cdot 5 \cdot \cdots \cdot 1} = \binom{49}{6} = 13983816.$$

Example 1.12 (Binomial distribution)**.** *Suppose there is an urn containing N balls, R of them are red and $N-R$ are white. Consecutively, we take n balls out of the urn, where the respective ball is put back after every draw. The balls are assumed to be enumerated. Without loss of generality (W.l.o.g. for short), the first R balls are the red ones.*

- *Sample space:* $\Omega = \{(\omega_1, \ldots, \omega_n) : 1 \leq \omega_i \leq N \text{ for } i = 1, \ldots, n\}$
- *Probability distribution on $(\Omega, \mathcal{P}(\Omega))$: Laplace distribution (everey n-tuple is equally probable)*

Question: How large is the probability that there are r red balls in the sample?

- *Event* $E_r = \{(\omega_1, \ldots, \omega_n) : |\{i : 1 \leq \omega_i \leq R\}| = r\}$

In order to determine the cardinality of E_r, we rewrite it as union of disjoint sets E_I, where $I \subset \{1, \ldots, n\}$ contains the number of those draws where the taken ball happened to be a red one:

$$E_I = \Big\{(\omega_1, \ldots, \omega_n) : \omega_i \in \{1, \ldots, R\} \text{ for } i \in I \text{ and } \omega_i \in \{R+1, \ldots, N\} \text{ for } i \in I^c\Big\}$$

and

$$E_r = \bigcup_{\substack{I \subset \{1,\ldots,n\}: \\ |I| = r}} E_I.$$

As there are $\binom{n}{r}$ of such subsets and all of them having cardinality $R^r(N-R)^{n-r}$,

$$|E_r| = \binom{n}{r} R^r (N-R)^{n-r}.$$

Using that $\mathcal{P}$ is the Laplace distribution and $|\Omega| = N^n$, we find

$$\mathbb{P}(E_r) = \frac{|E_r|}{|\Omega|} = \binom{n}{r} \Big(\frac{R}{N}\Big)^r \Big(1 - \frac{R}{N}\Big)^{n-r}.$$

Since the sample space is a disjoint union of $E_0, \dots, E_n$,

$$p(r) := \mathbb{P}(E_r),\ r \in \{0, \dots, n\},$$

defines a probability mass function on $\{0, \dots, n\}$. The corresponding distribution on $(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}))$ is called binomial distribution.

2 Independence, product spaces and conditional probability

2.1 Stochastic independence

Definition 2.1 (Stochastic independence)**.** *Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a discrete probability space. Two events $A, B \in \mathcal{P}(\Omega)$ are called stochastically independent if*

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

Example 2.2. *(i) We roll a (fair) dice and define the events*

- *$A =$ "Outcome is even"*
- *$B =$ "Outcome can be divided by 3".*

These events are independent by the above definition, because

$$\mathbb{P}(A) = \frac{1}{2},\ \mathbb{P}(B) = \frac{1}{3} \quad \text{and} \quad \mathbb{P}(\underbrace{A \cap B}_{\{6\}}) = \frac{1}{6}.$$

(ii) We draw twice with putting back out of an urn which contains 3 red and 5 white balls and study the events

- *$A =$ "First ball is red"*
- *$B =$ "Second ball is white".*

Then

$$\mathbb{P}(A) = \frac{3}{8} \cdot \frac{8}{8} = \frac{3}{8},\ \mathbb{P}(B) = \frac{8}{8} \cdot \frac{5}{8} = \frac{5}{8} \quad \text{and} \quad \mathbb{P}(A \cap B) = \frac{3}{8} \cdot \frac{5}{8}.$$

Thus, the two events A and B are independent.

We now extend the notion of independence to more than two events.

Definition 2.3. *Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a discrete probability space. The events $A_1, \dots, A_n$ are called stochastically independent if*

$$\mathbb{P}\big(A_{i_1} \cap \dots \cap A_{i_k}\big) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_k})$$

for every $k \in \{1, \dots, n\}$ and each collection of indices $1 \le i_1 < \dots < i_k \le n$.

With this definition, we ensure the desirable property that a subfamily of a family of independent events is itself independent.

Example 2.4 (Twice flipping a coin)**.** *Sample space* $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$*, Laplace experiment. We study the three events*

- $A = \{(H,H),(H,T)\}$ *"first throw head":* $\mathbb{P}(A) = \frac{2}{4} = \frac{1}{2}$,
- $B = \{(T,T),(H,T)\}$ *"second throw number":* $\mathbb{P}(B) = \frac{1}{2}$,
- $C = \{(H,T),(T,H)\}$ *"head exactly once":* $\mathbb{P}(C) = \frac{1}{2}$.

Are these events jointly independent?

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(B) \;\Leftrightarrow\; A, B \text{ independent} \\
\mathbb{P}(A \cap C) &= \frac{1}{4} = \mathbb{P}(A) \cdot \mathbb{P}(C) \;\Leftrightarrow\; A, C \text{ independent} \\
\mathbb{P}(B \cap C) &= \frac{1}{4} = \mathbb{P}(B) \cdot \mathbb{P}(C) \;\Leftrightarrow\; B, C \text{ independent}
\end{aligned}$$

but

$$\mathbb{P}(A \cap B \cap C) = \underbrace{\mathbb{P}(\{(H,T)\})}_{=\frac{1}{4}} \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C).$$

$\Longrightarrow$ *The sets are pairwise stochastically independent, but* A, B, C *are not jointly stochastically independent.*

Theorem 2.5. *Assume that* $A_1, \ldots, A_n$ *are stochastically independent. Then* $B_1, \ldots, B_n$ *with* $B_i \in \{A_i, A_i^c\}$ *are also stochastically independent.*

Proof. Induction on n.

$\underline{n = 2 \text{ (Initial case):}}$

$$\begin{aligned}
& (A_1 \cap A_2) + (A_1 \cap A_2^c) = A_1 \\
\Rightarrow\; & \underbrace{\mathbb{P}(A_1 \cap A_2)}_{=\mathbb{P}(A_1)\cdot\mathbb{P}(A_2)} + \mathbb{P}(A_1 \cap A_2^c) = \mathbb{P}(A_1) \\
\Rightarrow\; & \mathbb{P}(A_1 \cap A_2^c) = \mathbb{P}(A_1) \cdot \Big(1 - \mathbb{P}(A_2)\Big) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2). \\
\Rightarrow\; & A_1, A_2^c \text{ are stochastically independent.}
\end{aligned}$$

The proof for the independence of A_1^c, A_2 is carried out analogously.

$\underline{\text{Induction step } n-1 \to n\text{:}}$ Grant the claim for $n-1$ and let $A_1, \ldots, A_n$ be stochastically independent. Let $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ for $m = n-1$. Then by definition,

$A_{i_1}, \ldots, A_{i_m}$ are stochastically independent. By the induction hyposthesis, $B_{i_1}, \ldots, B_{i_m}$ are stochastically independent. Hence, it remains to show:

$$\mathbb{P}\big(B_1 \cap \ldots B_n\big) = \mathbb{P}(B_1) \cdot \ldots \cdot \mathbb{P}(B_n). \tag{2.1}$$

W.l.o.g., let $B_1 = A_1^c, \ldots, B_k = A_k^c, B_{k+1} = A_{k+1}, \ldots, B_n = A_n$. We prove (2.1) itself by induction on k:

$\underline{k = 1 \text{ (Initial case):}}$

$\mathbb{P}(A_1^c \cap A_2 \cap \cdots \cap A_n) + \mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_2 \cap \cdots \cap A_n)$

$$\begin{aligned} \Rightarrow \quad \mathbb{P}(A_1^c \cap A_2 \cap \cdots \cap A_n) &= \big(1 - \mathbb{P}(A_1)\big) \cdot \mathbb{P}(A_2) \cdot \ldots \cdot \mathbb{P}(A_n) \\ &= \mathbb{P}(A_1^c) \cdot \mathbb{P}(A_2) \cdot \ldots \cdot \mathbb{P}(A_n). \end{aligned}$$

$\underline{\text{Induction step } k-1 \to k\text{:}}$

$$\begin{aligned} &\mathbb{P}(A_1^c \cap \cdots \cap A_k^c \cap A_{k+1} \cap \cdots \cap A_n) + \mathbb{P}(A_1^c \cap \cdots \cap A_{k-1}^c \cap A_k \cap \cdots \cap A_n) \\ &\quad = \mathbb{P}(A_1^c \cap \cdots \cap A_{k-1}^c \cap A_{k+1} \cap \cdots \cap A_n) \\ &\quad \overset{\text{Ind. hyp.}}{=} \mathbb{P}(A_1^c) \cdot \ldots \cdot \mathbb{P}(A_{k-1}^c) \cdot \mathbb{P}(A_{k+1}) \cdot \ldots \cdot \mathbb{P}(A_n). \end{aligned}$$

$$\begin{aligned} \Rightarrow \quad &\mathbb{P}(A_1^c \cap \cdots \cap A_k^c \cap A_{k+1} \cap \cdots \cap A_n) \\ &\quad = \mathbb{P}(A_1^c) \cdot \ldots \cdot \mathbb{P}(A_{k-1}^c) \cdot \big(1 - \mathbb{P}(A_k)\big) \cdot \mathbb{P}(A_{k+1}) \cdot \ldots \cdot \mathbb{P}(A_n) \\ &\quad = \mathbb{P}(A_1^c) \cdot \ldots \cdot \mathbb{P}(A_k^c) \cdot \mathbb{P}(A_{k+1}) \cdot \ldots \cdot \mathbb{P}(A_n). \end{aligned}$$

□

2.2 Product spaces and product experiments

Motivation 2.6. *Two students – Student* 1 *from Freiburg, Student* 2 *from Konstanz – are doing an experiment at the same time: Student* 1 *is flipping a (fair) coin, Student* 2 *is rolling a (fair) dice. Both have their own probability spaces,* $\Omega_1 = \{H, T\}$ *with Laplace distribution* $\mathbb{P}_1$ *on* $\mathcal{P}(\Omega_1)$ *and* $\Omega_2 = \{1, \ldots, 6\}$ *with Laplace distribution* $\mathbb{P}_2$ *on* $\mathcal{P}(\Omega_2)$*. As concerns the probability of the event "Student* 1 *throws a head, Student* 2 *gets a "*6*"", we are facing the problem that the event neither belongs to* $\mathcal{P}(\Omega_1)$ *nor* $\mathcal{P}(\Omega_2)$*.*

Definition 2.7 (Product space)**.** *Let* $(\Omega_1, \mathcal{P}(\Omega_1), \mathbb{P}_1), \ldots, (\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$ *be discrete probability spaces. The* $\underline{\text{product space}}$ $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ *is the discrete probability space with*

$$\Omega = \Omega_1 \times \cdots \times \Omega_n = \big\{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i \text{ for all } i = 1, \ldots, n\big\}$$

and probability mass function $p(\omega_1, \ldots, \omega_n) = p_1(\omega_1) \cdot \ldots \cdot p_n(\omega_n)$. *The probability measure with* p *as probability mass function is called product probability measure and often denoted as* $\mathbb{P} = \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$.

For product spaces, we have the following result.

Theorem 2.8. *Let* $(\Omega_1, \mathcal{P}(\Omega_1), \mathbb{P}_1), \ldots, (\Omega_n, \mathcal{P}(\Omega_n), \mathbb{P}_n)$ *be discrete probability spaces with product space* $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. *Let* $A_i \subset \Omega_i$ *and* $A_i' := \{\omega \in \Omega : \omega_i \in A_i\}$, $i = 1, \ldots, n$. *Then*

$$\mathbb{P}(A_i') = \mathbb{P}_i(A_i)$$

for $i = 1, \ldots, n$ *and the events* $A_1', \ldots, A_n'$ *are stochastically independent.*

Proof. Note that subsequently, rearrangement of the potentially infinite summations is admissible because all summands are non-negative.

$$\begin{aligned}
\mathbb{P}(A_i') &= \sum_{\omega \in A_i'} p(\omega) \\
&= \sum_{(\omega_1, \ldots, \omega_n) : \omega_i \in A_i} p_1(\omega_1) \cdot \ldots \cdot p_n(\omega_n) \\
&= \underbrace{\left(\sum_{\omega_1 \in \Omega_1} p_1(\omega_1) \right)}_{=1} \cdot \ldots \cdot \left(\sum_{\omega_i \in A_i} p_i(\omega_i) \right) \cdot \ldots \cdot \underbrace{\left(\sum_{\omega_n \in \Omega_n} p_n(\omega_n) \right)}_{=1} \\
&= \sum_{\omega_i \in A_i} p_i(\omega_i) \\
&= \mathbb{P}_i(A_i).
\end{aligned}$$

For indices $1 \leq i_1 < \cdots < i_k \leq n$, we obtain

$$\begin{aligned}
\mathbb{P}(A_{i_1}' \cap \cdots \cap A_{i_k}') &= \sum_{\substack{\omega \in \Omega : \\ \omega_1 \in A_{i_1}, \ldots, \omega_{i_k} \in A_{ik}}} p_{i_1}(\omega_{i_1}) \cdot \ldots \cdot p_{i_k}(\omega_{i_k}) \prod_{j \notin \{i_1, \ldots, i_k\}} p_j(\omega_j) \\
&= \left(\sum_{\omega_{i_1} \in A_{i_1}} p_{i_1}(\omega_{i_1}) \right) \cdot \ldots \cdot \left(\sum_{\omega_{i_k} \in A_{i_k}} p_{i_k} \right) \\
&= \mathbb{P}_{i_1}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}_{i_k}(A_{i_k}) \\
&= \mathbb{P}(A_{i_1}') \cdot \ldots \cdot \mathbb{P}(A_{i_k}').
\end{aligned}$$

Hence, the events $A_{i_1}', \ldots, A_{i_k}'$ are independent. □

With the identity $A_1' \cap \cdots \cap A_n' = A_1 \times \cdots \times A_n$ is the meaning of Theorem 2.8, that

$$\mathbb{P}(A_1 \times \cdots \times A_n) = \mathbb{P}_1(A_1) \cdot \ldots \cdot \mathbb{P}_n(A_n)$$

on the product space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. More general, this formula can be taken as starting point for the definition of the product probability measure on not necessarily discrete spaces. Then, it has to be shown that $\mathbb{P}$ is uniquely determined via the probability of such Cartesian products. $\rightarrow$ Measure theory

Example 2.9 (Again the two students from Freiburg and Konstanz)**.** *With the notation of Motivation 2.6, define* $\Omega := \Omega_1 \times \Omega_2$ *and let* $\mathbb{P}$ *be the product measure on* $\mathcal{P}(\Omega)$ *with components* $\mathbb{P}_1$ *and* $\mathbb{P}_2$*. Then the event* $\{(H,6)\} =$*'Student* 1 *throws a head, Student* 2 *gets a "*6*"" is a one-element subset of* Ω *and*

$$\begin{aligned}\mathbb{P}(\{H,6)\}) &= \mathbb{P}\Big(\big(\{H\} \times \Omega_2\big) \cap \big(\Omega_1 \times \{6\}\big)\Big) \\ &= \mathbb{P}(\{H\} \times \Omega_2) \cdot \mathbb{P}(\Omega_1 \times \{6\}) \\ &= \mathbb{P}_1(\{H\}) \cdot \mathbb{P}_2(\{6\}) = \frac{1}{2} \cdot \frac{1}{6}.\end{aligned}$$

That is, under the product measure, the events in Experiment 1 *are independent of the events in Experiment* 2*.*

2.3 Conditional probabilities

Our next goal is to develop a concept to describe dependencies of events. What is the probability for an event A if we now that another event B has occurred?

Definition 2.10. *Let* $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ *be a discrete probability space,* $A, B \in \mathcal{P}(\Omega)$ *with* $\mathbb{P}(A) > 0$*. Then the conditional probability of* B *given* A *is defined as*

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

Example 2.11 (Twice tossing a dice)**.** *Consider the events*

$$A = \big\{(\omega_1, \omega_2) \in \{1, \ldots, 6\} \times \{1, \ldots, 6\} : \omega_1 + \omega_2 = 12\big\} \text{ and}$$

$$B = \big\{(\omega_1, \omega_2) \in \{1, \ldots, 6\} \times \{1, \ldots, 6\} : \omega_1 = 5\big\}.$$

Then $\mathbb{P}(A) = \frac{1}{36}$, $\mathbb{P}(B) = \frac{1}{6} \cdot \frac{6}{6} = \frac{1}{6}$ *and*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(\overbrace{A \cap B}^{=\emptyset})}{\mathbb{P}(B)} = 0.$$

Remark 2.12. *For independent events* A *and* B *with* $\mathbb{P}(A) > 0$*, we find*

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

that is, the probability of B given A is equal to the probability of B. Vice versa, $\mathbb{P}(B|A) = \mathbb{P}(B)$ implies

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) = \mathbb{P}(A) \cdot \mathbb{P}(B),$$

i.e. the two events are independent.

Theorem 2.13 (Law of the total probability)**.** *Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be a discrete probability space, $B_1, \ldots, B_n \in \mathcal{P}(\Omega)$ with $\cup_{i=1}^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ whenever $i \neq j$; finally $\mathbb{P}(B_j) > 0$ for $j = 1, \ldots, n$. Then*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)$$

for all $A \in \mathcal{P}(\Omega)$.

Proof.

$$\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}\Big(\sum_{i=1}^{n} (A \cap B_i) \Big) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i).$$

□

Theorem 2.14 (Bayes formula)**.** *Let $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ be as above, $A, B_1, \ldots, B_n \in \mathcal{P}(\Omega)$ with $\cup_{i=1}^n B_i = \Omega$ and $B_i \cap B_j = \emptyset$ whenever $i \neq j$; finally $\mathbb{P}(A) > 0$ and $\mathbb{P}(B_j) > 0$ for $j = 1, \ldots, n$. Then*

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}.$$

Proof.

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j \cap A)}{\mathbb{P}(A)} \overset{\text{Thm 2.13}}{=} \frac{\mathbb{P}(A|B_j) \cdot \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i) \cdot \mathbb{P}(B_i)}.$$

□

Example 2.15. *Suppose there are 6 urns, each of them containing five balls. There are 5 white balls in the first urn, 4 white balls and one black ball in the second one, and so on.*

○ ○ ○ ○ ○	*Urn 1*
○ ○ ○ ○ ●	*Urn 2*
○ ○ ○ ● ●	*Urn 3*
○ ○ ● ● ●	*Urn 4*
○ ● ● ● ●	*Urn 5*
● ● ● ● ●	*Urn 6*

The experiment is as follows: We roll a dice and choose the urn corresponding to the dots obtained. Then we draw consecutively a ball out of this chosen urn with putting it back after each draw.

(i) Consider the events

- $B_i =$ *"We draw out of the i'th urn"*
- $A_1 =$ *"The first ball taken out is black"*
- $A_2 =$ *"The second ball taken out is black".*

$\mathbb{P}(A_1)$*?,* $\mathbb{P}(A_2|A_1)$*?*

$B_1, \ldots, B_6$ *form a disjoint decomposition of the sample space,* $\mathbb{P}(B_i) = \frac{1}{6}$*. Moreover,* $\mathbb{P}(A_1|B_i) = \frac{i-1}{5}$*. Hence, by the law of the total probability,*

$$\mathbb{P}(A_1) = \sum_{i=1}^{6} \mathbb{P}(A_1|B_i) \cdot \mathbb{P}(B_i) = \sum_{i=1}^{n} \frac{i-1}{5} \cdot \frac{1}{6} = \frac{1}{30} \sum_{i=0}^{6} = \frac{1}{2}.$$

[Heuristically, this was clear for reasons of symmetry!]

Next, $\mathbb{P}(A_2|A_1) = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = 2\mathbb{P}(A_1 \cap A_2)$*. As we draw (independently) with replacement,*

$$\mathbb{P}(A_1 \cap A_2)|B_i) = \Big(\frac{i-1}{5}\Big)^2,$$

whence $\mathbb{P}(A_1 \cap A_2) = \frac{1}{6} \sum_{i=1}^{6} \left(\frac{i-1}{5}\right)^2 = \frac{1}{150} \sum_{i=0}^{5} i^2 = \frac{11}{30}$ *and*

$$\mathbb{P}(A_2|A_1) = \frac{11}{15} \quad \textit{(much(!) larger than } 1/2\textit{)}.$$

(ii) How large is the conditional probability, that we draw out of the i'th urn given the first ball taken out is black?

$$\mathbb{P}(B_i|A_1) = \frac{\mathbb{P}(A_1|B_i) \cdot \mathbb{P}(B_i)}{\sum_{k=1}^{6} \mathbb{P}(A_1|B_k) \cdot \mathbb{P}(B_k)} = \frac{\frac{i-1}{5} \cdot \frac{1}{6}}{\sum_{k=1}^{6} \frac{k-1}{5} \cdot \frac{1}{6}} = \frac{i-1}{15} \neq \frac{1}{6}.$$

3 Discrete random variables

Example 3.1 (Thrice flipping a coin)**.** $\Omega = \big\{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{0, 1\}, i = 1, 2, 3\big\}$, $\mathbb{P} =$ *Laplace distribution on* $(\Omega, \mathcal{P}(\Omega))$*, i.e.*

$$\mathbb{P}(A) = \frac{|A|}{|\omega|} = \frac{|A|}{8}.$$

Question: How large is the probability to obtain exactly k times Head?

Two possibilities:

(i) $A_k = \Big\{(\omega_1,\omega_2,\omega_3) \in \Omega : \sum_{i=1}^3 \omega_i = k\Big\}$ *and* $\mathbb{P}(A_k) = \frac{|A_k|}{8}$.

(ii) As derived in Example 1.11, the count of Heads is distributed according to $\mathrm{Bin}(3,\frac{1}{2})$ *(binomial distribution). Hereby, we use another sample space and another probability measure, namely*

$$\Omega^X = \{0,1,2,3\} \quad \text{and} \quad \mathbb{P}^X(\{k\}) = \binom{3}{k}\Big(\frac{1}{2}\Big)^k\Big(\frac{1}{2}\Big)^{3-k} = \binom{3}{k}\Big(\frac{1}{2}\Big)^3.$$

The probability space $(\Omega^X, \mathcal{P}(\Omega^X), \mathbb{P}^X)$ *can be obtained from* $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ *via the following map:*

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto \sum_{i=1}^3 \omega_i \ . \end{aligned}$$

Ω^X *is* $\operatorname{Image} X = \big\{x \in \mathbb{R} : \exists\, \omega \in \Omega \text{ with } x = X(\omega)\big\}$. $\mathbb{P}^X$ *is the probability measure on* $(\Omega^X, \mathcal{P}(\Omega^X))$ *with*

$$\mathbb{P}^X(B) := \mathbb{P}\big(\underbrace{X^{-1}(B)}_{\substack{\text{Preimage}\\ \text{of } B\\ \text{under } X}}\big) \quad \forall\, B \in \mathcal{P}(\Omega^X).$$

Recall the definition of the preimage: $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \subset \Omega$.

For instance:

$$\begin{aligned} \mathbb{P}^X(\{1\}) = \mathbb{P}\Big(X^{-1}(\{1\})\Big) &= \mathbb{P}\Big(\{\omega \in \Omega : X(\omega = 1)\}\Big) \\ &= \mathbb{P}\Big(\{(1,0,0),(0,1,0),(0,0,1)\}\Big) = \frac{3}{8}. \end{aligned}$$

Here, $\mathbb{P}^X = \mathrm{Bin}(3,\frac{1}{2})$. X *is called random variable,* $\mathbb{P}^X$ *the distribution on* $(\Omega^X, \mathcal{P}(\Omega^X))$ *induced by* X *(or just the distribution of* X*).*

Lemma and Definition 3.2. *Let* $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ *be a discrete probability space. A map* $X : \Omega \to \mathbb{R}$ *is called discrete random variable. The distribution* $\mathbb{P}^X$ *induced by* X*, given by* $\mathbb{P}^X(A) := \mathbb{P}\big(X^{-1}(A)\big)$ *for* $A \in \mathcal{P}(\Omega^X)$*, is a probability on* $(\Omega^X, \mathcal{P}(\Omega^X))$*. The function*

$$\begin{aligned} F_X : \mathbb{R} &\to [0,1] \\ x &\mapsto \mathbb{P}^X\big((-\infty, x] \cap \Omega^X\big) = \mathbb{P}(X \le x) = \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \le x\}\big) \end{aligned}$$

is called distribution function of X*.*

Proof. Homework: Prove that $\mathbb{P}^X$ is discrete a probability measure on $(\Omega^X, \mathcal{P}(\Omega^X))$ (axioms of Kolmogorov). □

Notation: If a random variable X on a probability space is distributed according to some probability measure μ, i.e. $\mathbb{P}^X = \mu$, one also writes $X \sim \mu$.

Remark 3.3. *One easily shows that F is monotonically increasing and right-continuous with $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. (Homework!)*

Poisson distribution and Poisson limit theorem. In "real world" applications, there may arise the problem to determine numerical values of the distribution function. For the binomial distribution $\mathrm{Bin}(n,p)$ with large parameter n, for instance, this evaluation for can be very costly, because (most of) the binomial coefficients become rather huge (Example 1.12).

Theorem 3.4 (Poisson limit theorem)**.** *Let $X_n \sim \mathrm{Bin}(n, p_n)$, $n \in \mathbb{N}$. If there exists $\lambda \in (0,\infty)$ with $np_n \to \lambda$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \forall\, k \in \mathbb{N}_0.$$

p_λ, given by $p_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k \in \mathbb{N}_0$, defines a probability mass function on $\mathbb{N}_0$.

Proof. For any fixed $k \in \mathbb{N}_0$, the following identity holds true.

$$\begin{aligned}
\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k!}\frac{1}{n^k}(np_n)^k\Big(1-\frac{np_n}{n}\Big)^n(1-p_n)^k \\
&= \underbrace{\Big(\frac{n}{n}\Big)\cdot\Big(\frac{n-1}{n}\Big)\cdot\ldots\cdot\Big(\frac{n-k+1}{n}\Big)}_{\to 1}\underbrace{(1-p_n)^{-k}}_{\to 1}\underbrace{\frac{(np_n)^k}{k!}}_{\to\frac{\lambda^k}{k!}}\underbrace{\Big(1-\frac{np_n}{n}\Big)^n}_{\to e^{-\lambda}}
\end{aligned}$$

Since $np_n \to \lambda$ by assumption and hence $p_n \to 0$,

$$\mathbb{P}(X_n = k) \longrightarrow e^{-\lambda}\frac{\lambda^k}{k!},$$

where we have used that $\lim_{n\to\infty}\big(1-\frac{x_n}{n}\big)^n = e^x$ for $x_n \to x$. With the series expansion of the exp-function

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^\lambda,$$

we obtain $\sum_{k=0}^{\infty} p_\lambda(k) = 1$, i.e. p_λ is a probability mass function on $\mathbb{N}_0$. □

Remark 3.5. *The probability measure on $(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$ with probability mass function p_λ is called Poisson distribution with parameter λ.*

Examples 3.6 (Some discrete distributions)**.**

(i) Bernoulli-distribution $B(p)$ with parameter $p \in [0,1]$:
This is a probability distribution on $(\{0,1\}, \mathcal{P}(\{0,1\})$ with probability mass function $f(1) = p,\ f(0) = 1-p$.

(ii) Binomial distribution $\mathrm{Bin}(n,p)$ with parameters $n \in \mathbb{N}$ and $p \in [0,1]$:
This is a probability distribution on $\big(\{0,1,\dots,n\}, \mathcal{P}(\{0,1,\dots,n\})\big)$ with probability mass function

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0,1,\dots,n.$$

(iii) Geometric distribution $G(p)$ with parameter $p \in [0,1]$:
This is a probability distribution on $(N, \mathcal{P}(\mathbb{N}))$ with probability mass function

$$f(k) = (1-p)^{k-1} p \quad \text{for } k \in \mathbb{N}.$$

[Distribution for the number of attempts until the first success]

3.1 Expectation and variance

The profit in gambling is random – what is the "expected" profit?

Definition 3.7. *Let X be a discrete random variable with probability mass function p_X, i.e. $\mathbb{P}^X(\{x\}) = \mathbb{P}(X^{-1}(\{x\})) = p_X(x)$. We say that the expectation of X exists if*

$$\sum_{x \in \Omega^X} |x| \cdot p_X(x) < \infty. \tag{3.1}$$

In this case, the expectation of X is defined as

$$\mathbb{E}X := \sum_{x \in \Omega^X} x \cdot p_X(x). \tag{3.2}$$

Remark. *The absolute summability requirement (3.1) ensures that the expression in (3.2) is well-defined, i.e. for any arrangement of the summands into a sequence, the series is finitely summable and its value is invariant under rearrangement of the summation order. Condition (3.1) is typically abbreviated as $\mathbb{E}|X| < \infty$.*

Examples 3.8. *(i) Let $X \sim B(p)$. Then $\mathbb{E}X = 0 \cdot (1-p) + 1 \cdot p = p$.*

(ii) Let X be Laplace distributed on $\{1, 2, \ldots, N\}$, i.e. $\mathbb{P}(X = k) = \frac{1}{N}$ for $k = 1, \ldots, N$. Then

$$\mathbb{E}X = \sum_{j=1}^{N} j \cdot \frac{1}{N} = \frac{1}{N} \frac{N(N+1)}{2} = \frac{N+1}{N}.$$

(iii) Let $X \sim \mathrm{Bin}(n, p)$, i.e. $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$. Then

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np \underbrace{\sum_{l=0}^{n-1} \binom{n-1}{l} p^l (1-p)^{(n-1)-l}}_{\substack{\text{Sum over probability mass function} \\ \text{of the } \mathrm{Bin}(n-1,p)\text{-distribution}}} \\
&= np.
\end{aligned}$$

Theorem 3.9 (Transformation formula for the expectation)**.** *Let X be a discrete random variable with probability mass function p_X and let $u : \Omega^X \to \mathbb{R}$ be a map with*

$$\sum_{x \in \Omega^X} |u(x)| p_X(x) < \infty.$$

Then the expectation $\mathbb{E}\big(u(X)\big)$ exists and is equal to

$$\mathbb{E}\big(u(X)\big) = \sum_{x \in \Omega^X} u(x) p_X(x).$$

Proof. The probability mass function of the random variable $Y = u \circ X$ is

$$p_Y(y) = \mathbb{P}(u \circ X = y) = \sum_{\substack{x \in \Omega^X: \\ u(x) = y}} p_X(x).$$

Plugging this expression into the definition of $\mathbb{E}Y$, we get

$$\mathbb{E}Y = \sum_y y p_Y(y) = \sum_y y \sum_{\substack{x \in \Omega^X: \\ u(x) = y}} p_X(x) = \sum_y \sum_{\substack{x \in \Omega^X: \\ u(x) = y}} u(x) p_X(x) = \sum_{x \in \Omega^x} u(x) p_X(x).$$

□

Theorem 3.10 (Properties of the expectation)**.** *Let X and Y be random variables on a joint probability space for which the expectations exist. Then, for all $a, b \in \mathbb{R}$, the following identities are satisfied:*

(i) $\mathbb{E}(aX) = a\mathbb{E}X$

(ii) $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$

(iii) $\mathbb{E}b = b$

(iv) $|\mathbb{E}X| \leq \mathbb{E}|X|$.

Proof. (i) With $u(x) = a \cdot x$, the transformation formula (Theorem 3.9) reveals

$$\mathbb{E}(aX) = \sum_x (ax) p_X(x) = a \sum_x x p_X(x) = a\mathbb{E}X.$$

(ii) With $\{x_1.x_2, \dots\} = \Omega^X$, $\{y_1, y_2, \dots\} = \Omega^Y$, $A_i := X^{-1}(\{x_i\})$, $B_i := Y^{-1}(\{y_i\})$, we obtain

$$\begin{aligned}
\mathbb{E}X + \mathbb{E}Y &= \sum_i x_i \mathbb{P}(A_i) + \sum_j y_j \mathbb{P}(B_j) \\
&= \sum_i x_i \mathbb{P}\Big(A_i \cap \sum_j B_j\Big) + \sum_j y_j \mathbb{P}\Big(\Big(\sum_i A_i\Big) \cap B_j\Big) \\
&= \sum_{i,j} x_i \mathbb{P}(A_i \cap B_j) + \sum_{i,j} y_j \mathbb{P}(A_i \cap B_j) \\
&= \sum_{i,j} (x_i + y_j) \mathbb{P}(A_i \cap B_j) \\
&= \sum_u \sum_{\substack{i,j \in \mathbb{N}: \\ x_i + y_j = u}} (x_i + y_j) \mathbb{P}(A_i \cap B_j) \\
&= \sum_u u \cdot \underbrace{\sum_{\substack{i,j \in \mathbb{N}: \\ x_i + y_j = u}} \mathbb{P}(A_i \cap B_j)}_{\mathbb{P}(X+Y=u)} = E(X + Y).
\end{aligned}$$

(iii) Clear.

(iv) $\mathbb{E}|X| = \sum_{x \in \Omega^X} |x| p_X(x) = \sum_{x \in \Omega^X} \big| x p_X(x) \big| \geq \Big| \sum_{x \in \Omega^X} x p_X(x) \Big| = |\mathbb{E}X|.$ □

Remark. *(i) In general, $\mathbb{E}g(X) = g(\mathbb{E}X)$ is NOT true!*

(ii) Prove by induction that $\mathbb{E}\Big(\sum_{i=1}^n X_i\Big) = \sum_{i=1}^n \mathbb{E}X_i$ for discrete random variables $X_1, \dots, X_n$ with existing expectations.

Example 3.11. *(i) Binomial distribution:* $X \sim \mathrm{Bin}(n,p)$

X is the number of successes in n independent Bernoulli experiments – for instance, X counts how often 6 appears when n times rolling a dice (in this case, $p = 1/6$).

$$\begin{aligned} A_i' :=& \text{“The outcome in the } i\text{'th experiment is success”} \\ =& \big\{\omega \in \Omega_1 \times \cdots \times \Omega_n : \omega_i = 1\big\}, \end{aligned}$$

$$\Omega_j = \{0,1\}\ \forall\, i = 1,\dots,n \text{ and } \mathbb{1}_{A_i'}(\omega) := \begin{cases} 1 & \text{if } \omega \in A_i' \\ 0 & \text{otherwise.} \end{cases}$$

Then $X = \sum_{i=1}^n \mathbb{1}_{A_i'}$ and therefore, $\mathbb{E}X = \sum_{i=1}^n \mathbb{E}\mathbb{1}_{A_i'} = np$.

(ii) Hypergeometric distribution: $Y \sim \mathcal{H}(N,R,n)$

Y is the number of collected red balls when drawing n times without replacement from an urn which contains R red and $N-R$ white balls

$$\mathbb{P}(Y = r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, \quad \text{where } \binom{n}{k} := 0 \text{ for } k < 0 \text{ and } k > n.$$

Two possibilities for evaluating the expectation:

(a)

$$\begin{aligned} \mathbb{E}Y &= \binom{N}{n}^{-1} \sum_{r=1}^{n} r \binom{R}{r}\binom{N-R}{n-r} \\ &= \binom{N}{n}^{-1} \sum_{r=1}^{n} R \binom{R-1}{r-1}\binom{N-R}{n-r} \\ &= R\binom{N}{n}^{-1} \sum_{r=1}^{n} \binom{R-1}{r-1}\binom{N-R}{n-r} \\ &= R\binom{N}{n}^{-1} \sum_{r=0}^{n-1} \binom{R-1}{r-1}\binom{N-1-(R-1)}{(n-1)-r} \\ &= R\binom{N}{n}^{-1} \binom{N-1}{n-1} = n\frac{R}{N}, \end{aligned}$$

(b) We define the event $B_i :=$ “The i'th drawn ball is red”. Then $Y = \mathbb{1}_{B_1} + \cdots + \mathbb{1}_{B_n}$ and

$$\mathbb{E}X = \sum_{i=1}^n \mathbb{E}\mathbb{1}_{B_i} = \sum_{i=1}^n \mathbb{P}(B_i) \overset{\text{Symmetry}}{=} n\frac{R}{N}.$$

Definition 3.12. *Let X be some discrete random variable with $\mathbb{E}(X^2) < \infty$. Then its variance is defined as*

$$\mathrm{Var}(X) := \mathbb{E}\big((X - \mathbb{E}X)^2\big).$$

Theorem 3.13 (Properties of the variance)**.** *Let X be a discrete random variable with $\mathbb{E}(X^2) < \infty$ and let $a, b \in \mathbb{R}$. Then*

(i) $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$

(ii) $\operatorname{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$

(iii) $\mathbb{E}\big((X-a)^2\big) = \operatorname{Var}(X) + \big(\mathbb{E}X - a\big)^2 \geq \operatorname{Var}(X)$.

Proof. (i)

$$\begin{aligned}\operatorname{Var}(aX+b) &= \mathbb{E}\Big(\big(aX + b - \mathbb{E}(aX+b)\big)^2\Big)\\ &= \mathbb{E}\Big(\big(ax + b - a\mathbb{E}X - b\big)^2\Big)\\ &= \mathbb{E}\Big(\big(a(X - \mathbb{E}X)\big)^2\Big) = a^2 \operatorname{Var}(X).\end{aligned}$$

(ii)

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}\Big(\big(X - \mathbb{E}X\big)^2\Big)\\ &= \mathbb{E}\big(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2\big)\\ &= \mathbb{E}(X^2) - 2\mathbb{E}\big(X\mathbb{E}X\big) + \mathbb{E}\big(\mathbb{E}X\big)^2\big)\\ &\overset{\text{Thm 3.10}}{=} \mathbb{E}(X^2) - 2(\mathbb{E}X)(\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2.\end{aligned}$$

(iii) Because of $\mathbb{E}\big(X - \mathbb{E}X\big) = \mathbb{E}X - \mathbb{E}X = 0$, we have

$$\begin{aligned}\mathbb{E}\big((X-a)^2\big) &= \mathbb{E}\Big(\big(X - \mathbb{E}X + \mathbb{E}X - a\big)^2\Big)\\ &= \mathbb{E}\big((X - \mathbb{E}X)^2\big) + 2\mathbb{E}\Big((X - \mathbb{E}X)(\mathbb{E}X - a)\Big) + \mathbb{E}\big((\mathbb{E}X - a)^2\big)\\ &= \operatorname{Var}(X) + 2\underbrace{\mathbb{E}(X - \mathbb{E}X)}_{=0}(\mathbb{E}X - a) + (\mathbb{E}X - a)^2\\ &= \operatorname{Var}(X) + \underbrace{(\mathbb{E}X - a)^2}_{\geq 0}.\end{aligned}$$

□

Example 3.14 (Variance of the binomial distribution)**.** *Let $X \sim \operatorname{Bin}(n, p)$. Then*

$$\begin{aligned}\mathbb{E}\big(X(X-1)\big) &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k (1-p)^{n-k}\\ &= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{(n-2)-(k-2)}\\ &= n(n-1)p^2.\end{aligned}$$

With $\mathbb{E}(X^2) = \mathbb{E}\big(X(X-1)\big) + \mathbb{E}X = n(n-1)p^2 + np$, *we deduce*

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

Theorem 3.15 (Markov and Chebyshev inequalities)**.** *Let X be some discrete random variable.*

(i) *Markov inequality: If $\mathbb{E}|X| < \infty$, then*

$$\mathbb{P}\big(|X| \geq \eta\big) \leq \frac{\mathbb{E}|X|}{\eta} \quad \text{for every } \eta > 0.$$

(ii) *Chebyshev inequality: If $\mathbb{E}(X^2) < \infty$, then we have for all $\eta > 0$ and $a \in \mathbb{R}$*

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

For $a = \mathbb{E}X$, we get in particular

$$\mathbb{P}\big(|X - \mathbb{E}X| \geq \eta\big) \leq \frac{\mathrm{Var}(X)}{\eta^2}.$$

Proof. (i)

$$\mathbb{P}\big(|X| \geq \eta\big) = \sum_{x:|x|\geq\eta} p_X(x) \overset{\frac{|x|}{\eta}\geq 1}{\leq} \sum_{x:|x|\geq\eta} \frac{|x|}{\eta} p_X(x) \leq \frac{1}{\eta} \sum_x |x| p_X(x) = \frac{\mathbb{E}|X|}{\eta}.$$

(ii)

$$\mathbb{P}\big(|X-a| \geq \eta\big) = \mathbb{P}\big(|X-a|^2 \geq \eta^2\big) \overset{(i)}{\leq} \frac{1}{\eta^2}\mathbb{E}\big(|X-a|^2\big).$$

□

3.2 Multidimensional distributions

Suppose that we have two random variables, for instance X = size of a fish, Y = age of that fish. Can we conclude from values of X to values of Y?

Definition 3.16. *Let $X_1, \ldots, X_n$ be random variables on a joint discrete probability space and*

$$F(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}\Big(X_1^{-1}((-\infty, x_1]) \cap \cdots \cap X_n^{-1}((-\infty, x_n])\Big)$$

the joint distribution function of $X_1, \ldots, X_n$. The function $p : \mathbb{R}^n \to [0,1]$, given by

$$p(x_1,\dots,x_n) := \mathbb{P}(X_1 = x_1,\dots,X_n = x_n),$$

is called joint probability mass function. For $1 \le i_1 < i_2 < \cdots < i_k \le n$, the joint distribution of $(X_{i_1},\dots,X_{i_k})$ is called a k-dimensional marginal distribution of $X_1,\dots,X_n$.

Remark 3.17. *For $I \subset \{1,\dots,n\}$, we have*

$$p_{(X_i:\, i\in I)}\big((x_i)_{i\in I}\big) = \mathbb{P}(X_i = x_i \text{ for } i \in I) = \sum_{j\in I^c} \sum_{x_j \in \Omega^{X_j}} p(X_1,\dots,x_n).$$

Example 3.18 (Thrice flipping a coin)**.** *Let 0 standing for "Head" and 1 for "Number", $\Omega = \{0,1\}^3$, $\mathbb{P}$ = Laplace distribution on $(\Omega, \mathcal{P}(\Omega))$. Define the random variables*

$X : \Omega \to \mathbb{R}$, $\omega = (\omega_1,\omega_2,\omega_3) \mapsto \omega_1$ *with* $\Omega^X = \{0,1\}$ *and*

$Y : \Omega \to \mathbb{R}$, $\omega = (\omega_1,\omega_2,\omega_3) \mapsto \omega_1+\omega_2+\omega_3$ *with* $\Omega^Y = \{0,1,2,3\}$.

The joint distribution $\mathbb{P}^{(X,Y)}$ of X, Y is a probability measure on $\Big(\Omega^X\times\Omega^Y, \mathcal{P}\big(\Omega^X\times\Omega^Y\big)\Big)$ with $\mathbb{P}^{(X,Y)}(C) = \mathbb{P}\big((X,Y)^{-1}(C)\big)$ for $C \in \mathcal{P}\big(\Omega^X\times\Omega^Y\big)$, where

$$(X,Y) : \Omega \to \Omega^X \times \Omega^Y, \quad \omega \mapsto \big(\omega_1, \omega_1+\omega_2+\omega_3\big).$$

For instance, if $C = \{(0,1),(1,2)\}$, we find

$$(X,Y)^{-1}(C) = \Big\{(0,1,0),(0,0,1),(1,1,0),(1,0,1)\Big\}$$

and

$$\mathbb{P}^{(X,Y)}(C) = \frac{|(X,Y)^{-1}(C)|}{|\Omega|} = \frac{4}{8} = \frac{1}{2}.$$

The joint probability mass function takes the values $p_{X,Y}(0,0) = \frac{1}{8}$, $p_{X,Y}(0,1) = \frac{2}{8}$, $p_{X,Y}(0,2) = \frac{1}{8}$, $p_{X,Y}(0,3) = 0$, $p_{X,Y}(1,0) = 0$, $p_{X,Y}(1,1) = \frac{1}{8}$, $p_{X,Y}(1,2) = \frac{2}{8}$, $p_{X,Y}(1,3) = \frac{1}{8}$. The marginal probability mass function p_X can be extracted from the joint probability mass function $p_{X,Y}$ by summing over all values of Y:

$$p_X(x) = \mathbb{P}(X=x) = \mathbb{P}(X=x, Y\in\Omega^Y) = \sum_{y\in\Omega^Y} \mathbb{P}(X=x,Y=y) = \sum_{y\in\Omega^Y} p_{X,Y}(x,y).$$

Definition 3.19. *Let $X_1,\dots,X_n$ be random variables on a joint discrete probability space. $X_1,\dots,X_n$ are called stochastically independent, if*

$$\underbrace{\mathbb{P}(X_1 = x_1,\dots,X_n = x_n)}_{p_{X_1,\dots,X_n}(x_1,\dots,x_n)} = \prod_{i=1}^{n} \underbrace{\mathbb{P}(X_i = x_i)}_{p_{X_i}(x_i)} \quad \forall\, x_1,\dots,x_n \in \mathbb{R}.$$

Remark 3.20. *(i) Equivalent is*

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

(ii) For stochastically independent random variables $X_1,\dots,X_n$, we have more general

$$\mathbb{P}(X_1 \in B_1,\dots,X_n \in B_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i) \quad \forall \underbrace{\textit{measurable}}_{\to later!} \; B_i \subset \mathbb{R}$$

In case of discrete random variables as treated up to now, it is sufficient consider $B_i \in \mathcal{P}(\Omega^{X_i})$, $i = 1,\dots,n$.

In the situation of Definition 3.19, one says that the joint density $p_{X_1,\dots,X_n}$ "factorizes".

Example 3.21 (Convolution)**.** *Let $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Bin}(m,p)$ be stochastically independent. $X+Y \sim ?$*

$$\begin{aligned}
\mathbb{P}(X+Y=k) &= \sum_{l=0}^{k} \mathbb{P}(X+Y=k, X=l) \\
&= \sum_{l=0}^{k} \mathbb{P}(Y=k-l, X=l) \\
&= \sum_{l=0}^{k} \binom{m}{k-l} p^{k-l}(1-p)^{m-(k-l)} \binom{n}{l} p^{l}(1-p)^{n-l} \\
&= \sum_{l=0}^{k} \binom{n}{l}\binom{m}{k-l} p^{k}(1-p)^{n+m-k} \\
&= \binom{m+n}{k} p^{k}(1-p)^{n+m-k},
\end{aligned}$$

where the last equality is Exercise 2 b), Homework 3. That is, $X+Y \sim \mathrm{Bin}(n+m,p)$. By induction, one can show in this way:

$$X_1,\dots,X_n \overset{iid}{\sim} \mathrm{Bin}(1,p) \implies \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n,p).$$

["iid" stands for independent identically distributed.]

Theorem 3.22. *Let X and Y be independent random variables on a joint discrete probability space with $\mathbb{E}(X^2) < \infty$, $\mathbb{E}(Y^2) < \infty$. Then*

(i) $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$ *and*

(ii) $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof. (i) With $U = XY$, we find

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_u u p_U(u) \\
&= \sum_u u \sum_{\substack{x\in\Omega^X, y\in\Omega^Y: \\ xy=u}} \mathbb{P}(X=x, Y=y) \\
&= \sum_u \sum_{\substack{x\in\Omega^X, y\in\Omega^Y: \\ xy=u}} xy\mathbb{P}(X=x)\mathbb{P}(Y=y) \\
&= \sum_{x\in\Omega^X, y\in\Omega^Y} xy p_X(x) p_Y(y) \\
&= \Big(\sum_{x\in\Omega^X} x p_X(x)\Big) \cdot \Big(\sum_{y\in\Omega^Y} y p_Y(y)\Big) = (\mathbb{E}X)(\mathbb{E}Y).
\end{aligned}$$

(ii) By linearity of the expectation and (i), we have

$$\mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) = 0.$$

Consequently,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}\Big(\big(X+Y-\mathbb{E}(X+Y)\big)^2\Big) \\
&= \mathbb{E}\Big(\big((X-\mathbb{E}X)+(Y-\mathbb{E}Y)\big)^2\Big) \\
&= \mathbb{E}\Big((X-\mathbb{E}X)^2 + 2(X-\mathbb{E}X)(Y-\mathbb{E}Y) + (Y-\mathbb{E}Y)^2\Big) \\
&= \mathbb{E}\big((X-\mathbb{E}X)^2\big) + 2\underbrace{\mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big)}_{=0} + \mathbb{E}\big((Y-\mathbb{E}Y)^2\big) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y).
\end{aligned}$$

□

Theorem 3.23 (Weak law of large numbers)**.** *Let $X_1, \dots, X_n$ be iid (discrete) random variables on a joint probability space with finite variance $\mathrm{Var}(X_1) = \sigma^2$. Then*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i)\Big| > \varepsilon\Big) \underset{n\to\infty}{\longrightarrow} 0 \quad \textit{for every } \varepsilon > 0.$$

Proof. By the Chebyshev inequality,

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i)\Big| > \varepsilon\Big) \le \frac{1}{\varepsilon^2}\,\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \overset{3.22}{=} \frac{1}{n^2\varepsilon^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

□

Remark 3.24. *The type of convergence of the random variables* $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$ *towards* $\mathbb{E}X_1$*, i.e.*

$$\mathbb{P}(|Y_n - \mathbb{E}X_1| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0 \quad \text{for every } \varepsilon > 0,$$

is called convergence in probability or stochastic convergence. There also exists a so-called strong law of large numbers, whose notion of convergence is stronger than in Theorem 3.23. Grant the conditions of the latter, our proof of the weak law of large numbers actually reveals convergence in quadratic mean $\mathbb{E}\big((Y_n - \mathbb{E}X_1)^2\big) \to 0$*, which by Chebyshev's inequality implies stochastic convergence.*

Definition 3.25. *Let* X *and* Y *be random variables on a joint discrete probability space with* $\mathbb{E}(X^2) < \infty$*,* $\mathbb{E}(Y^2) < \infty$*. Then*

$$\begin{aligned}\mathrm{Cov}(X,Y) &:= \mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big)\\ &= \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)\end{aligned}$$

is called covariance of X *and* Y*. If* $\mathrm{Cov}(X,Y) = 0$*, the random variables* X *and* Y *are called uncorrelated.*

Theorem 3.26. *Let* $X_1, \ldots, X_n$ *be random variables on a joint discrete probability space with* $\mathbb{E}(X_i^2) < \infty$*,* $i = 1, \ldots, n$*. Then*

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i,j:i<j} \mathrm{Cov}(X_i, X_j).$$

If $X_1, \ldots, X_n$ are independent, then $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i)$ by Theorem 3.22. That is, independence implies uncorrelatedness (provided $\mathbb{E}(X_i^2) < \infty$). The reverse direction is not true in general.

Proof. Immediate consequence of the linearity of the expectation (Theorem 3.10). □

Definition 3.27. *Let* X *and* Y *be random variables on some joint discrete probability space with* $\mathbb{E}(X^2) < \infty$*,* $\mathbb{E}(Y^2) < \infty$*,* $\mathrm{Var}(X) > 0$*,* $\mathrm{Var}(Y) > 0$*. The quantity*

$$\rho(X,Y) := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

is called correlation coefficient of X *and* Y*.*

The correlation coefficient is a measure for the linear relation between two random variables X and Y in the following sense.

Theorem 3.28. *Let* X *and* Y *be random variables on some joint discrete probability space with* $\mathbb{E}(X^2) < \infty$*,* $\mathbb{E}(Y^2) < \infty$*,* $\mathrm{Var}(X) > 0$*,* $\mathrm{Var}(Y) > 0$*. Then* $|\rho(X,Y)| \leq 1$ *and*

$\rho(X,Y) = \pm 1$ *if and only if there ecists* $a, b \in \mathbb{R}$ *such that* $\mathbb{P}(Y = aX + b) = 1$. *Here,* $b > 0$ *if* $\rho(X,Y) = 1$ *and* $b < 0$ *if* $\rho(X,Y) = -1$.

Proof. The statement $|\rho(X,Y)| \leq 1$ is a consequence of the Cauchy-Schwarz inequality. For $\rho(X,Y) = 1$, we find

$$\begin{aligned}
&\operatorname{Var}\left(\frac{X}{\sqrt{\operatorname{Var}(X)}} - \frac{Y}{\sqrt{\operatorname{Var}(Y)}}\right) \\
&\quad = \operatorname{Var}\left(\frac{X}{\sqrt{X}}\right) + \operatorname{Var}\left(\frac{Y}{\sqrt{\operatorname{Var}(Y)}}\right) - 2\operatorname{Cov}\left(\frac{X}{\sqrt{\operatorname{Var}(X)}}, \frac{Y}{\sqrt{\operatorname{Var}(Y)}}\right) \\
&\quad = 1 + 1 - 2 = 0.
\end{aligned}$$

This is equivalent to

$$\mathbb{P}\left(\frac{X}{\sqrt{\operatorname{Var}(X)}} - \frac{Y}{\sqrt{\operatorname{Var}(Y)}} = c\right) = 1$$

for some $c \in \mathbb{R}$, whence $\mathbb{P}(Y = a + bX) = 1$ with

$$b = \frac{\sqrt{\operatorname{Var}(X)}}{\sqrt{\operatorname{Var}(Y)}} > 0.$$

The claim for $\rho(X,Y) = -1$ follows analogously with $b = -\frac{\sqrt{\operatorname{Var}(X)}}{\sqrt{\operatorname{Var}(Y)}} < 0$. □

Definition 3.29. *Let* $X_1, \ldots, X_n$ *be random variables on a joint discrete probability space with* $\mathbb{E}(X_i^2) < \infty$, $i = 1, \ldots, n$. *Then the matrix*

$$\Sigma = \Big(\operatorname{Cov}(X_i, X_j)\Big)_{i,j=1,\ldots,n}$$

is called <u>covariance matrix</u> *of* $X = (X_1, \ldots, X_n)'$ *(* $'$ *stands for transposition).*

For any $\gamma = (\gamma_1, \ldots, \gamma_n)' \in \mathbb{R}^n$, bilinearity of Cov implies

$$\operatorname{Var}\left(\gamma' X\right) = \gamma' \Sigma \gamma.$$

The matrix Σ is symmetric and positive semidefinite (since $\operatorname{Var}\left(\gamma' X\right) \geq 0$ for all $\gamma \in \mathbb{R}^n$).

3.3 Conditional distribution and conditional expectation

Lemma and Definition 3.30. *Let* X *and* Y *be random variables on some discrete probability space* $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$. *Let* $x \in \mathbb{R}$ *such that* $\mathbb{P}(X = x) > 0$. *The probability measure* $A \mapsto \mathbb{P}(Y \in A | X = x)$ *on* $(\Omega^Y, \mathcal{P}(\Omega^Y))$ *is the* <u>conditional distribution of Y given $X = x$</u>.

The corresponding conditional probability mass function of Y given $X = x$ is given by

$$p_{Y|X=x}(y) := \mathbb{P}(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

Proof. Provided that $\mathbb{P}(X = x) > 0$, we have $p_{Y|X=x}(y) \geq 0$ and

$$\sum_{y\in\Omega^Y} p_{Y|X=x}(y) = \sum_{y\in\Omega^Y} \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{1}{p_X(x)} \sum_{y\in\Omega^Y} p_{x,y}(x, y) = \frac{p_X(x)}{p_X(x)} = 1.$$

Hence, $p_{Y|X=x} : \Omega^Y \to [0, 1]$ is a probability mass function. Its associated probability measure is the conditional distribution of Y given $X = x$. □

Remark. *One also writes $\mathbb{P}^{Y|X=x}(A)$ for $\mathbb{P}(Y \in A|X = x)$.*

Example 3.31. *Let $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ be independent. Then we find for the conditional probability mass function of X given $X + Y = k$:*

$$\begin{aligned}
p_{X|X+Y=k}(j) &= \mathbb{P}(X = j|X + Y = k) \\
&= \frac{\mathbb{P}(X = j, X + Y = k)}{\mathbb{P}(X + Y = k)} \\
&= \frac{\mathbb{P}(X = j, Y = k - j)}{\mathbb{P}(X + Y = k)} \\
&\overset{\substack{X,Y \\ \text{indep.}}}{=} \frac{\mathbb{P}(X = j)\mathbb{P}(Y = k - j)}{\mathbb{P}(X + Y = k)} \\
&\overset{3.21}{=} \frac{\binom{n}{j}p^j(1-p)^{n-j}\binom{m}{k-j}p^{k-j}(1-p)^{m-(k-j)}}{\binom{m+n}{k}p^k(1-p)^{n+m-k}} = \frac{\binom{n}{j}\binom{m}{k-j}}{\binom{n+m}{k}}.
\end{aligned}$$

That is, the conditional distribution of X given $X + Y = k$ is hypergeometric (Homework 1, Exercise 4 c)) with parameters $n + m$, n and k.

Definition 3.32. *Let X and Y be random variables on some discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$. Let $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) > 0$.*

(i) If $\sum_{y\in\Omega^Y} |y|p_{Y|X=x}(y) < \infty$, the conditional expectation of Y given $X = x$ is defined as

$$\mathbb{E}(Y|X = x) := \sum_{y\in\Omega^Y} y p_{Y|X=x}(y).$$

(ii) If $\sum_{y\in\Omega^Y} y^2 p_{Y|X=x}(y) < \infty$,

$$\mathrm{Var}(Y|X = x) := \mathbb{E}\Big(\big(Y - \mathbb{E}(Y|X = x)\big)^2 \,\Big|\, X = x\Big)$$

is called the conditional variance of Y given $X = x$.

In other words, conditional expectation and variance of Y given $X = x$ are expectation and variance corresponding to the conditional distribution of Y given $X = x$. Therefore, all theorems and identities which have been shown so far for expectation and variance transfer to the respective conditional versions.

Definition 3.33. *Let X and Y be random variables on some discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$, $\sum_{y \in \Omega^Y} |y| p_{Y|X=x}(y) < \infty$ for all $x \in \Omega^X$ with $p_X(x) > 0$. Let*

$$g(x) := \begin{cases} \mathbb{E}(Y|X = x) & \text{if } p_X(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then the random variable $g \circ X = g(X)$ is called conditional expectation of Y given X and is denoted by $\mathbb{E}(Y|X)$. Correspondingly, the conditional variance of Y given X is defined as composition $h \circ X = h(X)$, where

$$h(x) := \begin{cases} \mathrm{Var}(Y|X = x) & \text{if } p_X(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

With this definition, conditional expectation $\mathbb{E}(Y|X)$ and conditional variance $\mathrm{Var}(Y|X)$ are random variables themselves!

Theorem 3.34 (Properties of $\mathbb{E}(Y|X)$)**.** *Let X, Y, Z be random variables on some discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$, $\mathbb{E}|X| < \infty$, $\mathbb{E}|Y| < \infty$, $\mathbb{E}|Z| < \infty$. Then*

(i) $\mathbb{E}\big(\mathbb{E}(Y|X)\big) = \mathbb{E}Y$

(ii) *X, Y stochastically independent* $\Longrightarrow \mathbb{E}(Y|X) = \mathbb{E}Y$ *with probability* 1

(iii) Linearity: For $a, b \in \mathbb{R}$, $\mathbb{E}(aY + bZ|X) = a\mathbb{E}(Y|X) + b\mathbb{E}(Z|X)$ with probability 1.

[If a statement holds true with $\mathbb{P}$-probability 1*, one also says it holds "$\mathbb{P}$-almost surely".]*

Proof. (i)

$$\begin{aligned} \mathbb{E}\big(\mathbb{E}(Y|X)\big) &= \sum_{\substack{x \in \Omega^X: \\ p_X(x) > 0}} \mathbb{E}(Y|X = x) p_X(x) \\ &= \sum_{\substack{x \in \Omega^X: \\ p_X(x) > 0}} \sum_{y \in \Omega^Y} y \underbrace{p_{Y|X=x} \cdot p_X(x)}_{= p_{X,Y}(x,y)} \\ &= \sum_{y \in \Omega^Y} y \underbrace{\sum_{x \Omega^X} p_{X,Y}(x, y)}_{= p_Y(y)} \\ &= \mathbb{E}Y. \end{aligned}$$

(ii) For all $x \in \Omega^X$ with $p_X(x) > 0$, we find

$$\mathbb{E}(Y|X=x) = \sum_{y\in\Omega^Y} y p_{Y|X=x}(y) = \sum_{y\in\Omega^Y} y \frac{p_{X,Y}(x,y}{p_X(x)} = \sum_{y\in\Omega^Y} y \frac{p_X(x)p_Y(y)}{p_X(x)} = \mathbb{E}(Y),$$

where we have used that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ by independence.

(iii) Homework 5, Exercise 2. □

Theorem 3.35. *Let X and Y be stochastically independent random variables on some discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$, $f : \mathbb{R}^2 \to \mathbb{R}$ some function with $\mathbb{E}|f(X,Y)| < \infty$. Then*

$$\mathbb{E}\Big(f(X,Y)\,\Big|\,X = x_0\Big) = \mathbb{E}\big(f(x_0,Y)\big) \tag{3.3}$$

for all x_0 with $\mathbb{P}(X = x_0) > 0$.

Proof. For the random variable $Z = f(X,Y)$, we find

$$\begin{aligned}
\mathbb{E}(Z|X=x_0) = \sum_{z\in\Omega^Z} z p_{Z|X=x_0)}(z) &= \sum_{z\in\Omega^Z} z \frac{\mathbb{P}(Z=z, X=x_0)}{\mathbb{P}(X=x_0)} \\
&= \sum_{z\in\Omega^Z} z \frac{\mathbb{P}(f(x_0,Y)=z, X=x_0)}{\mathbb{P}(X=x_0)} \\
&= \sum_{z\in\Omega^Z} z \sum_{y:f(x_0,y)=z} \frac{\mathbb{P}(Y=y, X=x_0)}{\mathbb{P}(X=x_0)} \\
&\overset{\substack{X,Y\\ \text{indep.}}}{=} \sum_{z\in\Omega^Z} \sum_{y:f(x_0,y)=z} f(x_0,y) \frac{\mathbb{P}(Y=y)\mathbb{P}(X=x_0)}{\mathbb{P}(X=x_0)} \\
&= \sum_{y\in\Omega^Y} f(x_0,y) p_Y(y) = \mathbb{E}\big(f(x_0,Y)\big).
\end{aligned}$$

□

Remark 3.36. *If f in Theorem 3.35 is of the form $f(x,y) = y \cdot h(x)$, then (3.35) reveals*

$$\mathbb{E}\Big(Y h(X)\,\Big|\,X\Big) = h(X) \cdot \mathbb{E}(Y|X) \quad \mathbb{P}\text{-a.s.}$$

Moreover, $\mathbb{E}(Y|X) = \mathbb{E}Y$ $\mathbb{P}$-a.s. if X and Y are stochastically independent.

Applying the above considerations, we deduce a discrete version of Fubini's theorem.

Theorem 3.37 (Theorem of Fubini, discrete version)**.** *Let X and Y be stochastically independent random variables on some discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$, $f : \mathbb{R}^2 \to \mathbb{R}$ some function with $\mathbb{E}|f(X,Y)| < \infty$. Let $g(x) := \mathbb{E}\big(f(x,Y)\big)$ and $h(y) := \mathbb{E}\big(f(X,y)\big)$. Then*

$$\mathbb{E}\big(f(X,Y)\big) = \mathbb{E}g(X) = \mathbb{E}h(Y).$$

Proof. The identity is an immediate consequence of Theorem 3.34 (i) and Theorem 3.35. Alternatively, we may verify directly

$$\mathbb{E}\big(f(X,Y)\big) = \sum_{x,y} f(x,y)p_{X,Y}(x,y)$$
$$\overset{\substack{X,Y\\ \text{indep.}}}{=} \sum_x \Big(\sum_y f(x,y)p_Y(y)\Big)p_X(x) = \mathbb{E}\Big(\mathbb{E}\big(f(X,Y)\,\big|\,X\big)\Big).$$

□

If X is a discrete random variable with $\mathbb{E}\big(X^2\big) < \infty$, then

$$\mathbb{E}\big((X-a)^2\big) \geq \mathrm{Var}(X)$$

for every $a \in \mathbb{R}$ by Theorem 3.13. That is, $\mathbb{E}X$ minimizes the deviation in quadratic mean. In this sense, one talks of $\mathbb{E}X$ as the "best constant predictor" for the random variable X. A similar optimality property is true for the conditional expectation.

Theorem 3.38. *Let X and Y be random variables on some discrete probability space $(\Omega, \mathcal{P}(\Omega), \mathcal{P})$, $\mathbb{E}\big(X^2\big) < \infty$ and $\phi : \mathbb{R} \to \mathbb{R}$ some function with $\mathbb{E}\big(\phi(X)^2\big) < \infty$. Then*

$$\mathbb{E}\Big(\big(Y-\phi(X)\big)^2\Big) \geq \mathbb{E}\Big(\big(Y-\mathbb{E}(Y|X)\big)^2\Big)$$

with equality if and only if $\phi(X) = \mathbb{E}(Y|X)$ almost surely.

Proof. Define

$$g(x) := \begin{cases} \mathbb{E}\big((Y-\phi(X))^2\,\big|\,X=x\big) & \text{if } p_X(x) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}\big((Y-\phi(X))^2\big) = \mathbb{E}g(X)$ by Theorem 3.34 (i). We prove subsequently:

$$\mathbb{E}\Big(\big(Y-\phi(X)\big)^2\,\Big|\,X=x\Big) = \mathbb{E}\Big(\big(Y-\phi(x)\big)^2\,\Big|\,X=x\Big). \tag{3.4}$$

Denoting $f(x,y) := (y-\phi(x))^2$ and $Z := f(X,Y) = (Y-\phi(X))^2$, we find as in the proof of Theorem 3.35 for x_0 with $p_X(x_0) > 0$:

$$\mathbb{E}(Z|X=x_0) = \sum_{y\in\Omega^Y} f(x_0,y)p_{Y|X=x_0}(y) = \mathbb{E}\big(f(x_0,Y)\,\big|\,X=x_0\big),$$

i.e. (3.4). Correspondingly,

$$\mathbb{E}\Big(\big(Y-\mathbb{E}(Y|X)\big)^2\,\Big|\,X=x\Big) = h(x)$$

with

$$h(x) = \mathbb{E}\Big(\big(Y - \mathbb{E}(Y|X=x)\big)^2 \,\Big|\, X = x \Big).$$

In order to prove the claim of the theorem, it is sufficient by Homework 3, Exercise 1 b), to show that $g(x) \geq h(x)$ for all x with $p_X(x) > 0$ with equality if and only if $\phi(x) = \mathbb{E}(Y|X=x)$. But this follows from Theorem 3.13. □

4 The need of continuous random variables and some examples

Example 4.1. *Student S is waiting for his girlfriend. As she planned to arrive at four o'clock in the afternoon, he considers all time points between four and five as equally probable arrival times.*

First idea: (Approximation) Decompose the interval $[a,b] = [4,5]$ into n subintervals of equal length with mitpoints $x_1^n, \dots, x_n^n$ and assign to each of them the probability $1/n$. If the arrival time X could only take values in $\{x_1^n, \dots, x_n^n\}$, then

$$\mathbb{P}(c \leq X \leq d) = \sum_{c \leq x_i^n \leq d} \frac{1}{n} = \frac{b-a}{n} \sum_{c \leq x_i^n \leq d} \frac{1}{b-a} \underset{n\to\infty}{\longrightarrow} \int_c^d \frac{1}{b-a} dx.$$

Definition 4.2. *Let $f : \mathbb{R} \to [0,\infty)$ be integrable with $\int f(x)dx = 1$. Then the assignment*

$$\mathbb{P}([a,b]) = \int_a^b f(x)dx \quad \text{for all intervals } [a,b] \subset \mathbb{R}$$

defines a probability distribution on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (where $\mathcal{B}(\mathbb{R})$ is the Borel-σ-field $\to$ measure theory!). $\mathbb{P}$ is called continuous distribution and f its probability density.

The fact that $\mathbb{P}$ is uniquely defined and a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is proven in the course on measure theory.

Examples 4.3.

(i) Uniform distribution $\mathcal{U}[a,b]$ on some interval $[a,b]$ with $f(x) = \frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$.

(ii) Exponential distribution $\mathcal{E}(\lambda)$ with parameter λ and probability density

$$f(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,\infty)}(x).$$

(iii) Normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters μ, σ^2 and probability density

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