

Probability theory

Gergely Alpár

Institute for Computing and Information Sciences – Digital Security

Radboud University

Version: Autumn 2015



Outline

Probability

Combinatorics

Distributions

Random variables

Continuous random variables

Conditional probability and Bayes' rule

Final remarks



Historical background

“Probability” is the part of mathematics that looks for laws governing random events. It has its origins in games of chance *i.e.* in gambling. Chevalier de Méré (1607-1684) was a famous gambler and a friend of Blaise Pascal, who started to develop probability theory.

Example (Question about rolling dices)

What is more likely to get:

- (1) at least one ace (1) in 4 rolls of one dice
- (2) at least one pair (1,1) in 24 simultaneous rolls of two dice?

Chevalier expected (2), and lost money as a result.

- ▶ $p_1 = 1 - \left(\frac{5}{6}\right)^4 \approx 0.518$ (or 51.8% chance)
- ▶ $p_2 = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.491$



Currently we are here...

Probability

Combinatorics

Distributions

Random variables

Conditional probability and Bayes' rule

Experiments and their sample spaces

- ▶ An experiment is called **random** if the result will vary even if the conditions are the same
- ▶ A **sample space** consists of all possible outcomes of a random experiment, usually denoted with the letter S or Ω

Example (What are the relevant sample spaces?)

- (1) coin tossing once: $S = \{T, H\}$
- (2) coin tossing twice: $S = \{TT, TH, HT, HH\}$
- (3) dice tossing: $S = \{1, 2, 3, 4, 5, 6\}$
- (4) lifetime of a bulb: $S = \{t \mid 0 \leq t \leq 10 \text{ years}\}$

(Oxford dictionary: Historically, dice is the plural of die, but in modern standard English dice is both the singular and the plural)



Events

Definition

An **event** is a subset of outcomes of a random experiment, that is, a subset of the sample space.

We write the powerset $\mathcal{P}(S) = \{A \mid A \subseteq S\}$ for the set of events.

We use the notation: union \cup , intersection \cap , empty set \emptyset , set minus \setminus and not \neg .

Example (for sample space S)

- ▶ the entire subset $S \subseteq S$ is *the certain* event
- ▶ $\emptyset \subseteq S$ is the *impossible* event
- ▶ two events A and B are **mutually exclusive** if $A \cap B = \emptyset$.



Probability measure

Definition

A **probability measure** P for a sample space S is a function that gives for each event $A \subseteq S$ a probability $P(A) \in [0, 1]$, with:

- (1) Axiom 1: $P(S) = 1$
- (2) Axiom 2: for mutually exclusive events $A, B \subseteq S$ (i.e. $A \cap B = \emptyset$)
 $P(A \cup B) = P(A) + P(B)$

A probability measure on S is thus a function $P: \mathcal{P}(S) \rightarrow [0, 1]$ satisfying (1) and (2).

It is called **discrete** if the sample space S is finite; this implies that there are only finitely many events.

(Officially, discrete spaces can also be countable, but we shall not use those here)



Properties of probability measures

Theorem

Let P be a probability measure on space S , and let A, A_i, B , be events.
Then:

- (1) $A \subseteq B \Rightarrow P(A) \leq P(B)$
- (2) $P(\emptyset) = 0$
- (3) $P(\neg A) = 1 - P(A)$, where $\neg A = S - A = \{s \in S \mid s \notin A\}$
- (4) For pairwise mutually exclusive events A_1, A_2, \dots, A_n (i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$ and $n \geq 2$) one has
$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$
- (5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- (6) $P(A) = P(A \cap B) + P(A \cap \neg B)$

The points can all be derived from the axioms (1) and (2) for a probability measure P .



Example: proof of point (1)

Proof.

- ▶ Required: $A \subseteq B \implies P(A) \leq P(B)$
 - Assumption: $A \subseteq B$
 - Claim: $P(A) \leq P(B)$
- ▶ We can write B as **disjoint union** $B = A \cup (B \setminus A)$, where:
 - $B \setminus A = B \cap \neg A = \{s \in S \mid s \in B \text{ and } s \notin A\}$
 - $A \cap (B \setminus A) = \emptyset$
- ▶ By Axiom 2 we get: $P(B) = P(A) + P(B \setminus A)$
- ▶ Since $P(B \setminus A) \geq 0$, by definition it is in $[0, 1]$, we get $P(B) \geq P(A)$.

□



Currently we are here...

Probability

Combinatorics

Distributions

Random variables

Conditional probability and Bayes' rule

Combinatorics = smart counting

Combinatorics is a branch of mathematics that studies **counting**, typically in finite structures, of objects satisfying certain criteria.

Example (Counting permutations)

- ▶ A **permutation** of n objects is a rearrangement in some order
- ▶ **Question:** how many different permutations are there of n objects?
 - Try to think of the answer for $n = 2, 3, 4, \dots$
- ▶ The **answer** is $n! = n \cdot (n - 1) \cdot (n - 2) \cdots 2 \cdot 1$
 - Pronounce: $n!$ as “ n factorial”
 - For those who like recursion: $n! = n \cdot (n - 1)!$ and $0! = 1$.



Fundamental principle of (successive) counting

- ▶ Suppose that a task involves a sequence of k successive choices
 - let n_1 be the number of options at the first stage;
 - let n_2 be the number of options at the second stage, after the first stage has occurred;
 - ...
 - let n_k term be the number of options at the k -th stage, after the previous $k - 1$ stages have occurred.
- ▶ Then the **total number of different ways** the task can occur is:

$$n_1 \cdot n_2 \cdot \dots \cdot n_k = \prod_{1 \leq i \leq k} n_i$$



Simple counting example

A company places a 6-symbol code on each unit of its products, consisting of:

- ▶ 4 digits, the first of which is the number 3,
- ▶ followed by 2 letters, the first of which is NOT a vowel.

How many different codes are possible?

Using the basic counting principle:

- ▶ there is 1 option for digit 1 (namely, '3')
- ▶ there are 10 options (decimals) for digits 2, 3, 4
- ▶ 5 of the letters in the alphabet are vowels (a, e, i, o, u), so that means there are $26 - 5 = 21$ options for letter 1
- ▶ there are 26 letters in the alphabet, 26 options for letter 2

Altogether there are $1 \cdot 10 \cdot 10 \cdot 10 \cdot 21 \cdot 26 = 546,000$ different codes.



Ordered samples with repetition

Question

- ▶ Suppose you have n objects, and you take an **ordered** sample **with repetition** of r out of them (with $r \leq n$)
- ▶ This means that the **order** of the selected r elements matters, and the same element may be selected **multiple** times
- ▶ How many such samples are there?

Example (2-samples out of 3 elements, say $\{1, 2, 3\}$)

- ▶ samples: 11, 12, 13, 21, 22, 23, 31, 32, 33
- ▶ number of samples: $9 = 3^2$

Lemma

There are n^r ordered samples with repetition.



Unordered samples without repetition

Recall two things:

- ▶ **permutation** of all items: there are $n!$ ways to order/permute n items
- ▶ there are $\frac{n!}{(n-r)!}$ ordered samples without repetition ($0 < r \leq n$)
 - What's happening if $r = n$? ($0! = 1$)

Combining these two yields:

Lemma

There are $\frac{n!}{r!(n-r)!}$ unordered samples without repetition.

One writes $\binom{n}{r} = \frac{n!}{r!(n-r)!}$. This is called the **binomial coefficient**.

It is pronounced as “ n choose r ”.

An unordered sample is sometimes called a **combination**.



Examples (of unordered samples without repetition)

Example (Lotto with 49 numbered balls)

How many possible outcomes are there if we consecutively pick 6 balls?

Answer: $\binom{49}{6} = 13,983,816$

Example

Find the number of ways to form a committee of 5 people from a set of 9.

Answer: $\binom{9}{5} = 126$. (what is the difference if it is permutation?)

Example

How many symmetric keys are needed so that n people can all communicate directly with each other?

Answer: $\binom{n}{2} = \frac{n(n-1)}{2} = (n-1) + (n-2) + \cdots + 2 + 1$



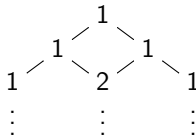
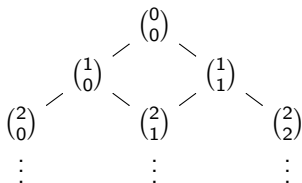
Calculation rules for binomial coefficients

$$(1) \binom{n}{r} = \binom{n}{n-r}$$

$$(2) \sum_{r=0}^n \binom{n}{r} = 2^n$$

$$(3) \binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}$$

Recall also Pascal's triangle



Binomial expansion of powers of sums

- ▶ Recall: $(x + y)^2 = x^2 + 2xy + y^2$
 $= \binom{2}{0}x^2y^0 + \binom{2}{1}x^1y^1 + \binom{2}{2}x^0y^2$
- ▶ Similarly: $(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$
 $= \binom{3}{0}x^3y^0 + \binom{3}{1}x^2y^1 + \binom{3}{2}x^1y^2 + \binom{3}{3}x^0y^3$

Lemma

For arbitrary $n \in \mathbb{N}$,

$$(x + y)^n = \sum_{i=0}^{i=n} \binom{n}{i} x^{n-i} y^i$$



Computing probability from combinatorics

When

- ▶ the number of ways for an event to occur and
- ▶ the number of all outcomes

can be counted (often by combinatorics), we can compute the probability as follows:

$$P(\text{event } A) = \frac{\text{\#ways } A \text{ can occur}}{\text{\#all possible outcomes}} = \frac{\text{\#POSITIVE}}{\text{\#ALL}}$$



Birthday paradox

- (1) What is the probability that at least 2 of r randomly selected people have the same birthday?
- (2) How large must r be so that the probability is greater than 50%?



Solution, part 1

- ▶ Assume all birthdays are equally distributed (also, we exclude leap days, that is, presume no one is born on 29 February)
 - $n = 365$, and samples of r ordered with repetition (once a birthday occurs, it is not excluded, since it can occur again)
 - $n^r = 365^r$ birthday options for r people
- ▶ Look at r birthdays, all at **different days**
 - number of options:
 $365 \cdot 364 \cdots (365 - (r - 1)) = \frac{365!}{(365-r)!} = \binom{365}{r} r!$
 - take fraction: the probability that r people have their birthday on **different days** is:

$$\frac{\frac{365!}{(365-r)!}}{365^r} = \frac{365!}{(365-r)! \cdot 365^r}$$

- ▶ Therefore, the probability that **at least 2 people out of r have their birthday on the same day** is $P(r) = 1 - \frac{365!}{(365-r)! \cdot 365^r}$



Solution, part 2

Some values for $p(r) = 1 - \frac{365!}{(365-r)! \cdot 365^r}$, depending on r .

r	$P(r)$
10	0.117
20	0.411
22	0.476
23	0.507
30	0.706
50	0.970
68	0.999

Hence for $r = 23$ the probability of birthday-coincidence is $\geq 50\%$.



Application: Birthday attacks on hash functions

- ▶ SHA1 with a 160 bit output requires brute-force work of at most 2^{80} operations
 - (although because of weaknesses in SHA1 collisions are found already in around 2^{60} steps)
- ▶ In general hash functions used for signature schemes should have the number of output bits n large enough such that $2^{n/2}$ computations are impractical

Note: With \$8M budget an 80-bit key can be retrieved in a year (2011).



Currently we are here...

Probability

Combinatorics

Distributions

Random variables

Conditional probability and Bayes' rule

Discrete sample space example

Recall that a sample space S is called **discrete** if it is **finite**

Example (One dice)

- ▶ $S = \{1, 2, 3, 4, 5, 6\}$, with **events** $A \subseteq S$
- ▶ The probability measure $P: \mathcal{P}(S) \rightarrow [0, 1]$ is easy:
 - $P(\{1, 3, 5\}) = \frac{1}{2}$
 - $P(\{1, 6\}) = \frac{1}{3}$
- ▶ We see that P is determined by what it does on **elementary events** $\{i\} \subseteq S$
- ▶ This is typical for finite (and countable) sample spaces.



Discrete sample spaces

Let S be a **discrete** (i.e. finite) sample space, with probability measure $P: \mathcal{P}(S) \rightarrow [0, 1]$.

- ▶ An event $A \subseteq S$ is then also finite, say $A = \{x_1, \dots, x_n\}$
- ▶ Hence we can write it as **disjoint union of elementary events**:

$$A = \{x_1\} \cup \dots \cup \{x_n\}$$

- ▶ Hence $P(A) = P(\{x_1\}) + \dots + P(\{x_n\})$, by Axiom 2.
- ▶ Thus, P is entirely determined by its values $P(\{x\})$ on elementary events, for $x \in S$.
- ▶ The function $f: S \rightarrow [0, 1]$ with $f(x) = P(\{x\})$ is called the underlying **distribution**
- ▶ It satisfies $\sum_{x \in S} f(x) = 1$ since:

$$\sum_{x \in S} f(x) = \sum_{x \in S} P(\{x\}) = P(\bigcup_{x \in S} \{x\}) = P(S) = 1$$



The uniform distribution

Fix a number $n \in \mathbb{N}$ and take as sample space $S = \{1, 2, \dots, n\}$.

- ▶ The simplest distribution is the **uniform** distribution $u_n: S \rightarrow [0, 1]$, which assigns the same probability to each $i \in S$
- ▶ Since the sum of probabilities must be 1, the only option is:

$$u_n(i) = \frac{1}{n}$$

- ▶ More generally, on each finite set X we can define $u: X \rightarrow [0, 1]$ as $u(x) = \frac{1}{\#X}$, where $\#X \in \mathbb{N}$ is the number of elements of X .



The binomial distribution

Fix $n \in \mathbb{N}$ with $S = \{0, 1, \dots, n\}$ and $p \in [0, 1]$.

- Define the **binomial distribution** $b: S \rightarrow [0, 1]$ as:

$$b(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- Read $b(k)$ as:

*the probability of exactly k successes after n trials,
each with chance p*

Briefly: $b(k) = P(k \text{ out of } n)$.

- This is well-defined distribution by **binomial expansion**:

$$\sum_k b(k) = \sum_k \binom{n}{k} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1^n = 1$$



Example binomial expansion

Suppose we have a **biased coin**, which comes up head with probability $p \in [0, 1]$ (not necessarily $p = \frac{1}{2}$).

Example (Toss the coin $n = 5$ times)

What is the probability of getting head k times (for $0 \leq k \leq 5$)?

- ▶ If $k = 0$, then: $(1 - p)^5$
 - via the formula: $b(0) = \binom{5}{0} p^0 (1 - p)^{5-0} = (1 - p)^5$
- ▶ If $k = 1$, then: $5p(1 - p)^4$
 - $b(1) = \binom{5}{1} p^1 (1 - p)^{5-1} = 5p(1 - p)^4$
- ▶ In general: $b(k) = \binom{5}{k} p^k (1 - p)^{5-k}$.

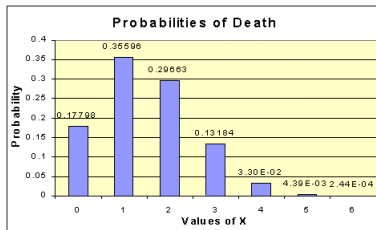
What happens if $p = \frac{1}{2}$?



Another binomial distribution example

Hospital records show that of patients suffering from a certain disease, 75% die of it. What is the probability that of 6 randomly selected patients, 4 will recover?

- ▶ We have $n = 6$, with recovery probability $p = \frac{1}{4}$.
- ▶ Hence $b(4) = \binom{6}{4}(\frac{1}{4})^4(\frac{3}{4})^2 \approx 0,0329595$
- ▶ Picture of all (recovery) probabilities in a **histogram**



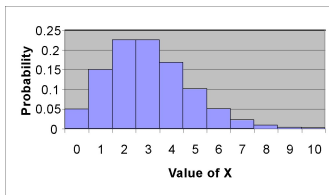
(source: intmath.com)



Other distributions

There are many other standard distributions, like:

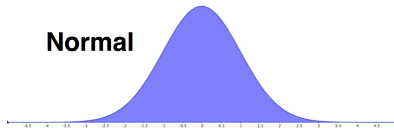
- ▶ **Normal** distribution (see later in the continuous case)
- ▶ **Hypergeometric** distribution
- ▶ **Poisson distribution**
 - for independent occurrences, where some average μ is known
 - then $p(k) = e^{-\mu} \cdot \frac{\mu^k}{k!}$, for $k \in \mathbb{N}$. For instance, for $\mu = 3$,



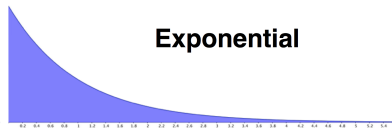
We will not discuss these distributions here. Look up the details, later in your life, when you need them.

Examples for continuous distributions

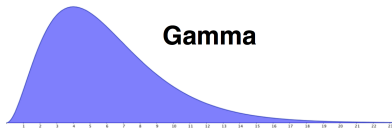
Normal



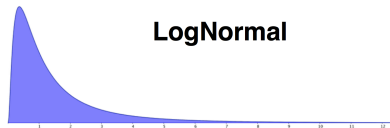
Exponential



Gamma



LogNormal



Partitions

Recall the two axioms (A(1), A(2)) and the six theorems (T(1)–T(6)) w.r.t. a probability measure P defined over a sample space S . (See Slides 5 and 6)

Definition

A **partition** of a sample space S is a collections of events $A_1, \dots, A_n \subseteq S$ with both:

$$A_1 \cup \dots \cup A_n = S \quad \text{and} \quad A_i \cap A_j = \emptyset, \text{ for } i \neq j$$

In particular, a binary partition is given by $A, \neg A$.

Therefore, for any probability measure P and a partition A_1, \dots, A_n , $P(A_1) + \dots + P(A_n) = 1$, because:

$$P(A_1) + \dots + P(A_n) \stackrel{T(4)}{=} P(A_1 \cup \dots \cup A_n) \stackrel{\text{by def.}}{=} P(S) \stackrel{A(1)}{=} 1$$



Currently we are here...

Probability

Combinatorics

Distributions

Random variables

Continuous random variables

Conditional probability and Bayes' rule

Random variable

Associating a value with an experiment

- ▶ Suppose, we throw a coin, so the sample space is $S = \{H, T\}$
- ▶ If the outcome is H , I get 100€, otherwise I lose 100€
- ▶ This situation is described via a **random variable** $X: S \rightarrow \mathbb{R}$
 - $X(H) = 100, X(T) = -100$

Definition

Let S be a sample space. A **random variable** is a real function defined on S , of the form $X: S \rightarrow \mathbb{R}$.

A random variable is also called a **stochastic** variable



More about random variables

- ▶ A random variable $X: S \rightarrow \mathbb{R}$ that takes on a finite (or a countably infinite) number of values is called **discrete**
 - this means that the range $R(X) \subseteq \mathbb{R}$ is finite (or countable)
 - otherwise we have a **non-discrete** or **continuous** random variable
 - (Note that if S is discrete, then so is X .)
- ▶ If we have two random variables, say $X, Y: S \rightarrow \mathbb{R}$, then we can also define the random variables $X + Y, X - Y, rX$ in the obvious, pointwise manner:

$$\begin{aligned}(X + Y)(s) &= X(s) + Y(s) & (rX)(s) &= r \cdot X(s) \\ (X - Y)(s) &= X(s) - Y(s)\end{aligned}$$



Random variables and events

Definition

Let $X: S \rightarrow \mathbb{R}$ be a random variable, and $x \in \mathbb{R}$ be an outcome.

There is an event $(X = x) \subseteq S$, understood as “outcome is x ”, namely:

$$(X = x) = \{s \in S \mid X(s) = x\}.$$

If there is also a **probability measure** $P: \mathcal{P}(S) \rightarrow [0, 1]$, then we write $P(X = x) \in [0, 1]$ for the probability of this event $(X = x) \subseteq S$.

Lemma

If X is a discrete random variable, say with outcomes $\{x_1, \dots, x_n\}$, then the events $(X = x_1), \dots, (X = x_n)$ form a **partition**.



Overview: measures / distributions / random variables

Let S be a sample space. Recall:

- ▶ a **probability measure** P is a function from events to probabilities:

$$\mathcal{P}(S) \xrightarrow{P} [0, 1]$$

- ▶ If S is finite, P corresponds to a **distribution** from samples to probabilities:

$$S \xrightarrow{f} [0, 1] \quad \text{via} \quad f(s) = P(\{s\})$$

- ▶ A **random variable** is a function from samples to values:

$$S \xrightarrow{X} \mathbb{R}$$

It gives rise to events $(X = x) \subseteq S$, with probability $P(X = x) \in [0, 1]$.



Example

A (fair) coin is tossed twice times; we count the number of heads.

- ▶ the **sample space** is $S = \{HH, HT, TH, TT\}$
- ▶ We have a **uniform distribution** $f: S \rightarrow [0, 1]$ namely $f(s) = \frac{1}{4}$
- ▶ There is a “sum of heads” random variable $X: S \rightarrow \mathbb{R}$:

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0$$

We are in the discrete case, with range $R(X) = \{0, 1, 2\} \subseteq \mathbb{R}$.

- ▶ Example events with probabilities:

$$P(X = 0) = \frac{1}{4}, \quad P(X = 1) = \frac{1}{2}, \quad P(X = 2) = \frac{1}{4}.$$



Expectation and variance (discrete case)

Definition

Let $X: S \rightarrow \mathbb{R}$ be a discrete random variable, with values (range) $\{x_1, \dots, x_n\} \subseteq \mathbb{R}$.

- ▶ The **expectation** or **expected value** or **weighted mean** $E(X) \in \mathbb{R}$ is:

$$E(X) = P(X = x_1) \cdot x_1 + \dots + P(X = x_n) \cdot x_n$$

- ▶ The **variance** $Var(X) \in \mathbb{R}$ describes the spread

$$Var(X) = E((X - E(X))^2) = \sum_i P(X = x_i) \cdot (x_i - E(X))^2$$

- ▶ The **standard deviation** is: $\sigma_X = \sqrt{Var(X)}$.
 - an outcome in $[E(X) - \sigma_X, E(X) + \sigma_X]$ is considered “normal”



Dice example

We have $S = \{1, 2, 3, 4, 5, 6\}$ with $X: S \rightarrow \mathbb{R}$ simply $X(s) = s$

- ▶ $P(X = i) = \frac{1}{6}$, for $i = 1, 2, 3, 4, 5, 6$
- ▶
$$\begin{aligned} E(X) &= \sum_i P(X = i) \cdot i = \frac{1}{6} \cdot 1 + \dots + \frac{1}{6} \cdot 6 \\ &= \frac{1}{6} \cdot (1 + \dots + 6) = \frac{21}{6} = \frac{7}{2} = 3.5 \end{aligned}$$

The expectation is the **mean** for a uniform distribution

- ▶
$$\begin{aligned} \text{Var}(X) &= E((X - E(X))^2) = \sum_i P(X = i) \cdot (i - \frac{7}{2})^2 \\ &= \frac{1}{6}(1 - \frac{7}{2})^2 + \dots + \frac{1}{6}(6 - \frac{7}{2})^2 = \frac{35}{12} \end{aligned}$$
- ▶ $\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{\frac{35}{12}} \approx 1.71.$



Lottery example

Setting and question

- ▶ In a lottery there are 200 prizes of 50€, 20 of 250€, 5 of 1000€. Assuming that 10.000 tickets will be issued and sold, what is a fair price to pay for a ticket?
- ▶ **Answer:** a price that is just a bit more than the expected amount to be won.
- ▶ The sample space has 4 elements; we describe the random variable X for the amount to be won.

X	50	250	1000	0
$P(X = x)$	0.02	0.002	0.0005	0.9775

- ▶ $E(X) = \sum_i P(X = i) \cdot i = 0,02 \cdot 50 + 0,002 \cdot 250 + 0,0005 \cdot 1000 + 0 = 2$.
So, a ticket should cost (a bit more than) 2€.



Expectation for the binomial distribution

Recall the parameters are $n \in \mathbb{N}$ and $p \in [0, 1]$. Then:

- ▶ $S = \{0, 1, \dots, n\}$, to which we add $X: S \rightarrow \mathbb{R}$ with $X(i) = i$.
- ▶ $b(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

Lemma

The binomial distribution for $n \in \mathbb{N}$ and $p \in [0, 1]$ satisfies

$$E(X) = n \cdot p$$

Recall the hospital example, with $n = 6$ patients and $p = \frac{1}{4}$ recovery probability.

The expected number of survivors is $6 \cdot \frac{1}{4} = \frac{3}{2}$.



Expectation for the binomial distribution, proof

$$\begin{aligned} E(X) &= \sum_{k=0}^{k=n} P(X = k) \cdot k = \sum_{k=0}^{k=n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{k=n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{k=n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np \sum_{i=0}^{i=n-1} \binom{n-1}{i} p^i (1-p)^{(n-1)-i} \\ &= np(p + (1-p))^{n-1} = np \cdot 1^{n-1} = np. \quad \blacksquare \end{aligned}$$



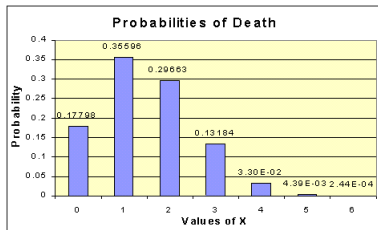
Events

- ▶ For a **discrete** random variable X we have seen probabilities of the form $P(X = x)$
- ▶ In the **continuous** case we shall look at $P(X \leq x)$ or also at $P(x \leq X \leq y)$
- ▶ They describe the probability of the **events**:
 - $(X \leq x) = \{s \in S \mid X(s) \leq x\} \subseteq S$
 - $(x \leq X \leq y) = \{s \in S \mid x \leq X(s) \leq y\} \subseteq S$



Probability, surface, and integration: discrete case

Recall the hospital records histogram used earlier:

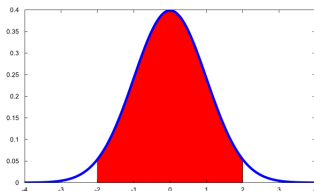


- ▶ We can read it as a function $f: \mathbb{R} \rightarrow \mathbb{R}$
- ▶ Suppose we are interested in the probability $P(1 \leq X \leq 3)$
- ▶ This probability is obtained as $P(1 \leq X \leq 3) = \int_1^3 f(x)dx$



Probability, surface, and integration: continuous case

This idea can be generalised, as suggested in:



$$\begin{aligned} P(-2 \leq X \leq 2) \\ = \int_{-2}^2 f(x) dx \end{aligned}$$

- ▶ The functions $f: \mathbb{R} \rightarrow \mathbb{R}$ used in such a way are called **probability density functions** (pdf, *dichtheidsfunctie*)
- ▶ They should satisfy: $f(x) \geq 0$ and: $\int_{-\infty}^{+\infty} f(x) dx = 1$.
- ▶ This connects the first and second part of this course!



Example pdf

- ▶ Often a **continuous** random variable is defined directly via a probability density function (pdf)
- ▶ For instance, consider $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by:

$$f(x) = \begin{cases} \frac{1}{4}x & \text{if } 1 \leq x \leq 3 \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ Clearly $f(x) \geq 0$ and:

$$\int_{-\infty}^{+\infty} f(x)dx = \int_1^3 \frac{1}{4}xdx = \left[\frac{1}{8}x^2\right]_1^3 = \frac{1}{8}(3^2 - 1^2) = 1.$$

- ▶ Hence we can define a continuous random variable as:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

For instance $P(1 \leq X \leq 2) = \frac{3}{8}$.



Cumulative distribution function

- ▶ In practice, a pdf is often given directly, and the random variable is then defined accordingly (by definite integrals)
- ▶ But one can also obtain the pdf from a random variable X
- ▶ Define the **cumulative distribution function** $F: \mathbb{R} \rightarrow [0, 1]$ as:

$$F(x) = P(X \leq x)$$

Then " $F(-\infty)$ " = 0 and " $F(+\infty)$ " = 1.

That is, $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.

- ▶ The pdf f is then the derivative $f = F' = \frac{d}{dx}P(X \leq x)$.
Then indeed:

$$\begin{aligned} P(X \leq b) &= F(b) = F(b) - F(-\infty) = \int_{-\infty}^b f(x)dx \\ P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx = \int_a^b f(x)dx \end{aligned}$$



Uniform probability, in the continuous case

- ▶ Consider the interval $[a, b] \subseteq \mathbb{R}$, for given $a < b$
- ▶ Define a density function $u_{a,b}: \mathbb{R} \rightarrow \mathbb{R}$ as:

$$u_{a,b}(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ Then:

$$\int_{-\infty}^{\infty} u_{a,b}(x) dx = \int_a^b \frac{1}{b-a} dx = \left. \frac{1}{b-a} x \right|_a^b = \frac{1}{b-a} (b - a) = 1.$$



Expectation, in the continuous case

Definition

Let X be continuous random variable, given by pdf f . Then:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x) dx$$

$$\sigma_X = \sqrt{\text{Var}(X)}.$$



Example, in the uniform case

Recall $u_{a,b}: \mathbb{R} \rightarrow \mathbb{R}$ with value $\frac{1}{b-a}$ on $[a, b]$. Then:

$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} x u_{a,b}(x) dx = \int_a^b \frac{x}{b-a} dx \\ &= \frac{1}{2(b-a)} x^2 \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(a-b)(a+b)}{2(b-a)} = \frac{a+b}{2} \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{+\infty} (x - E(X))^2 u_{a,b}(x) dx = \int_a^b \frac{(x - \frac{a+b}{2})^2}{b-a} dx \\ &= \frac{1}{4(b-a)} \int_a^b (2x - (a+b))^2 dx \\ &= \frac{1}{4(b-a)} \int_a^b 4x^2 - 4(a+b)x + (a+b)^2 dx \\ &= \frac{1}{4(b-a)} \left[\frac{4}{3} x^3 - 2(a+b)x^2 + (a+b)^2 x \right]_a^b \\ &= \dots = \frac{1}{12} (a-b)^2. \end{aligned}$$



Properties, also for the discrete case

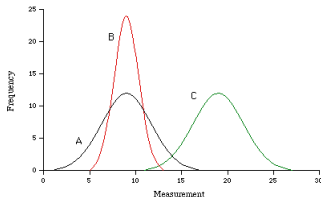
Lemma

- ▶ $E(X + Y) = E(X) + E(Y)$
- ▶ $E(r \cdot X) = r \cdot E(X)$
- ▶ $\text{Var}(X) = E(X^2) - E(X)^2$



Normal (Gaussian) distribution

An important class of probability density functions is of this form:



- ▶ Such **bell curves** are typical for normal/Gaussian distributions
- ▶ The curve is determined by two parameters
 - the **mean**, written as μ , which determines the location of the middle of the bell
 - the **variance**, written as σ^2 , which determines the width
- ▶ This distribution is very common in practice, when observations pile up around a particular value

Normal (Gaussian) distribution, definition

Definition

Given parameters μ for mean and σ for variance, consider the pdf

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

The associated random variable X is called **normal** or **Gaussian**. Its distribution function is thus:

$$P(a \leq X \leq b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

One writes $X \sim N(\mu, \sigma)$.

The integral cannot be computed exactly, so one uses tables of cumulative probabilities for a special normal distribution to calculate the probabilities.



Standardising normal distribution

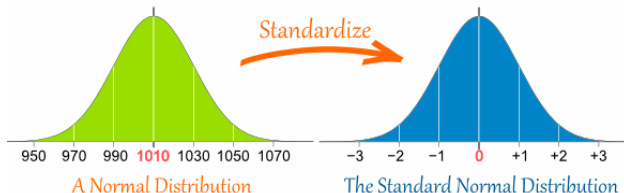
$N(0, 1)$ is often called the **standard** normal random variable; the pdf involved is $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$; it is centered around 0.

Theorem

If X is a normal random variable with mean μ and standard deviation σ , then $Z = \frac{X - \mu}{\sigma}$ is a standard normal random variable.

It satisfies $E(Z) = 0$ and $\sigma_Z = 1$.

In a picture:



Standardisation example

Setting and question

Suppose students can get 100 marks for their exam; the output is:

$marks = [85, 24, 63, 12, 87, 90, 33, 38, 25]$

Which ones should “**normally**” fail?

Via a small Python program we get $\mu = 50.78$ and $\sigma = 28.94$.

The list $[(i - \mu)/\sigma \text{ for } i \text{ in } marks]$ of standardized marks is:

$[1.18, -0.93, 0.42, -1.34, 1.25, 1.36, -0.61, -0.44, -0.89]$

By construction this list has **mean 0** and **deviation 1**.

Only one entry deviates < -1 ; this one should fail “normally”; it corresponds to the result 12 in the original list.



Currently we are here...

Probability

Combinatorics

Distributions

Random variables

Conditional probability and Bayes' rule

Final remarks

Conditional probability intro

Example (Suppose you throw one dice)

- ▶ Of course, the probability of 4 is $\frac{1}{6}$
- ▶ But what is the probability of 4, if you already know that the outcome is even?
- ▶ Intuitively it is clear it should be: $\frac{1}{3}$.
- ▶ We write $P(4) = \frac{1}{6}$ and $P(4 \mid \text{even}) = \frac{1}{3}$

Conditional probability is about **updating** probabilities in the light of **given** (aka **prior**) information.



Conditional probability example

Assume a group of students for which:

- ▶ The probability that a student does mathematics and computer science is $\frac{1}{10}$
- ▶ The probability that a student does computer science is $\frac{3}{4}$.

Question: What is the probability that a student does mathematics, given that we know that (s)he does computer science?

Answer: We have $P(M \cap CS) = \frac{1}{10}$ and $P(CS) = \frac{3}{4}$.

We seek the conditional probability $P(M | CS) = \text{"M, given CS"}$

The formula is:

$$P(M | CS) = \frac{P(M \cap CS)}{P(CS)} = \frac{\frac{1}{10}}{\frac{3}{4}} = \frac{4}{30} = \frac{2}{15}.$$



Basic definitions

Definition

For two events A, B , the **conditional probability** $P(A \mid B)$ = “the probability of A , given B ”, is

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

Alternatively, $P(A \mid B) \cdot P(B) = P(A \cap B)$.

Definition

Two events A, B are **independent** if $P(A \cap B) = P(A) \cdot P(B)$.

Equivalently, $P(A \mid B) = P(A)$.



Conditional probability, for multiple events

► Recall $P(A_1 \cap A_2) = P(A_1 | A_2) \cdot P(A_2)$

► Hence

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1 | A_2 \cap A_3) \cdot P(A_2 \cap A_3) \\ &= P(A_1 | A_2 \cap A_3) \cdot P(A_2 | A_3) \cdot P(A_3). \end{aligned}$$

► Alternatively:

$$P(A_1 | A_2 \cap A_3) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_2 \cap A_3)} = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_2 | A_3) \cdot P(A_3)}$$

► This can be generalised to A_1, \dots, A_n .



Partitions and the total probability lemma

Lemma (Total probability)

For a partition A_1, \dots, A_n and arbitrary event B ,

$$P(B) = P(B \mid A_1) \cdot P(A_1) + \dots + P(B \mid A_n) \cdot P(A_n).$$

Because:

$$\begin{aligned} & P(B \mid A_1) \cdot P(A_1) + \dots + P(B \mid A_n) \cdot P(A_n) \\ &= P(B \cap A_1) + \dots + P(B \cap A_n) \\ &= P((B \cap A_1) \cup \dots \cup (B \cap A_n)) \\ &= P(B \cap (A_1 \cup \dots \cup A_n)) \\ &= P(B \cap S) \\ &= P(B). \end{aligned}$$



Total probability illustration

Example (Two boxes with long & short bolts)

- ▶ In box 1, there are 60 short bolts and 40 long bolts. In box 2, there are 10 short bolts and 20 long bolts. Take one of the boxes at random, and pick a bolt also randomly. What is the probability that you choose a short bolt?
- ▶ Write B_i for the event that box i is chosen, for $i = 1, 2$
- ▶ The solution is:

$$\begin{aligned}P(\text{short}) &= P(\text{short} \mid B_1)P(B_1) + P(\text{short} \mid B_2)P(B_2) \\&= \frac{60}{100} \cdot \frac{1}{2} + \frac{10}{30} \cdot \frac{1}{2} \\&= \frac{3}{10} + \frac{1}{6} \\&= \frac{7}{15}.\end{aligned}$$

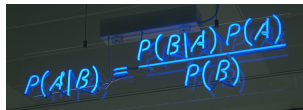


Bayes' Rule/Theorem

Theorem

For events E, H we have:

$$P(H | E) = \frac{P(E | H) \cdot P(H)}{P(E)}.$$



A photograph of a chalkboard with the formula $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$ written in blue chalk.

Terminology:

- ▶ E = evidence, H = hypothesis
- ▶ $P(H)$ = prior probability, $P(H | E)$ = posterior probability

Proof

$$P(E | H) \cdot P(H) = P(E \cap H) = P(H \cap E) = P(H | E) \cdot P(E).$$



Bayes' Rule/Theorem for partitions

Theorem

Suppose we have a partition H_1, \dots, H_n . Then:

$$P(H_i | E) = \frac{P(E | H_i) \cdot P(H_i)}{\sum_j P(E | H_j) \cdot P(H_j)}.$$

Proof Since $P(E) = \sum_j P(E | H_j) \cdot P(H_j)$ by the total probability lemma.



Machine example

Setting and question

- ▶ There are 3 machines M_1, M_2, M_3 producing items, with defect probabilities 0,01, 0,02, 0,03 respectively.
- ▶ 20% of items come from M_1 , 30% from M_2 , 50% from M_3
- ▶ Find the probability that a defect item comes from M_1 .

Solution

Question rephrased: Provided that an item is defect, what is the probability of M_1 : $P(M_1 | D)$?

- ▶ We have $P(M_1) = 0,2$, $P(M_2) = 0,3$, $P(M_3) = 0,5$ and $P(D | M_1) = 0,01$, $P(D | M_2) = 0,02$, $P(D | M_3) = 0,03$
- ▶ Using the total probability lemma, we compute $P(D)$ as:
$$P(D | M_1) \cdot p(M_1) + P(D | M_2) \cdot p(M_2) + P(D | M_3) \cdot p(M_3)$$
$$= 0,01 \cdot 0,2 + 0,02 \cdot 0,3 + 0,03 \cdot 0,5 = 0,023$$
- ▶ Then: $P(M_1 | D) = \frac{P(D|M_1) \cdot P(M_1)}{P(D)} = \frac{0,01 \cdot 0,2}{0,023} = 0,087$



Rain-and-umbrella example

Setting

- ▶ Prior knowledge $P(\text{rain}) = \frac{1}{5}$
- ▶ $P(\text{umbrella} \mid \text{rain}) = \frac{7}{10}$ and $P(\text{umbrella} \mid \neg\text{rain}) = \frac{1}{10}$
- ▶ Suppose you see someone with an umbrella. What is the probability that it rains?

Answer

$$\begin{aligned} &P(\text{rain} \mid \text{umbrella}) \\ &= \frac{P(\text{umbrella} \mid \text{rain}) \cdot P(\text{rain})}{P(\text{umbrella} \mid \text{rain}) \cdot P(\text{rain}) + P(\text{umbrella} \mid \neg\text{rain}) \cdot P(\neg\text{rain})} \\ &= \frac{\frac{7}{10} \cdot \frac{1}{5}}{\frac{7}{10} \cdot \frac{1}{5} + \frac{1}{10} \cdot \frac{4}{5}} = \frac{\frac{7}{50}}{\frac{7}{50} + \frac{4}{50}} = \frac{7}{11} \approx 0,64. \end{aligned}$$



Inference: learning from iterated observation

- ▶ In the previous example we started from $P(\text{rain}) = \frac{1}{5}$, and computed $P(\text{rain} \mid \text{umbrella}) = \frac{7}{11}$.
- ▶ Thus after observing this umbrella we may **update** our prior knowledge to $P'(\text{rain}) = \frac{7}{11}$.
- ▶ What if we see another, second umbrella? Surely, the probability of rain is even higher. How to compute it?
- ▶ We can play the same game with the updated rain probability $P'(\text{rain}) = \frac{7}{11}$.

$$P(\text{rain} \mid 2 \text{ umbrellas})$$

$$\begin{aligned} &= \frac{P(\text{umbrella} \mid \text{rain}) \cdot P'(\text{rain})}{P(\text{umbrella} \mid \text{rain}) \cdot P'(\text{rain}) + P(\text{umbrella} \mid \neg \text{rain}) \cdot P'(\neg \text{rain})} \\ &= \frac{\frac{7}{10} \cdot \frac{7}{11}}{\frac{7}{10} \cdot \frac{7}{11} + \frac{1}{10} \cdot \frac{4}{11}} = \frac{\frac{49}{110}}{\frac{49}{110} + \frac{4}{110}} = \frac{49}{53} \approx 0,92. \end{aligned}$$

- ▶ See courses on AI (esp. Machine Learning) for more information, esp. on Bayesian networks (graphical models)!



Overview

In this course we have:

- (1) seen/recalled the basics of mathematical calculus: derivation and integration
- (2) introduced the basics of probability theory

These areas are connected, through continuous random variables, where probabilities are computed using integration of a probability density function (pdf).



Preparing for the exam

- ▶ The emphasis is on being able to **calculate** things, not to prove properties
 - much like in the exercises
- ▶ The exam is “closed” book
- ▶ Hence, definitions and results must be known by heart, when they are relevant for doing calculations
- ▶ A simple calculator will be provided (only $+$, $-$, $*$, $/$)
- ▶ Assumed knowledge:
 - Essential rules and identities (algebraic, log, trigonometry, *etc.*)
 - New formulas and rules (functions, differentiation, integration, probability)
 - Slides (supported GeoGebra, and possibly by the lecture notes)



The role of the exercises

- ▶ The exercises form the **best preparation** for the exam!
- ▶ Assignment 8 is a **mock** exam
 - do it as a test for yourself, without the notes
 - hence you can see what you know and what you don't
 - an elaborated version will be put online, after the hand-in date
- ▶ To take the exam, you need
 - **five** sufficient assignments: min. 60 points
- ▶ There will be a second written exam (**resit**), where no precondition applies.



Summary of important points

► Derivation

- limits, derivatives and their interpretation as tangent, differentiation rules, elementary and special functions, function investigation, curve sketching

► Integration

- definite integral as area, indefinite integral as inverse to differentiation, improper integral, primitives of elementary and special functions, substitution & integration by parts, arc length

► Probability

- combinatorics, binomials, (discrete) probability measure, uniform & binomial distributions, random variable, expectation, standard deviation, pdf, normal distributions and standardisation, conditional probability, total probability, Bayes' rule



The exam itself

- ▶ **Date and time:** Monday, 26 Oct, 12:30 – 15:30 (arrive 15 minutes earlier!)
 - LIN 3, for the strong group (Bram's)
 - LIN 8, for the not-so-strong group (Joost's)
 - HG00.086, for “extra time” students (until 16:30)
- ▶ If you qualify for extra time, let me know **in advance**, via email.
- ▶ Make sure (and double-check) that you are **registered**
 - do this today!
 - registration cannot be done on the spot; you will be **excluded**

Good luck with the preparation, and of course, with the exam!

