Winter 2019/20
By Timo Freiesleben

Script for Central Topics in the Philosophy of Science

This script contains the definitions and theorems that one needs for solving the problem sets of CTPS. The level of detail I gave is aimed to match the level we need for the philosophical applications we have in mind. Hence, it will be simpler compared to most of standard mathematics introductions to probability theory or Bayesian Networks. Nevertheless, I tried to ground the definitions solidly such that they are special cases of more complex definitions. If you are interested in reading on the more advanced or more general probability theory, I tried to point to the concept names you can google for.
All information is subject to change. There will be typos, misscalculations, and maybe even missconceptions in the script. Therefore, it is better if you ultimately rely on the lecture slides. If you find any mistakes, please write an email to timo.freiesleben@web.de.
Notice that this script does not give an intro to the philosophical parts of the CTPS-course.

1 Proof by Mathematical Induction

Mathematical induction is one of the standard methods of how to prove statements of the following form:

$$\forall n \in \mathbb{N}_0 \text{ with } n \geq k \text{ holds } A(n) \tag{1}$$

Where k is some element in $\mathbb{N}$ and A is a 1-ary predicate.
There are different ways of how this can be done, which are all equivalent. Two very common versions look like this:

$$A(k) \wedge \forall n \geq k : (A(n) \rightarrow A(n+1))$$

or

$$A(k) \wedge \forall n \geq k : ((\forall k \leq l \leq n : A(l)) \rightarrow A(n+1)).$$

The simple idea standing behind mathematical induction is the following.

1. Assume I can prove some statement for a particular natural number k.
2. Assume moreover I can prove that whenever our statement holds for some natural number n greater than k then it also holds for n+1.

Then, I can conclude that it holds for all natural numbers n greater than k. Why is that? Easy! I know the statement is true for k since (1). But, if it holds for k it also has to hold for k+1 due to (2). Now, I know it holds for k+1, but then it also has to hold for (k+1)+1, and so on and so forth. At some point, we will reach every natural number greater than k by this procedure.

Example 1.1. Prove that for all $n \in \mathbb{N}_0$ the following holds $\sum\limits_{i=1}^{n} i = \frac{n(n+1)}{2}$ by mathematical induction.

Proof. Let $A(l) :\equiv \sum\limits_{i=1}^{l} i = \frac{l(l+1)}{2}$.
Base Case:
Since the statement should hold for all $n \in \mathbb{N}_0$ our base case is to show that $A(0)$. $A(0) \Leftrightarrow \sum\limits_{i=1}^{0} i = \frac{0(0+1)}{2} \Leftrightarrow 0 = 0$
Induction Hypotheses:
for a fixed but arbitrary $n \in \mathbb{N}_0$ holds A(n).
Inductive Step:
$\sum\limits_{i=1}^{n+1} i = (n+1) + \sum\limits_{i=1}^{n} i \overset{I.H}{=} (n+1) + \frac{n(n+1)}{2} = 2\frac{(n+1)}{2} + \frac{n(n+1)}{2} = \frac{n(n+1)+2(n+1)}{2} = \frac{(n+1)(n+2)}{2}$

□

2 Basic Definitions and Analysis

The following Definitions one will need regularly:

Definition 2.1. Summation

$$\sum_{i=1}^{n} a_i = a_1 + \cdots + a_n$$

where $n \in \mathbb{N}_0$ and $a_1, \ldots, a_n \in \mathbb{R}$. For $n = 0$ $\sum\limits_{i=1}^{n} a_i = \sum\limits_{i=1}^{0} a_i = 0$..

Definition 2.2. Products

$$\prod_{i=1}^{n} a_i = a_1 \cdots a_n$$

where $n \in \mathbb{N}_0$ and $a_1, \ldots, a_n \in \mathbb{R}$. For $n = 0$ $\prod\limits_{i=1}^{n} a_i = \prod\limits_{i=1}^{0} a_i = 1$..

Definition 2.3. Factorial

$$n! = \prod_{i=1}^{n} i$$

where $n \in \mathbb{N}_0$.

Example 2.4. Simple Examples:

- $\sum\limits_{i=1}^{5} i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2$
- $\prod\limits_{i=1}^{5} i = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 5!$

Comment 2.5. *Index Trick*
Let $a_1, \ldots, a_n \in \mathbb{R}$ with $n \in \mathbb{N}$ and $k \in \mathbb{N}$, then

$$\sum_{i=1}^{n} a_i = \sum_{i=1+k}^{n+k} a_{i-k}$$

This trick is used quite regularly in many contexts, especially in proofs by induction. Clearly, the same works with products.

For this course, one only needs a very basic knowledge of analysis. The following I consider relevant. Check for the proofs of the theorems online if you are interested. I'll always put the names of the theorems such that it is easy to find proofs.

Theorem 2.6. *(Chain Rule)*
Assume u and v are differentiable functions with $Image(v) \subseteq Domain(u)$ with $f(x) := u(v(x))$ for all $x \in Domain(v)$. If f is differentiable in point $x_0 \in Domain(v)$ (You usually don't have to care about differentiability that much) then:

$$f'(x_0) = (u \circ v)'(x_0) = u'(v(x_0))v'(x_0)$$

Or in short

$$f' = (u \circ v)' = (u' \circ v) \cdot v'$$

where f', u', v' denote the derivative of the respective functions.

Theorem 2.7. *(Product Rule)*
Assume u and v are functions from the domain $D \subseteq \mathbb{R}$ with $f(x) := u(x) \cdot v(x)$ for all $x \in D$. Assume f is differentiable in point $x_0 \in D$ then:

$$f'(x_0) = (uv)'(x_0) = u'(x_0)v(x_0) + u(x_0)v'(x_0)$$

Or in short

$$f' = (uv)' = u'v + uv'$$

where f', u', v' denote the derivative of the respective functions.

Theorem 2.8. *(Quotient Rule)*
Assume u and v are functions from the domain $D \subseteq \mathbb{R}$ with $f(x) := \frac{u(x)}{v(x)}$ for all $x \in D$ with $v(x) \neq 0$. Assume f is differentiable in point $x_0 \in D$ then:

$$f'(x_0) = (uv)'(x_0) = \frac{u'(x_0)v(x_0) - u(x_0)v'(x_0)}{v(x)^2}$$

Or in short

$$f' = (uv)' = \frac{u'v - uv'}{v^2}$$

where f', u', v' denote the derivative of the respective functions.

Theorem 2.9. *Let $f(x) := ln(g(x))$ where g is a differentiable function. Then,*

$$f'(x) = \frac{g'(x)}{g(x)}.$$

Proof. We make use of the chain rule. We define $u(z) := ln(z)$ and $v(x) = g(x)$. Then, $u'(z) = \frac{1}{z}$ and $v'(x) = g'(x)$. Thus, by the chain rule we get that $f'(x) = u'(v(x))v'(x) = \frac{1}{v(x)}v'(x) = \frac{g'(x)}{g(x)}$. □

Theorem 2.10. *Let $f(x) := e^{g(x)}$ where g is a differentiable function. Then,*

$$f'(x) = g'(x) \cdot e^{g(x)}.$$

Proof. We make use of the chain rule. We define $u(z) := e^z$ and $v(x) = g(x)$. Then, $u'(z) = e^z$ and $v'(x) = g'(x)$. Thus, by the chain rule we get that $f'(x) = u'(v(x))v'(x) = e^{v(x)}v'(x) = g'(x)e^{g(x)}$ □

Example 2.11. Let $f(x) := x \cdot \log(x)$. We make use of the product rule. We define $u(x) := x$ and $v(x) := \log(x)$. We obtain that $u'(x) = 1$ and $v'(x) := \frac{1}{x}$. Thus $f'(x) = u'(x)v(x) + u(x)v'(x) = log(x) + 1$.

Theorem 2.12. *(Laws for Logarithm)*
Let $a, x, y \in \mathbb{R}^+, \beta, z \in \mathbb{R}$ then:

- $log_a(x \cdot y) = log_a(x) + log_a(y)$
- $log_a(\frac{x}{y}) = log_a(x) - log_a(y)$
- $\beta\ log_a(x) = log_a(x^\beta)$
- $x = a^z \Rightarrow z = log_a(x)$

Usually, we will have that $a = e$ is the Euler constant with $e \approx 2,7183$.

Theorem 2.13. *(Laws for Exponential)*
Let $x, y, z, \beta \in \mathbb{R}$ then:

- $x^y \cdot x^z = x^{y+z}$
- $x^z \cdot y^z = (x \cdot y)^z$
- $(x^y)^\beta = x^{\beta y}$
- $\frac{1}{x^y} = x^{-y}$

Definition 2.14. Maxima/Minima
Let $f : D \to \mathbb{R};\ x \mapsto f(x)$ be a differentiable function with $D \subseteq \mathbb{R}$. We call $x_0 \in D$

- a local minimum of f if[1] $f'(x_0) = 0 \wedge f''(x_0) > 0$
- a local maximum of f if $f'(x_0) = 0 \wedge f''(x_0) < 0$
- a global minimum of f iff $\forall x \in D :\ f(x_0) \leq f(x)$
- a global maximum of f iff $\forall x \in D :\ f(x_0) \geq f(x)$

[1] Note that $f'(x_0) = 0$ is only a necessary condition and $f''(x_0) > 0$ is only a sufficient condition. However, usually nothing more will be relevant for you. To see that the latter is not necessary consider $f(x) = x^4$ which has a minimum at $x = 0$ however $f''(0) = 0$.

3 Combinatorics

The following are some very basic scenarios in combinatorics.
Assume you have an urn with n distinct elements. There are:

1. $\binom{n}{k} := \frac{n}{k!(n-k)!}$ ways to draw k elements without replacement without order
2. $\frac{n!}{(n-k)!}$ ways to draw k elements without replacement in order.
3. $\binom{n+k-1}{k}$ ways to draw k elements with replacement without order.[2]
4. n^k ways to draw k elements with replacement in order.

4 Basic Probability Theory

Here we introduce the basics of probability theory. If we work with a space of events Ω we will always assume that $|\Omega| < \infty$. This means that Ω is finite.[3]

Definition 4.1. Sample space
A sample space is a set of possible outcomes. In philosophical contexts, this is often a set of possible worlds. This set is usually denoted Ω. We demand that $\Omega \neq \emptyset$.

Definition 4.2. σ-Algebra
For us the σ-Algebra is just $\mathcal{F} := \mathcal{P}(\Omega) = 2^\Omega$ in every case. $\mathcal{P}(\Omega)$ denotes the powerset of Ω, which is the set of all subsets of Ω meaning $\mathcal{P}(\Omega) := \{A \mid A \subseteq \Omega\}$. The elements of $\mathcal{P}(\Omega)$ are usually called events.

Definition 4.3. Probability measure
A Probability measure P is a function

$$P : \mathcal{P}(\Omega) \to [0, \infty],\ A \mapsto P(A)$$

that satisfies the following two conditions:

- $P(\Omega) = 1$ and
- if $A, B \subseteq \Omega$ with $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

Definition 4.4. Probability Space
A probability space is a triple $(\Omega, P, \mathcal{P}(\Omega))$ where Ω is a sample space, P is a probability measure on Ω, and $\mathcal{P}(\Omega)$ the powerset of Ω.

Definition 4.5. Random Variable
A random variable X is a (measureable)[4] function $X : \Omega \to \mathbb{R}$, $\omega \mapsto X(\omega)$. We define for any $x \in \mathbb{R}$ $[X = x] := X^{-1}(x) \subseteq \Omega$[5] or more generally $[X = A] = X^{-1}(A) \subseteq \Omega$ for $A \subseteq \mathbb{R}$. Thus,

$$P(X = x) = P(X^{-1}(x)).$$

Theorem 4.6. *Let $A, B \subseteq \Omega$. The following are interesting properties of probability measures:*

- $P(\emptyset) = 0$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) \leq 1$
- $P(A) = \sum\limits_{\omega \in A} P(\{\omega\})$

[2] Only added for matters of completeness. Google Multiset if you want to know more.

[3] As always things get much more difficult but also interesting if we drop this assumption. However, this would demand a lot more work. If you are interested you should first acquire some basics in measurement theory. The σ-Algebra we are working with will in infinite cases usually not be the powerset of Ω. (Coolest thing in measurement theory related to that: The Banach-Tarski Paradoxon.)

[4] Again, in your case any function is measureable. The requirement is $X^{-1}(A) \in \mathcal{F}$ for all A in the Borel-σ-Algebra of $\mathbb{R}$.

[5] This is only well defined if X is a measureable function, so you can see how the puzzle fits together.

- *If* $\bigcup\limits_{i=1}^{n} A_i = \Omega$ *and* $\forall i, j \in \{1, \dots, n\}$ *with* $i \neq j$ *holds* $A_i \cap A_j = \emptyset$ *then for any set* $B \subseteq \Omega$ *holds* $P(B) = \sum\limits_{i=1}^{n} P(A_i \cap B)$

Definition 4.7. Joint Probability Mass Function and the Distribution of X
Let $(\Omega, P, 2^\Omega)$ be a probability space where $|\Omega| < \infty$, 2^Ω denotes the powerset of Ω, and let $X : \Omega \to \mathbb{R}$ be a random variable. The *distribution* P_X of X is defined by

$$P_X(A') := P(\{\omega \in \Omega : X(\omega) \in A'\}) \text{ for all } A' \subseteq \mathbb{R}$$

The so-called *probability mass function* $f_X : \mathbb{R} \to [0,1]$ of X is used among other things to visualise distributions. It is straight forwardly defined by

$$f_X : x \mapsto P(X = x).$$

For n random variables $X_1, \dots, X_n$ and real numbers $x_1, \dots, x_n \in \mathbb{R}$ with $n \in \mathbb{N}$ the joint probability mass function is defined as:

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) := P(X_1 = x_1 \text{ and } \cdots \text{ and } X_n = x_n)$$

Definition 4.8. Identically Distributed
Let X, Y be random variables. We say that X and Y are distributed identical if $\forall a \in \mathbb{R}$ holds:

$$f_X(a) = f_Y(a)$$

Notice that this does not imply that the random variables are identical. They just assign the same probabilities to particular values, that could be to completely different events. Also, on zero measure sets (events that "almost never" happen) the values could be completely off.

Example 4.9. Some Interesting Distributions
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space and X be a random variable. The following are some common and interesting distributions.

- We call P uniformly distributed on Ω ($P \sim Unif(\Omega)$) iff $P(\omega) = \frac{1}{|\Omega|} \quad \forall \omega \in \Omega$.
- We call P_X (often also X) Bernoulli distributed ($X \sim$ *Ber*(p)) with $p \in [0,1]$ iff P(X=1)=p=1-P(X=0).
- We call P_X (often also X) Binomially distributed ($X \sim$ *Bin*(p,n)) with $p \in [0,1], n \in \mathbb{N}$ iff $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.
- We call P_X (often also X) Geometrically distributed ($X \sim$ *Geo*(p)) with $p \in [0,1]$ iff $P(X = k) = p(1-p)^{k-1}$.

A standard Bernoulli trial with parameter p would be a coin flip with probability p showing heads. A Binomially distributed variable with parameters p and n would be the sum of n many independent Bernoulli trials with parameter p. A geometrically distributed variable with parameter p could be interpreted as assigning to each k the probability that after k Bernoulli trials it turns out heads for the first time.

Definition 4.10. Expected Value
Let X be a random variable as defined above. Then,

$$\mathbb{E}(X) := \sum_{x \in \mathbb{R}} P(X = x) \cdot x.$$

This is only one way to state the expected value. As one can prove the following is equivalent:

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} P(\{\omega\}) \cdot X(\omega).$$

Notice that we will regularly use $P(\omega)$ instead of $P(\{\omega\})$, which is only an abbreviation but does not change the fact that P is only defined for sets of outcomes.[6]
Intuitively the expected value is something like a weighted average of the outcomes.

[6] Also, we might use $\mathbb{E}[X]$ instead of $\mathbb{E}(X)$ from time to time, both notions are very common.

Theorem 4.11. *Linearity of Expected Value*
Let X, Y be random variables and $a, b \in \mathbb{R}$. Important properties of the Expected Value:

- $\mathbb{E}(a \cdot X + b) = a \cdot \mathbb{E}(X) + b$
- $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

Definition 4.12. Variance & Covariance
Let X, Y be random variables. The variance of X is defined as:

$$var(X) := \mathbb{E}((X - \mathbb{E}[X])^2)$$

The variance tells you about how far the outcomes of the random variable are spread from the average value. Moreover, we define the covariance of X, Y as follows:

$$cov(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Covariance is a lot harder to interpret. I would say it mainly shows the linear relationships between two random variables. Notice, that if the Covariance is zero we say that the two random variables are uncorrelated. This does not mean that they are independent. Independence on the other side implies a covariance of zero.

Theorem 4.13. *Properties of Variance & Covariance*

- $var(X) = cov(X, X)$.
- $var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
- $cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

Definition 4.14. Conditional Probabilities
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. For any $A, B \subseteq \Omega$ with $P(B) > 0$ we can define the probability of A given that B as follows:

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}$$

Definition 4.15. Independence
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Let $A, B \subseteq \Omega$. We call A independent of B if and only if

$$P(A \cap B) = P(A) \cdot P(B)$$

We will later see a further generalization of this definition.
One can derive by this that if $P(B) > 0$ then

$$A, B \text{ are independent iff } P(A \mid B) = P(A). \tag{2}$$

This has a lot of intuitive appeal! It says that event A is independent of event B if knowing that B happened does not tell us anything about whether event A happens.

Definition 4.16. Independence of Random Variables
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Let X and Y be two random variables. We call X independent of Y if and only if

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) \;\; \forall x, y \in \mathbb{R}$$

Very often we do not specify the sampling space exactly and instead we start with the random variables and the probability distributions of these random variables. There is a theorem showing that there exists a sample space on which these random variables are well defined.

Theorem 4.17. *Law of Large Numbers*
Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent and identically distributed random variables (i.i.d) with $E[X_1^2] < \infty$ then for $\overline{X}_n := \sum\limits_{i=1}^{n} \frac{X_i}{n}$ holds:

$$\lim_{n\to\infty} \overline{X}_n = E(X_1) \textit{ almost certainly.}$$

Theorem 4.18. *General Product Rule (Also called chain rule of probability)*
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. For any $A_1, \ldots, A_n \subseteq \Omega$ with $\bigcap_{i=1}^{n-1} A_i \neq \emptyset$ with $n \in \mathbb{N}$ holds the following:

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i \mid A_1 \cap \cdots \cap A_{i-1}).$$

Theorem 4.19. *Bayes Theorem*
Let $\bigcup_{i=1}^{n} H_i = H$ for some set H in Ω s.t. $\forall i, j \in \{1, \ldots, n\}$ with $i \neq j$ holds $H_i \cap H_j = \emptyset$. Let moreover $E \subseteq \Omega$. Then, the following holds:

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{\sum\limits_{i=1}^{n} P(H_i)P(E \mid H_i)}$$

Definition 4.20. Conditional Expectation
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Moreover, let X, Y be random variables and $y \in Image(Y)$ with $P(Y = y) > 0$. Then, we can define the conditional expectation of X given $Y = y$ as follows:

$$\mathbb{E}(X \mid Y = y) := \sum_{x \in \mathbb{R}} P(X = x \mid Y = y)x$$

Generally, we can define the function

$$\mathbb{E}(X \mid Y) : A \to \mathbb{R};\ y \mapsto \mathbb{E}(X \mid Y = y)$$

Where $A = \{y \in Image(Y) \mid P(Y = y) > 0\}$. If we define this more carefully[7] we can even get a random variable over all $\mathbb{R}$.
Intuitively the conditional expectation of X on $Y = y$ expresses the expected value of the random variable X given that the random variable Y happened to be y.

Let's see all this notions in action in an example:

Example 4.21.
Probability Space
Let $\Omega := \{w_1, \ldots, w_6\} \times \{w_1, \ldots, w_6\}$ be a probability space, where (w_i, w_j) describes the outcome that die one landed on i and die two landed on j with $i, j \in \{1, \ldots, 6\}$. Assume $P \sim Unif(\Omega)$ (meaning $P(w) := P(\{w\}) = \frac{1}{36}$ for all $w \in \Omega$) and our σ-algebra is $\mathcal{P}(\Omega)$. Then our probability space is $(\Omega, P, \mathcal{P}(\Omega))$.
Measuring an event
Let $A := \{(w_1, w_1), (w_2, w_2), (w_4, w_2), (w_6, w_3)\}$, then

$$P(A) = P(\{(w_1, w_1), (w_2, w_2), (w_4, w_2), (w_6, w_3)\}) = P((w_1, w_1)) + P((w_2, w_2)) + P((w_4, w_2)) + P((w_6, w_3)) = \frac{4}{36} = \frac{1}{9}.$$

Independence of two sets
Let $A := \{(w_1, w_1), (w_1, w_2), (w_1, w_3), (w_1, w_4), (w_1, w_5), (w_1, w_6)\}$ and
$B := \{(w_1, w_4), (w_2, w_4), (w_3, w_4), (w_4, w_4), (w_5, w_4), (w_6, w_4)\}$. Then, $A \cap B = \{(w_1, w_4)\}$. We can show that A and B are independent:

$$P(A \cap B) = P((w_1, w_4)) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(A)P(B).$$

Independence of Random Variables
Now, we have a look at random variables. Let,

$$X_1 : \Omega \to \mathbb{R},\ (\omega_i, \omega_j) \mapsto i \text{ for all } i, j \in \{1, \ldots, 6\}$$

and

$$X_2 : \Omega \to \mathbb{R},\ (\omega_i, \omega_j) \mapsto j \text{ for all } i, j \in \{1, \ldots, 6\}.$$

[7] In maths this is defined not constructively but by a condition on the function. Consult the wiki article of conditional expectation to learn more about it.

Informally, X_1 is a random variable describing the outcome of rolling die one and X_2 is a r.v. describing the outcome of rolling die two. The distribution of X is $P_{X_1}(i) = \begin{cases} \frac{1}{6} \text{ if } i \in \{1, \ldots, 6\} \\ 0 \text{ else} \end{cases}$. Notice that X_1 and X_2 are identically distributed since $f_{X_1}(i) = f_{X_2}(i)$ for all $i \in \mathbb{R}$.
X_1 and X_2 are also independent. One can actually show that, in order for doing so one has to show that for all $i, j \in \mathbb{R}$ holds $P(X_1 = i, X_2 = j) = P(X_1 = i)P(X_2 = j)$. We showed already above that $A := [X_1 = 1]$ and $B := [X_2 = 4]$ are independent. The same can be done for the other values of $i, j \in \mathbb{R}$. For $i \in \mathbb{R} \setminus \{1, \ldots, 6\}$ or $j \in \mathbb{R} \setminus \{1, \ldots, 6\}$ for example holds that $P(X_1 = i) = 0$ or $P(X_2 = j) = 0$ and thus $P(X_1 = i, X_2 = j) = P(\emptyset) = 0 = P(X_1 = i)P(X_2 = j)$.
The joint probability mass function is for all $i, j \in \{1, \ldots, 6\}$:

$$f_{X_1,X_2}(i, j) = f_{X_1}(i) f_{X_2}(j) = \frac{1}{36}$$

Expectation
The expectation can be calculated as follows:

$$\mathbb{E}(X_1) = \sum_{i=1}^{6} P(X_1 = i)\, i = \frac{1}{6}(6 + 5 + 4 + 3 + 2 + 1) = \frac{7}{2}$$

Variance
$\operatorname{Var}(X_1) = \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \sum\limits_{i=1}^{6} i^2 P(X = i) - \frac{7}{2} = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) - \frac{7}{2}^2 \approx 2.92.$
Covariance
$\operatorname{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) \overset{indep}{=} \mathbb{E}(X_1)\mathbb{E}(X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = 0$
Sum of Random Variables
Now, we define $X_3 := X_1 + X_2$. Then X_3 is not independent from X_1 since for instance $0 = P(X_3 = 12 \mid X_1 = 1) \neq P(X_3 = 12) = \frac{1}{36}$.
Conditional Expectation

$$\mathbb{E}(X_3 \mid X_1) : \{1, \ldots, 6\} \to \mathbb{R},\ i \mapsto \mathbb{E}(X_3 \mid X_1 = i) = i + 3.5$$

Law of Large Numbers
Let $(Y_i)_{i \in \mathbb{N}}$ be i.i.d. random variables with $f_{Y_i}(j) = \frac{1}{6}$ for all $i \in \mathbb{N}$ and $j \in \{1, \ldots, 6\}$. Then, $\mathbb{E}(Y_1^2) < \infty$ and $\overline{Y}_n := \sum\limits_{i=1}^{n} \frac{Y_i}{n}$

$$\lim_{n \to \infty} \overline{Y}_n = E(Y_1) = 3.5 \text{ almost certainly.}$$

5 Propositional Variables

Comment 5.1. *Let Ω be a sample space and P be a probability measure. Instead of classical random variables or events, it is often simpler in philosophical applications to work with propositional variables and ignore the underlying structure. However, we can interpret a proposition C in a language $\mathcal{L}$ also as the set of possible worlds in which the proposition is true. Formally a proposition is then nothing but a subset of Ω. We describe a propositional variable as a set $\mathcal{H} := \{H, \neg H\}$, where H is a proposition and $\neg H$ denotes the negation of H. Obviously, $\mathcal{H}$ partitions Ω. If we work with propositional variables we will usually not specify the sample space in full detail but instead, directly look at the probability of the propositions. However, for any propositional variable, we can easily construct a corresponding binary random variable. We can define the corresponding binary random variable to a propositional variable $\mathcal{H}$ as follows:*

$$\mathcal{H}_{binary} : \Omega \to \{0, 1\};\ \omega \mapsto \begin{cases} 1 \ \textit{if } \omega \in H \\ 0 \ \textit{else} \end{cases}$$

Then, $\mathcal{H}_{binary}(\omega) = 1$ if and only if H is true. One could also see their relation exactly by the pre-image, via $\mathcal{H}_{binary}^{-1}(1) = H$ and $\mathcal{H}_{binary}^{-1}(0) = \neg H$.
If we have more than one proposition for example n many with H_i proposition for all $i \in \{1, \ldots, n\}$ then we define $P(H_1, \ldots, H_n) := P(H_1 \wedge \cdots \wedge H_n)$

Comment 5.2. *For almost all concepts we defined before we have the corresponding concept in propositional language. Let A, B be propositions and $\top, \bot$ denote Verum, Falsum respectively.*

- $P(\top) = 1$
- $P(\bot) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
- $P(A) \leq 1$
- *If* $\vdash \bigvee\limits_{i=1}^{n} A_i$ *and* $\forall i, j \in \{1, \ldots, n\}$ *with* $i \neq j$ *holds* $A_i \wedge A_j \vdash \bot$ *then for any proposition* B *holds* $P(B) = \sum\limits_{i=1}^{n} P(A_i \wedge B)$

However, we can not get the concept of expectation as we have it with random variables since it is defined over the values in $\mathbb{R}$, but we can often use the corresponding binary random variable instead.
Assume you have $\mathcal{A}_1, \ldots, \mathcal{A}_n$ different propositional variables. In order to fully specify the joint distribution $P(\mathcal{A}_1, \ldots, \mathcal{A}_n)$ we need $2^n - 1$ different values. We need one value for each of the configurations $\mathcal{A}_1, \ldots, \mathcal{A}_n$ can have except for one which we can derive by it being a complement to the union of all the others.

6 Bayesian Networks and d-separation

Definition 6.1. DAG
A directed acyclic graph consists of vertices(also called nodes) V where V is a finite non-empty set and directed edges(these represent the arrows) E. Let $A, B \in V$ be nodes, then, an directed edge $e \in E$ is given by a tuple $e := (A, B) \in E$, where, e denotes that there is an edge from node A to node B. For the graph being acyclic there must not exist directed edges $(A_1, B_1), \ldots, (A_n, B_n) \in E$ with $n \in \mathbb{N}$ and $A_1, \ldots, A_n, B_1, \ldots, B_n \in V$ such that $\forall i \in \{1, \ldots, n-1\} : B_i = A_{i+1}$ and $A_1 = B_n$.

Definition 6.2. Root Node
Let $\langle V, E \rangle$ be a DAG. Then, we call $\mathit{root}(\langle V, E \rangle) := \{B \in V \mid \nexists A \in V : (A, B) \in E\}$ the root nodes. Informally, the root nodes are the nodes without any incoming edges.

Definition 6.3. Parents
Let $\langle V, E \rangle$ be a DAG. Then, for any given node $A \in V$ we call $\mathit{parents}(A) := \{B \in V \mid (B, A) \in E\}$ the parents of A. Informally, the parents of A are just the nodes which have an outgoing edge to A.

Definition 6.4. Children
Let $\langle V, E \rangle$ be a DAG. Then, we define $\mathit{children}(A) := \{B \in V \mid (A, B) \in E\}$ as the children of A. Informally, the children of A are all the nodes that have an incoming edge from A.

Definition 6.5. Descendants
Let $\langle V, E \rangle$ be a DAG. A descendant of A is defined as
$\mathit{descendants}(A) := \{B \in V \mid \exists (A_1, B_1), \ldots, (A_n, B_n) \in E : B_i = A_{i+1}$ and $A_1 = A$ and $B_n = B\}$
Informally, the descendants of A are all the nodes that can be reached by going along directed edges starting at A.

Definition 6.6. Chain
Let $\langle V, E \rangle$ be a DAG. If $e := (A, B) \in E$ then we call $e' := (B, A)$ the inverse edge of e. A chain between X and Y with $X, Y \in V$ is a sequence of edges $e_1, \ldots, e_n$ with $e_i := (A_i, B_i)$ such that

- $e_i \neq e_j$ for $i \neq j$
- $A_1 = X$ and $B_n = Y$
- $B_i = A_{i+1}$ for $i \in \{1, \ldots, n-1\}$
- for all e_i holds $e_i \in E$ or $e'_i \in E$.

A chain is simply a path between two nodes where we are ignoring the direction of the arrows. Condition one says we don't go any edge twice, two says we start at the right point and also end at the right node. Condition three guarantees that the nodes in between are actually connected and condition four says that the the paths we want to go are really on our DAG.

Definition 6.7. d-separation
Let $\mathbb{X}, \mathbb{Y}, \mathbb{Z} \subseteq V$ be sets of nodes in a DAG $\langle V, E\rangle$. Z d-separates $\mathbb{X}$ from $\mathbb{Y}$, if and only if for every chain $\shortmid$ connecting elements of $\mathbb{X}$ and $\mathbb{Y}$ there exists a node (variable) C such that

- $C \in \mathbb{Z}$ and the arrows in $\shortmid$ meet head-to-tail at C,
- $C \in \mathbb{Z}$ and the arrows in $\shortmid$ meet tail-to-tail at C,
- $C \notin \mathbb{Z}$ none of C's descendants is in $\mathbb{Z}$ and the arrows in $\shortmid$ at C meet head-to-head.

Definition 6.8. Conditional Independence
Let $\mathbb{X} := \{X_1, \ldots, X_k\}$, $\mathbb{Y} := \{Y_1, \ldots, Y_r\}$, and $\mathbb{Z} := \{Z_1, \ldots, Z_m\}$ be sets of random/propositional variables with $k, r, m \in \mathbb{N}$. Then, we define the conditional independence of $\mathbb{X}$ of $\mathbb{Y}$ given $\mathbb{Z}$ as follows:

$$\mathbb{X} \perp\!\!\!\perp \mathbb{Y} \mid \mathbb{Z} \text{ iff } P(X_1 = x_1, \ldots, X_k = x_k, Y_1 = y_1, \ldots, Y_r = y_r \mid Z_1 = z_1, \ldots, Z_m = z_m) =$$
$$P(X_1 = x_1, \ldots, X_k = x_k \mid Z_1 = z_1, \ldots, Z_m = z_m) \cdot P(Y_1 = y_1, \ldots, Y_r = y_r \mid Z_1 = z_1, \ldots, Z_m = z_m)$$

for all $x_1, \ldots, x_k, y_1, \ldots, y_r, z_1, \ldots, z_m \in \mathbb{R}$ or $\in \{true, false\}$ respectively for propositional variables. Moreover, we have to assume $P(Z_1 = z_1, \ldots, Z_m = z_m) > 0$.

Definition 6.9. Parental Markov Condition
Let $(\Omega, P, \mathcal{P}(\Omega))$ a probability space and $\langle V, E\rangle$ be a DAG, where each of the nodes in $A \in V$ represents a random/propositional variable. We say that $\langle V, E\rangle$ with $(\Omega, P, \mathcal{P}(\Omega))$ satisfies the parental markov condition (PMC) if for any $A \in V$ holds that A is conditionally independent of any (possibly combinations) of its non-descendants given all its parents.

Definition 6.10. Bayesian Network
Formally a Bayesian Network is a tuple $\langle \mathbb{G}, \mathbb{P}\rangle$ where $\mathbb{G} = \langle V, E\rangle$ is a directed acyclic graph, $\mathbb{P} = (\Omega, P, \mathcal{P}(\Omega))$ is a probability space and V is a set of random/propositional variables that satisfies the parental markov condition.
Intuitively a Bayesian Network is a DAG with a joint probability distribution on the nodes of the DAG that satisfies PMC.

Comment 6.11. *Important Graphical Structures*
The following three structures are the most important in Bayesian Networks and are named as follows:

- *$X \to Y \to Z$ is called a directed chain.*
- *$X \to Y \leftarrow Z$ is called a collider.*
- *$X \leftarrow Y \to Z$ is called a fork, also called common cause in causal language.*

Definition 6.12. Faithful
We call a Bayesian Network $\langle\langle V, E\rangle \mathbb{P}\rangle$ *faithful*, iff the PMC+semi graphoid axioms imposed on $\mathbb{G}$ entails all and only the conditional independences of $\mathbb{P}$.

Definition 6.13. Minimality
Let $\langle \mathbb{G}, \mathbb{P}\rangle$ be a Bayesian Network. We say that the DAG $\mathbb{G}$ is *minimal* with $\mathbb{P}$ if the following holds: if we remove any arrows from $\mathbb{G}$, the resultant DAG $\mathbb{G}'$ no longer satisfies the PMC with $\mathbb{P}$.

Definition 6.14. Semi Graphoid Axioms
The following four axioms together are called the semi graphoid axioms.[8]

1. Symmetry: $X \perp\!\!\!\perp Y \mid Z \;\to\; Y \perp\!\!\!\perp X \mid Z$.
2. Decomposition: $X \perp\!\!\!\perp Y, W \mid Z \;\to\; X \perp\!\!\!\perp Y \mid Z$.
3. Weak Union: $X \perp\!\!\!\perp Y, W \mid Z \;\to\; X \perp\!\!\!\perp Y \mid Z, W$.
4. Contraction: $(X \perp\!\!\!\perp Y \mid Z \;\&\; X \perp\!\!\!\perp W \mid Y, Z) \;\to\; X \perp\!\!\!\perp Y, W \mid Z$.

Theorem 6.15. *Power of d-separation*
If the given Bayesian Network $\langle\langle V, E\rangle, \mathbb{P}\rangle$ is faithful, for $X, Y, Z \subseteq V$ it holds that $X \perp\!\!\!\perp Y \mid Z$, if and only if Z d-separates X from Y. (Without the faithfulness assumption, only the $\Leftarrow$ direction holds.)

[8] I put commas between the (sets of) random variables since I find it easier to read. However, it means the same as Jürgen's definitions.

Theorem 6.16. *Product Rule for Bayes Nets*
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space and $\langle V, E\rangle$.
Define an ancestral ordering on variables, all parents have a smaller number than all of their children. (Count top-down, with roots at the top and leaves at the bottom.)
For any $A_1, \dots, A_n \subseteq \Omega$ with $\bigcap_{i=1}^{n-1} A_i \neq \emptyset$ with $n \in \mathbb{N}$ holds the following:

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i \mid \bigcap_{g=i+1}^{n} X_g = x_g \text{ where } X_g \text{ is a parent of } X_i)).$$

7 Entropy

The entropy is a measure of average information or surprise.

Definition 7.1. Information of an outcome
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Then, for every $\omega \in \Omega$ we can define

$$I(\omega) := -\log_2(P(\omega)).$$

Using this particular function is motivated by several properties of the negative log. It only assigns non-negative values in $[0;1]$. Independent information about two events is the same as the sum of the two pieces of information. The basis 2 of the logarithm is chosen for the outcome of a fair coin toss having exactly the information 1. Often also the natural logarithm to the basis e is used. The unit of information is Shannon[9] or sometimes also bit in computer science. We use the convention that $0\log(0) = 0$ which is a continuous extension of $x\log(x)$ in zero.

Definition 7.2. Entropy of a probability distribution
Let $(\Omega, P, \mathcal{P}(\Omega))$ be a probability space. Then, the *entropy* of P is defined as:

$$H(P) := -\sum_{\omega\in\Omega} P(\omega)\log_2(P(\omega)) = \sum_{\omega\in\Omega} P(\omega)I(\omega)$$

The entropy is informally nothing but the expected amount of information.
Much more often then the entropy of our measure we are interested in the entropy of a random variable. How much information will this random variable give me on average? Thus, let X be a random variable taking values $\{x_1, \dots, x_n\} \in \mathbb{R}$.[10] Then, the entropy of X is nothing but the entropy of P_X. Therefore,

$$H(X) := H(P_X) = -\sum_{x\in\{x_1,\dots,x_n\}} P(x)\log_2(P(x))$$

Theorem 7.3. *Properties of the Entropy*
Let Ω be a sample space and $\mathbb{P}$ be the set of well-defined probability measures on Ω. The following are interesting properties of the Entropy H.

- $\underset{P\in\mathbb{P}}{\mathit{argsup}}\, H(P) = P'$ *with* $P' \sim Unif(\Omega)$.
- *For all $P \in \mathbb{P}$ holds that if $P(\omega) = 1$ for some $\omega \in \Omega$ then $H(P) = 0$.*

Example 7.4. Let $\Omega := \{w_1, w_2, w_3, w_4, w_5, w_6\}$ be some sample space and $P \sim Unif(\Omega)$. Then, the entropy of P is:

$$H(P) = -\sum_{i=1}^{6} P(w_i)\log_2(P(w_i)) = -\sum_{i=1}^{6}\frac{1}{6}\log_2(\frac{1}{6}) = \log_2(6) \approx 2.585$$

Let $X(w_i) := \begin{cases} 0 \text{ if } i = 1,2,3 \\ 1 \text{ if } i = 4,5 \\ 2 \text{ if } i = 6 \end{cases}$ then, the Entropy of X is given by:

$$E(X) = -\sum_{i=0}^{2} P_X(i)\log(P_X(i)) = -(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{3}\log_2(\frac{1}{3}) + \frac{1}{6}\log_2(\frac{1}{6})) =\approx 1.459$$

I am not an expert on entropy nor familiar with maxEnt, thus I leave this open for now. If you have things to add feel welcome to send me an email.

[9] In honor of the founder of information theory Claude E. Shannon.
[10] This also works for $\{true, false\}$.