V7 Foundations of Probability Theory

„**Probability**“ : degree of confidence that an event of an uncertain nature will occur.

„Events“ : we will assume that there is
an agreed upon **space** Ω of possible outcomes („events“).

E.g. a normal die (*dt. Würfel*) has a space $\Omega = \{1,2,3,4,5,6\}$

Also we assume that there is a set of **measurable events S**
to which we are willing to assign probabilities.

In the die example, the event $\{6\}$ is the case where the die shows 6.

The event $\{1,3,5\}$ represents the case of an odd outcome.

Foundations of Probability Theory

Probability theory requires that the **event space** satisfy 3 basic properties:

- It contains the **empty event** $\varnothing$ and the **trivial event** Ω.

- It is **closed under union** $\rightarrow$ if $\alpha, \beta \in S$, then so is $\alpha \cup \beta \in S$,

- It is **closed under complementation** $\rightarrow$ if $\alpha \in S$, then so is $\Omega - \alpha \in S$

The requirement that the event space is closed under union and complementation implies that it is also closed under other Boolean operations, such as intersection and set difference.

Probability distributions

A **probability distribution** P *over* (Ω, S) is a mapping from events in S to real values that satisfies the following conditions:

(1) $P(\alpha) \geq 0$ for all $\alpha \in S$ → *Probabilities are not negative*

(2) $P(\Omega) = 1$ → *The probability of the trivial event which allows all possible outcomes has the maximal possible probability of 1.*

(3) If $\alpha, \beta \in S$ and $\alpha \cap \beta = 0$ then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$

Interpretation of probabilities

The **frequentist's** interpretation:

The **probability** of an event is the **fraction of times**
the event occurs if we repeat the experiment indefinitely.

E.g. throwing of dice, coin flips, card games, … where frequencies
will satisfy the requirements of proper distributions.

For an event such as „It will rain tomorrow afternoon“,
the frequentist approach does not provide a satisfactory interpretation.

Interpretation of probabilities

An alternative interpretation views probabilities as **subjective degrees of belief**.

E.g. the statement „the probability of rain tomorrow afternoon is 50 percent“ tells us that in the opinion of the speaker, the chances of rain and no rain tomorrow afternoon are the same.

When we discuss probabilities in the following we usually do not explicitly state their interpretation since both interpretations lead to the same mathematical rules.

Conditional probability

The **conditional probability** of β given α is defined as

$$P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$$

The probability that β is true given that we know α is the relative proportion of outcomes satisfying β among these that satisfy α.

From this we immediately see that

$$P(\alpha \cap \beta) = P(\alpha)P(\beta|\alpha)$$

This equality is know as the **chain rule** of conditional probabilities.

More generally, if α_1, α_2, … α_k are events, we can write

$$P(\alpha_1 \cap \alpha_2 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2|\alpha_1) \ldots P(\alpha_k|\alpha_1 \cap \cdots \cap \alpha_{k-1})$$

Bayes rule

Another immediate consequence of the definition of conditional probability is **Bayes‘ rule**.

Due to symmetry, we can swap the 2 variables α and β in the definition $P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$ and get the equivalent expression $P(\alpha|\beta) = \frac{P(\beta \cap \alpha)}{P(\beta)}$

If we rearrange, we get Bayes‘ rule $P(\beta|\alpha)P(\alpha) = P(\alpha|\beta)P(\beta)$ or

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

A more general conditional version of Bayes‘ rule where all probabilities are conditioned on some background event γ also holds:

$$P(\alpha|\beta \cap \gamma) = \frac{P(\beta|\alpha \cap \gamma)P(\alpha|\gamma)}{P(\beta|\gamma)}$$

Example 1 for Bayes rule

Consider a student population.
Let `Smart` denote smart students and `GradeA` denote students who got grade A.

Assume we believe that P(`GradeA|Smart`) = 0.6, and now we learn that a particular student received grade A.

Suppose that P (`Smart`) = 0.3 and P (`GradeA`) = 0.2

Then we have P(`Smart|GradeA`) = 0.6 × 0.3 / 0.2 = 0.9

In this model, an A grade strongly suggests that the student is smart.

On the other hand, if the test was easier and high grades were more common, e.g. P (`GradeA`) = 0.4, then we would get

P(`Smart|GradeA`) = 0.6 × 0.3 / 0.4 = 0.45 which is much less conclusive.

Example 2 for Bayes rule

Suppose that a tuberculosis skin test is 95% percent accurate.

That is, if the patient is TB-infected, then the test will be positive with probability 0.95 and if the patient is not infected, the test will be negative with probability 0.95.

Now suppose that a person gets a positive test result.

What is the probability that the person is infected?

Naive reasoning suggests that if the test result is wrong 5% of the time, then the probability that the subject is infected is 0.95.

That would mean that 95% of subjects with positive results have TB.

Example 2 for Bayes rule

If we consider the problem by applying Bayes‘ rule, we need to consider the prior probability of TB infection, and the probability of getting a positive test result.

Suppose that 1 in 1000 of the subjects who get tested is infected $\rightarrow$ P(TB) = 0.001

We see that 0.001×0.95 infected subjects get a positive result
and 0.999×0.05 uninfected subjects get a positive result.

Thus P(Positive) = $0.001 \times 0.95 + 0.999 \times 0.05 = 0.0509$

Applying Bayes‘ rule, we get P(TB|Positive) = P(TB) $\times$ P(Positive|TB) / P(Positive)
= $0.001 \times 0.95 / 0.0509 \cong 0.0187$

Thus, although a subject with a positive test is much more probable to be TB-infected than is a random subject, fewer than 2% of these subjects are TB-infected.

Random Variables

A **random variable** is defined by a function
that associates with each outcome in Ω a value.

For students in a class, this could be a function f_{grade} that maps
each student in the class (in Ω) to his or her grade (1, …, 5).

The event `grade = A` is a shorthand for the event $\{\omega \in \Omega: f_{grade}(\omega) = A\}$.

There exist categorical (or discrete) random values that take on
one of a few values, e.g. intelligence could be „high“ or „low“.

There also exist integer or real random variable that can take on
an infinite number of continuous values, e.g. the height of students.

By Val(X) we denote the set of values that a random variable X can take.

Random Variables

In the following, we will either consider categorical random variables or random variables that take real values.

We will use capital letters X, Y, Z to denote random variables.

Lowercase values will refer to the values of random variables.

E.g. $P(X = x) \geq 0 \text{ for all } x \in Val(X)$

When we discuss categorical random numbers, we will denote the *i-th* value as x^i.

Bold capital letters are used for sets of random variables: **X, Y, Z**.

Marginal Distributions

Once we define a random variable X, we can consider the **marginal distribution** *P(X)* over events that can be described using X.

E.g. let us take the two random variables `Intelligence` and `Grade` and their marginal distributions `P(Intelligence)` and `P(Grade)`

Let us suppose that

$$P(\text{Intelligence=high}) = 0.3$$
$$P(\text{Intelligence=low}) = 0.7$$

$$P(\text{Grade=A}) = 0.25$$
$$P(\text{Grade=B}) = 0.37$$
$$P(\text{Grade=C}) = 0.38$$

These marginal distributions are probability distributions satisfying the 3 properties.

Joint Distributions

Often we are interested in questions that
involve the values of several random variables.

E.g. we might be interested in the event „`Intelligence = high` and `Grade = A`“.

In that case we need to consider the **joint distribution** $P(X_1, \ldots, X_n)$
over these two random variables.

The joint distribution of 2 random variables has to be consistent
with the marginal distribution in that $P(x) = \sum_y P(x, y)$.

		Intelligence		
		low	*high*	
Grade	*A*	0.07	0.18	0.25
	B	0.28	0.09	0.37
	C	0.35	0.03	0.38
		0.7	0.3	1

Conditional Probability

The notion of conditional probability extends to induced distributions over random variables.

$P(\text{Intelligence}|\text{Grade}=\text{A})$ denotes the conditional distribution over the events describable by `Intelligence` given the knowledge that the student's grade is A.

Note that the conditional probability $P(\text{Intelligence}=\text{high}|\text{Grade}=\text{A}) = \frac{0.18}{0.25} = 0.72$ is quite different from the marginal distribution $P(\text{Intelligence}=\text{high}) = 0.3$.

We will use the notation $P(X|Y)$ to present a set of conditional probability distributions.

Bayes' rule in terms of conditional probability distributions reads

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)}$$

Independence

We usually expect $P(\alpha \mid \beta)$ to be different from $P(\alpha)$.

Learning that β is true typically changes our probability over α.

However, in some situations $P(\alpha \mid \beta) = P(\alpha)$.

Definition: We say that an event α is **independent** of event β in P, denoted as $(\alpha \perp \beta)$ if $P(\alpha \mid \beta) = P(\alpha)$ or if $P(\beta) = 0$.

We will now provide an alternative definition for this concept of independence.

Independence

Proposition: A distribution *P* satisfies $(\alpha \perp \beta)$ if and only if $P(\alpha \cap \beta) = P(\alpha)P(\beta)$

Proof

If $P(\beta) = 0 \rightarrow (\alpha \perp \beta)$. Also $P(\alpha \cap \beta) = 0$ so that $P(\alpha \cap \beta) = P(\alpha)P(\beta)$ is fulfilled.

Let now $P(\beta) \neq 0$

$\Rightarrow$ From the chain rule we get $P(\alpha \cap \beta) = P(\alpha|\beta)P(\beta)$

Since α is independent of β, $P(\alpha|\beta) = P(\alpha)$. Thus we get $P(\alpha \cap \beta) = P(\alpha)P(\beta)$

$\Leftarrow$ Suppose that $P(\alpha \cap \beta) = P(\alpha)P(\beta)$. Then, by definition we have

$$P(\alpha|\beta) = \frac{P(\alpha \cap \beta)}{P(\beta)} = \frac{P(\alpha)P(\beta)}{P(\beta)} = P(\alpha)$$

which is what needs to be shown.

Note that $(\alpha \perp \beta)$ implies $(\beta \perp \alpha)$

Independence of Random Variables

Definition: Let **X, Y, Z** be sets of random variables.

We say that **X** is conditionally independent of **Y**
given **Z** in a distribution *P*
if *P* satisfies

$$(\mathbf{X} = \mathbf{x} \perp \mathbf{Y} = \mathbf{y} | \mathbf{Z} = \mathbf{z})$$

for all values $\boldsymbol{x} \in Val(\boldsymbol{X})$, $\mathbf{y} \in Val(\mathbf{Y})$, $\mathbf{z} \in Val(\mathbf{Z})$.

As before we can give an alternative characterization of conditional independence

Proposition: The distribution *P* satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if

$$P(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z}) = P(\mathbf{X} \mid \mathbf{Z})\, P(\mathbf{Y} \mid \mathbf{Z})$$

Independence properties of distributions

Symmetry $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z})$

Decomposition $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$

Weak union $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$

Contraction $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Z}, \mathbf{Y}) \,\&\, (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z})$

Probability Density Functions

A function $p: \mathbb{R} \to \mathbb{R}$

is a **probability density function** (PDF) for X

if it is a nonnegative integrable function so that $\int_{Val(X)} p(x)dx = 1$

The function $\mathrm{P}(X \leq a) = \int_{-\infty}^{a} p(x)dx$ is the **cumulative distribution** for X.

By using the density function we can evaluate the probability of other events. E.g.

$$\mathrm{P}(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$

Uniform distribution

The simplest PDF is the **uniform distribution**

Definition: A variable X has a uniform distribution over [a,b] denoted X ~ Unif[a,b] if it has the PDF

$$p(x) = \begin{cases} \dfrac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise} \end{cases}$$

Thus the probability of any subinterval of [*a,b*] is proportional to its size relative to the size of [*a,b*].

If $b - a < 1$, the density can be greater than 1.

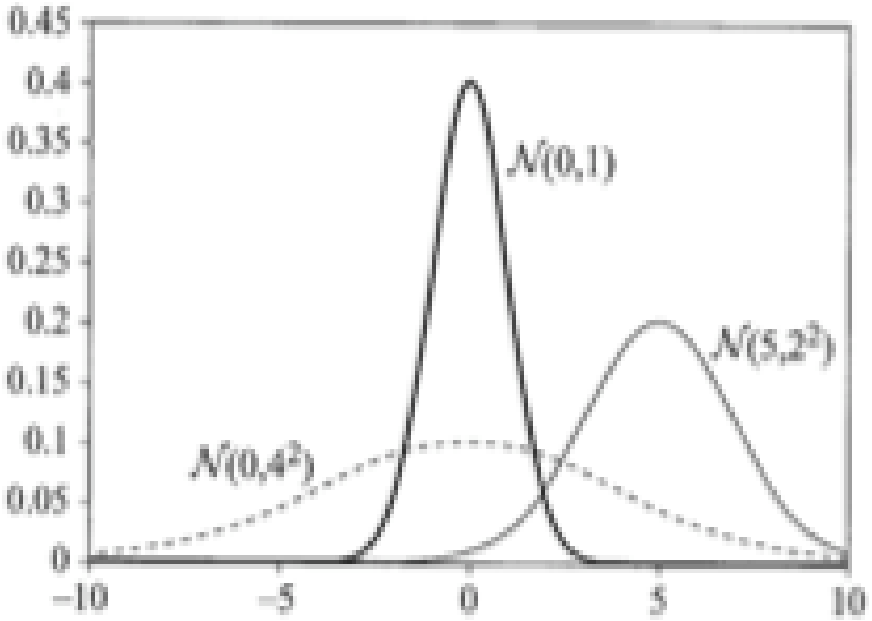
We only have to satisfy the constraint that the total area under the PDF is 1.

Gaussian distribution

A random variable X has a Gaussian distribution with mean μ and variance σ^2 , denoted $X \sim N(\mu;\sigma^2)$ if it has the PDF

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

A standard Gaussian has mean 0 and variance 1.



Joint density functions

Let P be a joint distribution over continuous random variables $X_1, \ldots X_n$.
A function $p(x_1, \ldots x_n)$ is a **joint density function** of $X_1, \ldots X_n$ if

- $p(x_1, \ldots x_n) \geq 0$ for all values $x_1, \ldots x_n$ of $X_1, \ldots X_n$
- p is an integrable function
- for any choice of $a_1, \ldots a_n$ and $b_1, \ldots b_n$

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} p(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

From the joint density we can derive the marginal density of any random variable by integrating out the other variables. E.g. if $p(x,y)$ is the joint density of X and Y

$$p(x) = \int_{-\infty}^{\infty} p(x, y) dy$$

Conditional density functions

We now want to be able to describe conditional distributions of continuous variables.

Applying the previous definition $P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)}$ is problematic because the probability of an isolated point $P(X = x)$ is zero.

Thus we define $P(Y|x) = \lim_{\varepsilon \to 0} P(Y|x - \varepsilon \leq X \leq x + \varepsilon)$

If there exists a continuous joint density function *p(x,y)* then we can derive the form of this term.

Let us consider some event on Y, say $a \leq Y \leq b$.

Conditional density functions

$$P(a \le Y \le b | x-\varepsilon \le X \le x+\varepsilon) = \frac{P(a \le Y \le b, x-\varepsilon \le X \le x+\varepsilon)}{P(x-\varepsilon \le X \le x+\varepsilon)}$$

$$= \frac{\int_a^b \int_{x-\varepsilon}^{x+\varepsilon} p(x',y)\,dy\,dx'}{\int_{x-\varepsilon}^{x+\varepsilon} p(x')\,dx'}$$

When ε is sufficiently small, we can assume that *p(x)* = const in this interval and approximate

$$\int_{x-\varepsilon}^{x+\varepsilon} p(x')\,dx' \approx 2\varepsilon\, p(x)$$

Using a similar approximation for *p(x',y)* , we get

$$P(a \le Y \le b, x-\varepsilon \le X \le x+\varepsilon) \approx \frac{\int_a^b 2\varepsilon\, p(x,y)\,dy}{2\varepsilon\, p(x)} = \int_a^b \frac{p(x,y)}{p(x)}\,dy$$

We conclude that $\frac{p(x,y)}{p(x)}$ is the density of *P(Y | X = x)*

Conditional density functions

Let *p(x,y)* be the joint density of *X* and *Y.*

The **conditional density function** of *Y* given *X* is defined as

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

When *p(x)* = 0, the conditional density is undefined.

The properties of joint distributions and conditional distributions carry over to joint and conditional density functions.

In particular, we have the **chain rule**

$$p(x,y) = p(x)p(y|x)$$

and **Bayes‘ rule**

$$p(x|y) = \frac{p(x)p(y|x)}{p(y)}$$

Conditional density functions

Definition:

Let **X, Y** and **Z** be sets of continuous random variables with joint density *p(***X, Y**,**Z***).*

We say that **X** is conditionally independent of **Y** given **Z**

if $p(x|z) = p(x|y,z)$ for all *x,y, z* such that *p(z) > 0*

Expectation

Let X be a discrete random variable that takes numerical values.

Then, the **expectation** of X under the distribution P is

$$\mathbf{E}_P[X] = \sum_x x \cdot P(x)$$

If X is a continuous variable,
then we use the density function

$$\mathbf{E}_P[X] = \int x \cdot p(x) dx$$

E.g. if we consider X to be the outcome of rolling a good die with probability 1/6 for each outcome, then **E**[X] = 1 · 1/6 + 2 · 1/6 + … + 6 · 1/6 = 3.5

Properties of the expectation of a random variable

$\mathbf{E}[a \cdot X + b] = a\,\mathbf{E}[X] + b$

Let X and Y be two random variables

$\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$

Here, it does not matter whether X and Y are independent or not.

What can be say about the expectation value of a product of two random variables?

In the general case very little.

Consider 2 variables X and Y that take each on the values +1 and -1 with probabilities 0.5.

If X and Y are independent, then $\mathbf{E}[X \cdot Y] = 0$.

If they always take the same value (they are correlated), then $\mathbf{E}[X \cdot Y] = 1$.

Properties of the Expectation of a random variable

If X and Y are independent then

E[X · Y]= **E**[X] · **E**[Y]

The **conditional expectation** of X given y is

$$E_P[X|y] = \sum_x x \cdot P(x|y)$$

Variance

The expectation of X tells us the mean value of X. However, it does not indicate how far X deviates from this value. A measure of this deviation is the **variance** of X:

$$Var_P[X] = \mathbf{E}_P[(X - \mathbf{E}_P[X])^2]$$

The variance is the expectation of the squared difference between X and its expected value. An alternative formulation of the variance is

$$Var[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

If X and Y are independent, then $Var[X + Y] = Var[X] + Var[Y]$

$$Var[a \cdot X + b] = a^2 Var[X]$$

For this reason, we are often interested in the square root of the variance, which is called the **standard deviation** of the random variable. We define

$$\sigma_X = \sqrt{Var[X]}$$

Variance

Let X be a random variable with Gaussian distribution $N(\mu;\sigma^2)$.
Then **E**[X] = μ and Var[X] = σ^2.

Thus, the parameters of the Gaussian distribution specify the expectation and the variance of the distribution.
The form of the Gaussian distribution implies that the density of values of X drops exponentially fast in the distance (x - μ) / σ.

Not all distributions show such a rapid decline in the probability of outcomes that are distance from the expectation.
However, even for arbitrary distributions, one can show that there is a decline.

Chebyshev inequality states $P(|X - \mathbf{E}_P[X]| \geq t) \leq \frac{Var_P[X]}{t^2}$

or in terms of σ $P(|X - \mathbf{E}_P[X]| \geq k\sigma_X) \leq \frac{1}{k^2}$