Discrete Probability Theory

Jakob Piribauer

TU München

Summer 2023

Organisation

Lectures:
Wednesday, 3pm to 5:30pm, Forum (occasionally online, details will follow)

Exercise classes:
Thursday, 09:45am to 11:15pm, seminar room D.2.01
Thursday, 11:30am to 01:00pm, seminar room D.2.01

Exam:
written exam of 120 mins, date: TBA

Moodle course:
Slides and exercise sheets will be uploaded there.

Exercise sheets:
in-class exercises and homework exercises

Bonus system:
You may hand in homework exercises in groups of two.
If you hand in reasonable solutions to at 75% of the homework exercises, you will be awarded a bonus of 0,3 grades in the exam.

Literature

- S. Ross:
  A First Course in Probability (9th edition),
  Pearson Education, 2014.
- R. E. Walpole, R. H. Myers, S. L. Myers, K. Ye:
  Probability & Statistics for Engineers & Scientists (9th edition),
  Pearson Education, 2012.
- H. Gordon:
  Discrete Probability,
  Springer, 1997.
- M. Mitzenmacher, E. Upfal:
  Probability and Computing: Randomized Algorithms and Probabilistic Analysis,
  Cambridge University Press, 2005
- W. Feller:
  An introduction to probability theory and its applications,
  Wiley, 1971.

Outline

What is randomness?

Individual random events are unpredictable.

However, if the probability distribution is known, the frequency of different outcomes over repeated experiments (or "trials") is predictable.

Randomness in computer science

- randomized algorithms
- systems interact with a probabilistic environment
- failures of components occur randomly
- average case analysis: what is the expected run-time of an algorithm for a randomly chosen input?

Outline

Sample space

Definition (Sample space)

The set Ω of all possible outcomes of an experiment is called the *sample space*.

Example

When flipping a coin, we may say that the sample space is

$$\Omega = \{H, T\}$$

where the two outcomes represent heads and tails.

When flipping a coin three times, we may choose

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Sample space

Example

When rolling two dice, we might choose

$$\begin{aligned}\Omega_1 = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\end{aligned}$$

or

$$\Omega_2 = \{2,3,4,5,6,7,8,9,10,11,12\}.$$

Elements of Ω_1 contain a lot more information about the outcome.

Discrete probability space

Definition (Discrete probability space)

A *discrete probability space* consists of

- a finite or countably infinite sample space $\Omega = \{\omega_1, \omega_2, \dots\}$ and
- a probability function $\Pr\colon \Omega \to [0,1]$ such that

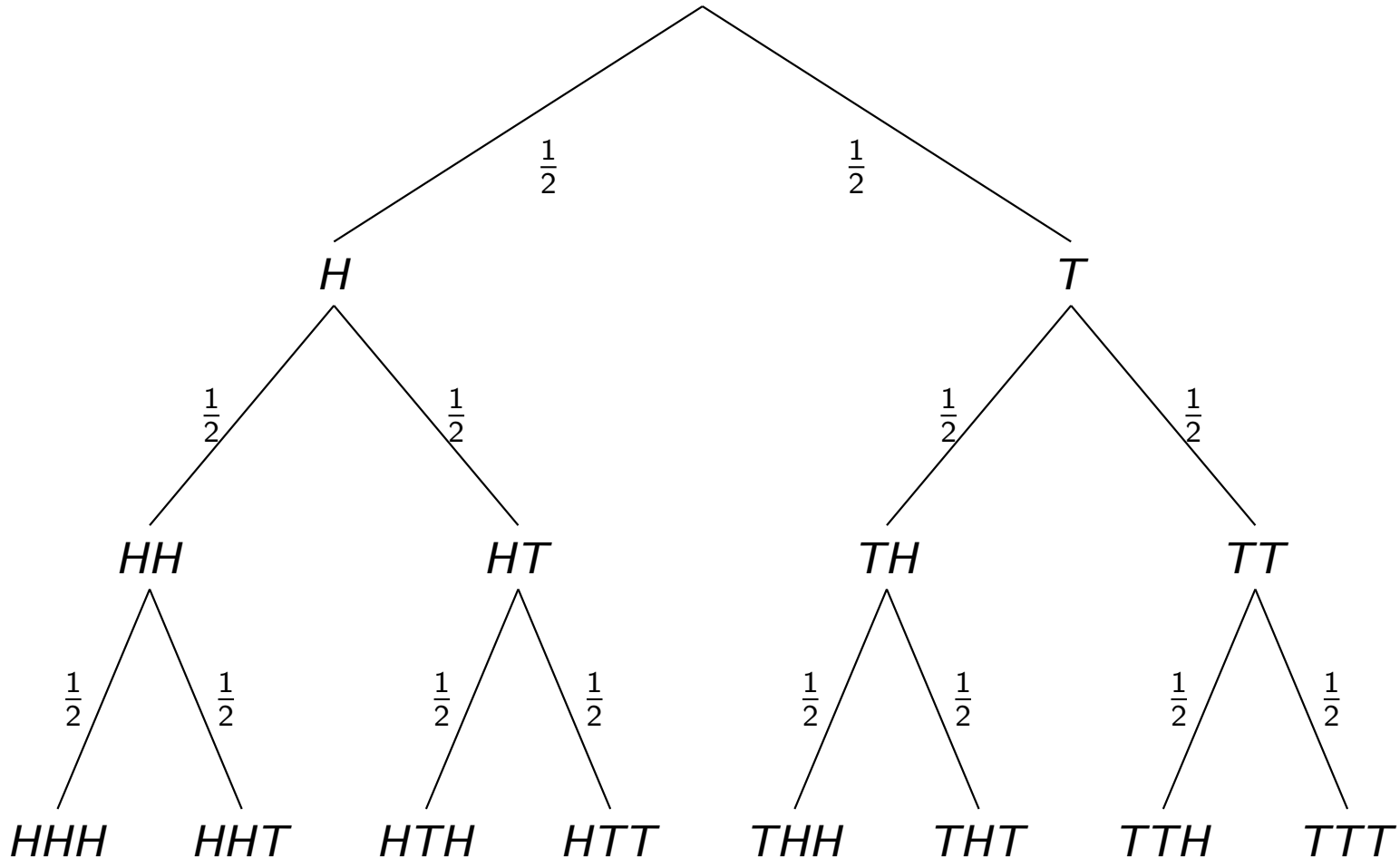
$$\sum_{\omega \in \Omega} \Pr(\omega) = 1.$$

Example

As probability function on $\Omega = \{H, T\}$ we can choose $\Pr(H) = \Pr(T) = \frac{1}{2}$.
We can also choose $\Pr(H) = p$ and $\Pr(T) = 1 - p$ for any $p \in [0,1]$.

On $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$, the probability function might be given by $\Pr(\omega) = \frac{1}{8}$ for all $\omega \in \Omega$.

Discrete probability space



Discrete probability space

Example

When rolling two dice, we might choose

$$\begin{aligned}\Omega_1 = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\end{aligned}$$

and $\Pr(\omega) = 1/36$ for all $\omega \in \Omega_1$.

Or

$$\Omega_2 = \{2,3,4,5,6,7,8,9,10,11,12\}$$

and

$$\Pr(2) = \frac{1}{36},\ \Pr(3) = \frac{2}{36},\ \ldots,\ \Pr(7) = \frac{6}{36},\ \Pr(8) = \frac{5}{36},\ \ldots,\ \Pr(12) = \frac{1}{36}.$$

Principle of indifference

The principle of indifference states that if there is no reason to believe otherwise, all possible outcomes should be assigned the same probability.

Formulated, e.g., by Jacob Bernoulli (1655 – 1705) and Pierre Simon Laplace (1749 – 1827).

Example

Suppose we are dealt two cards from a deck of 52 cards. What is the probability that we get the two black aces?

There are $\frac{52 \cdot 51}{2}$ combinations of two cards. There is no reason to believe that any of them is more likely than the others.
So, the probability for two black aces is

$$\frac{2}{52 \cdot 51} = \frac{1}{1326} \approx 0.00075.$$

Countably infinite probability space

Example

We toss a coin until heads comes up. The sample space is

$$\Omega = \{H, TH, TTH, TTTH, TTTTH, \dots\}.$$

If our coin shows tails with probability $p \in (0, 1)$, the probability of a sample point $T^i H$ is

$$\Pr(T^i H) = p^i \cdot (1 - p).$$

Is this a discrete probability space?

Yes, since

$$\sum_{i=0}^{\infty} \Pr(T^i H) = \sum_{i=0}^{\infty} p^i \cdot (1 - p) = (1 - p) \cdot \sum_{i=0}^{\infty} p^i = (1 - p) \cdot \frac{1}{1 - p} = 1.$$

Events

Definition (Event)

Let $(\Omega, \Pr)$ be a discrete probability space. An *event* is a set $E \subseteq \Omega$.
We can extend the probability function to events E by defining

$$\Pr(E) = \sum_{\omega \in \Omega} \Pr(\omega).$$

Example

Let $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ and $\Pr(\omega) = \frac{1}{8}$ for all $\omega \in \Omega$.

The event "exactly two throws are heads" is the set $E = \{HHT, HTH, THH\}$.

It's probability is $\Pr(E) = \sum_{\omega \in E} \Pr(\omega) = \frac{3}{8}$.

Events

We can use all set-theoretic operations to construct new events: For events $A, B \subseteq \Omega$, the following sets are also events:

- $A \cup B$,
- $A \cap B$,
- $A \setminus B$,
- $A^c = \Omega \setminus A$.

Two events A and B are called *disjoint* or *mutually exclusive* if $A \cap B = \emptyset$.

Events

Lemma

Let $(\Omega, \Pr)$ *be a discrete probability space and let* $A, B \subseteq \Omega$ *be events. Then,*

1. $\Pr(\emptyset) = 0$ *and* $\Pr(\Omega) = 1$.
2. $0 \leq \Pr(A) \leq 1$.
3. $\Pr(A^c) = 1 - \Pr(A)$.
4. *If* $A \subseteq B$, *then* $\Pr(A) \leq \Pr(B)$.
5. *If* A *and* B *are disjoint, then* $\Pr(A \cup B) = \Pr(A) + \Pr(B)$.

Proof.

1. $\Pr(\emptyset) = \sum_{\omega \in \emptyset} \Pr(\omega) = 0$ and $\Pr(\Omega) = \sum_{\omega \in \Omega} \Pr(\omega) = 1$ by the definition of probability spaces.

2. As $\Pr(\omega) \in [0, 1]$ for all $\omega \in \Omega$, we have $\Pr(A) = \sum_{\omega \in A} \Pr(\omega) \geq 0$ and $\Pr(A) = \sum_{\omega \in A} \Pr(\omega) \leq \sum_{\omega \in \Omega} \Pr(\omega) = 1$.

3. $1 = \sum_{\omega \in \Omega} \Pr(\omega) = \sum_{\omega \in A} \Pr(\omega) + \sum_{\omega \in A^c} \Pr(\omega) = \Pr(A) + \Pr(A^c)$.

4. If $A \subseteq B$, then

$$\Pr(A) = \sum_{\omega \in A} \Pr(\omega) \leq \sum_{\omega \in A} \Pr(\omega) + \sum_{\omega \in B \setminus A} \Pr(\omega) = \sum_{\omega \in B} \Pr(\omega) = \Pr(B).$$

5. If A and B are disjoint, then

$$\Pr(A \cup B) = \sum_{\omega \in A \cup B} \Pr(\omega) = \sum_{\omega \in A} \Pr(\omega) + \sum_{\omega \in B} \Pr(\omega) = \Pr(A) + \Pr(B). \qquad \square$$

Lemma

For pairwise disjoint events $A_1, A_2, \ldots, A_n$,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \Pr(A_i).$$

Analogously, for an infinite sequence $A_1, A_2, \ldots$ of pairwise disjoint events,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

Proof.

Follows as in the case of two disjoint events in the previous lemma. □

Events

What if events are not disjoint?

Lemma (Boole's inequality, George Boole (1815–1864))

Let $(\Omega, \Pr)$ *be a discrete probability space. For events* $A_1, A_2, \ldots, A_n \subseteq \Omega$,

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \Pr(A_i).$$

Analogously, for an infinite sequence $A_1, A_2, \ldots$ *of events,*

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \Pr(A_i).$$

Proof.
For the left hand side of the finite case, we get

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\omega \in \bigcup_{i=1}^{n} A_i} \Pr(\omega).$$

For the right hand side,

$$\sum_{i=1}^{n} \Pr(A_i) = \sum_{i=1}^{n} \sum_{\omega \in A_i} \Pr(\omega).$$

Everything that is summed up in the left hand side appears at least once in the sum for the right hand side. The infinite case follows analogously. □

Events

More precisely, we can show the following result:

Lemma

For two events A and B, we have

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

For three events A, B, and C, we have

$$\begin{aligned}\Pr(A \cup B \cup C) = &\Pr(A) + \Pr(B) + \Pr(C) \\ &- \Pr(A \cap B) - \Pr(A \cap C) - \Pr(B \cap C) \\ &+ \Pr(A \cap B \cap C).\end{aligned}$$

Proof.

For two events, we have $A \cup B = A \cup (B \setminus A)$ and A and $B \setminus A$ are disjoint. So,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B \setminus A).$$

But now, $B = (B \setminus A) \cup (A \cap B)$ and $B \setminus A$ and $A \cap B$ are disjoint. So,

$$\Pr(B) = \Pr(B \setminus A) + \Pr(A \cap B).$$

This means

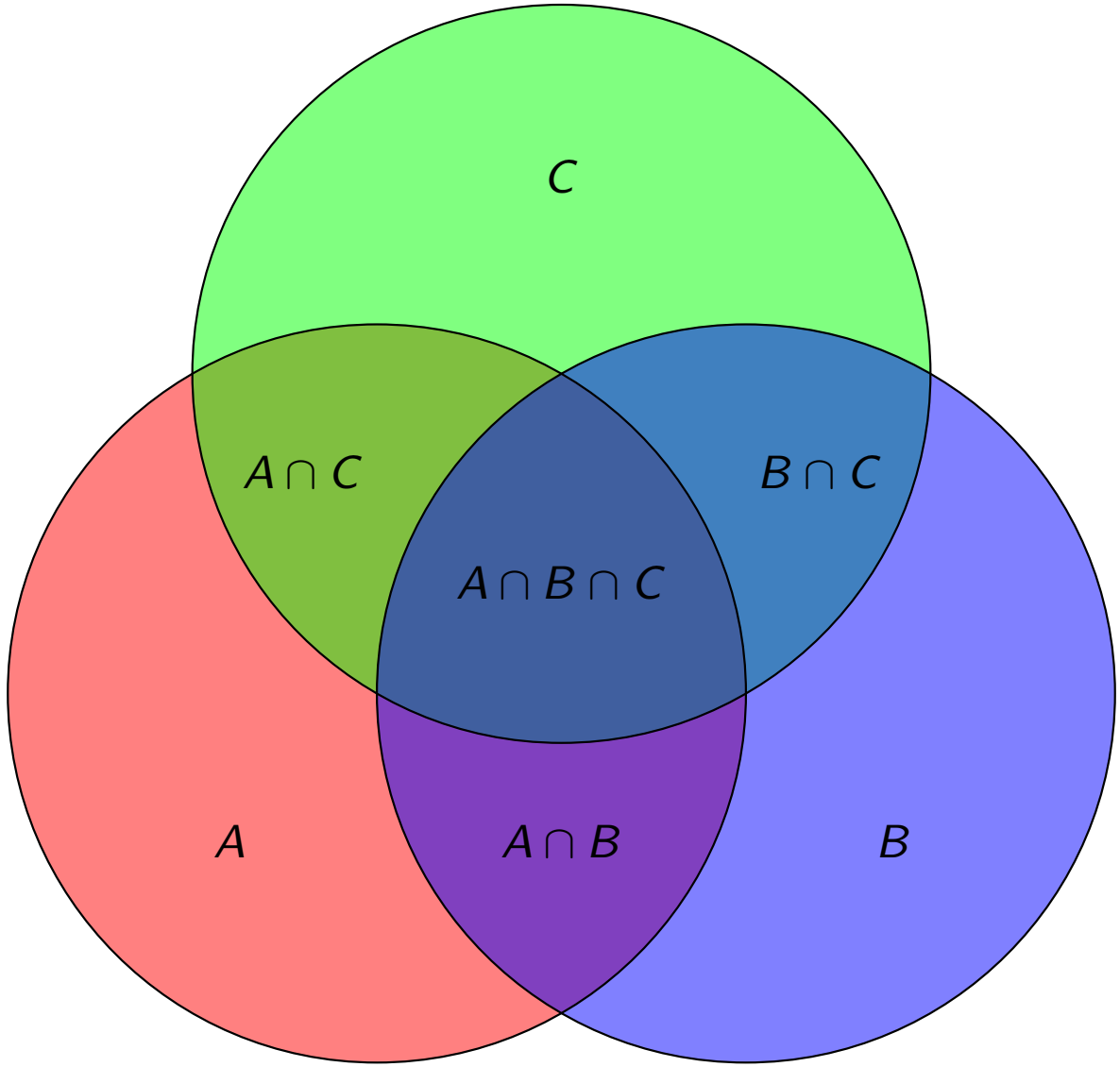
$$\Pr(B \setminus A) = \Pr(B) - \Pr(A \cap B)$$

and hence

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$$

Events

For three events, we illustrate the proof with the following picture.



In general, one can show (e.g., by induction):

Theorem (Inclusion–exclusion identity)

For events $A_1, \ldots, A_n$, we have

$$\Pr\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{r=1}^{n} (-1)^{r+1} \sum_{1 \le i_1 < \cdots < i_r \le n} \Pr\left(\bigcap_{j=1}^{r} A_{i_j}\right).$$

This theorem is sometimes also called the Poincaré-Sylvester-theorem, named after

Jules Henri Poincaré (1854–1912)

and

James Joseph Sylvester (1814–1897).

Conditional probabilities

Example

A flight departs on time with probability 0.8.
It arrives on time with probability 0.7.
It arrives and departs on time with probability 0.6.

What is the probability that a flight that departs on time arrives on time?

$$\Pr(\text{arrival on time} \mid \text{departure on time}) = \frac{0.6}{0.8} = 0.75.$$

What is the probability that a flight that arrives on time departed on time?

$$\Pr(\text{departure on time} \mid \text{arrival on time}) = \frac{0.6}{0.7} \approx 0.86.$$

Conditional probabilities

Definition (Conditional probability)

Let $(\Omega, \Pr)$ be a discrete probability space. Let A and B be events with $\Pr(B) > 0$. The *conditional probability* of A given B is defined as

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

Example

We toss a fair coin three times. Let $B = \{HHT, HTH, THH, HHH\}$ be the event that heads occurs at least two times. Let $A = \{HTT, THT, TTH, HHH\}$ be the event that heads occurs an odd number of times. Then,

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(\{HHH\})}{\Pr(\{HHT, HTH, THH, HHH\})} = \frac{1/8}{1/2} = \frac{1}{4}.$$

Conditional probabilities

Example

Assume a certain basketball player has a 50% chance to score on a free throw. The player is awarded two free throws. We do not see what happens, but due to the noise of the crowd we know that the player scored at least one of them. What is the chance that both throws were good?

Let $\Omega = \{OO, XO, OX, XX\}$ where X stands for a successful throw.

The probability of each sample point is $1/4$.

We are interested in the probability of $A = \{XX\}$ given that $B = \{XO, OX, XX\}$ is the case.

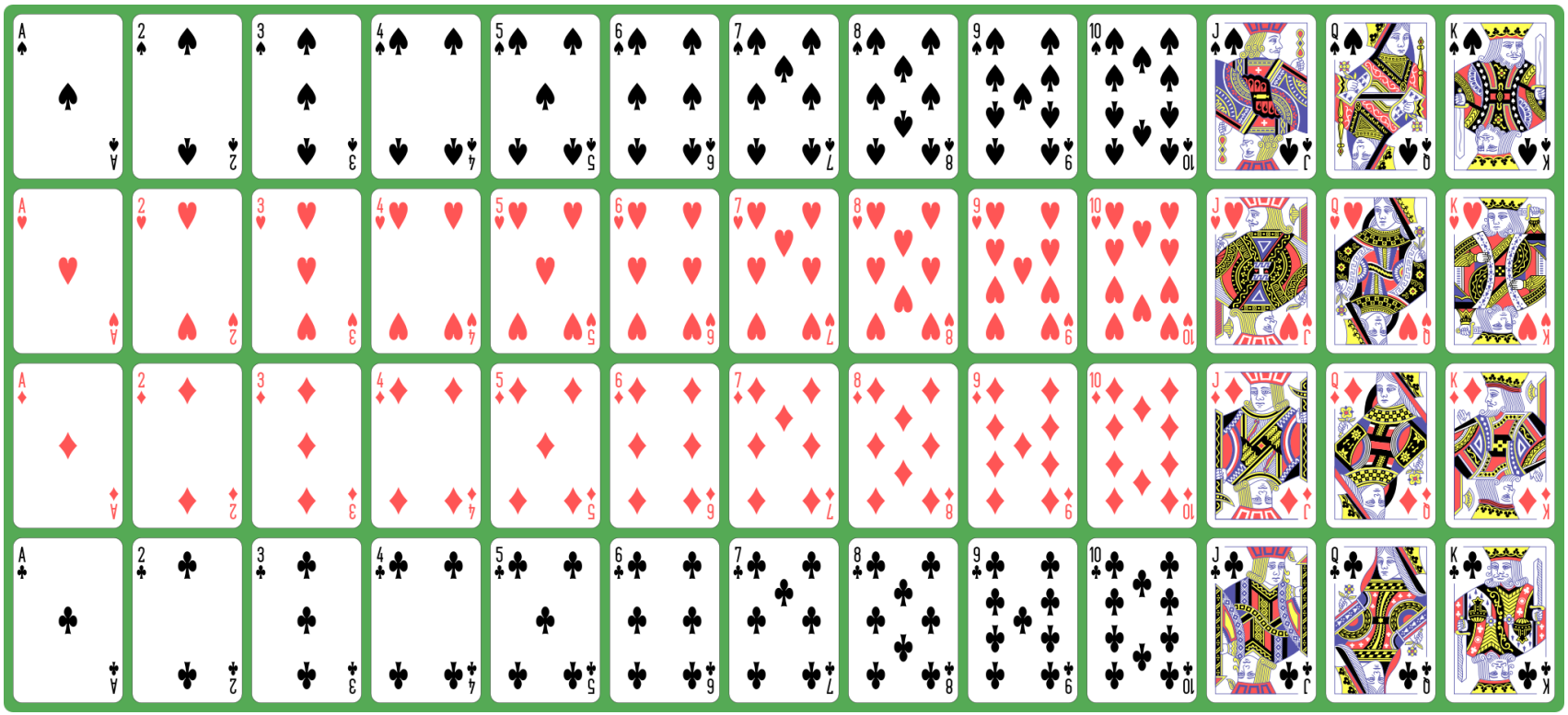
So, we compute

$$\Pr(A \mid B) = \frac{\Pr(\{XX\})}{\Pr(\{XO, OX, XX\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Conditional probabilities

Example

If you are dealt five cards from a deck of 52 cards, what is the probability of a straight flush (5 cards of the same suit with successive values)?



Picture: Dmitry Fomin (`https://commons.wikimedia.org/wiki/File:English_pattern_playing_cards_deck.svg`)

Conditional probabilities

In each suit, there are 10 possible straight flushes (starting from ace up to ten).

$$\Pr(\text{straight flush}) = \frac{4 \cdot 10}{\binom{52}{5}} \approx 1.5 \cdot 10^{-6}.$$

What is the probability of a straight flush given that we hold only cards with value ten or higher?

$$\Pr(\text{straight flush} \mid \text{only tens or higher}) = \frac{4}{\binom{20}{5}} \approx 2.6 \cdot 10^{-4}.$$

What is the probability of a straight flush given that we hold only spades?

$$\Pr(\text{straight flush} \mid \text{only spades}) = \frac{10}{\binom{13}{5}} \approx 7.8 \cdot 10^{-3}.$$

Conditional probabilities

For an event B in a discrete probability space $(\Omega, \Pr)$, we use the notation $\Pr(\cdot \mid B)$ to denote the function

$$\begin{aligned} \Pr(\cdot \mid B) \colon \Omega &\to [0,1] \\ \omega &\mapsto \Pr(\omega \mid B) := \Pr(\{\omega\} \mid B). \end{aligned}$$

Lemma

Let $(\Omega, \Pr)$ be a discrete probability space and let B be an event with $\Pr(B) > 0$. Then, $\Pr(\cdot \mid B)$ defines a probability function on Ω, i.e., $(\Omega, \Pr(\cdot \mid B))$ is a discrete probability space.

Proof.

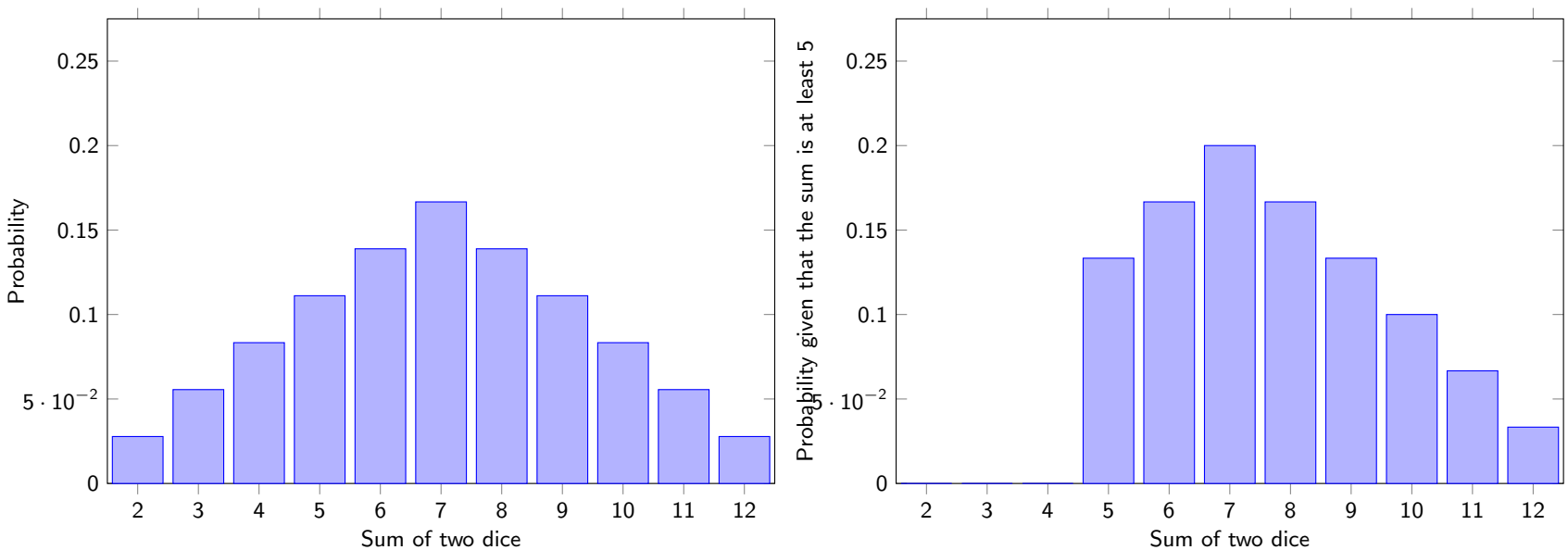
$$\begin{aligned} \sum_{\omega \in \Omega} \Pr(\omega \mid B) &= \sum_{\omega \in \Omega} \frac{\Pr(\{\omega\} \cap B)}{\Pr(B)} = \frac{1}{\Pr(B)} \sum_{\omega \in B} \Pr(\{\omega\} \cap B) \\ &= \frac{1}{\Pr(B)} \sum_{\omega \in B} \Pr(\omega) = 1. \end{aligned}$$

□

Conditional probabilities

Example

Illustration of $\Pr(\cdot \mid B)$:
Consider $\Omega = \{2, \ldots, 12\}$, the probability function Pr obtained from rolling two fair dice, and the event B that the sum is at least 5.



Conditional probabilities

Lemma

Let A and B be events with $\Pr(A) > 0$. Then,

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B \mid A).$$

Proof.

Immediate: $\Pr(B \mid A) = \Pr(A \cap B) / \Pr(A)$. □

Theorem

Let $A_1, \ldots, A_n$ be events with $\Pr(A_1 \cap \cdots \cap A_{n-1}) > 0$. Then,

$$\begin{aligned} &\Pr(A_1 \cap \cdots \cap A_n) \\ =\ &\Pr(A_1) \cdot \Pr(A_2 \mid A_1) \cdot \Pr(A_3 \mid A_1 \cap A_2) \cdot \cdots \cdot \Pr(A_n \mid A_1 \cap \cdots \cap A_{n-1}). \end{aligned}$$

Proof.

By induction using the previous lemma. □

Example (Birthday problem)

How large is the probability that two people share the same birthday in a group of m people?

Reformulation of the opposite event: If we throw m balls randomly into $n = 365$ boxes, what is the probability that no box contains two balls? Clearly, we can require that $m \leq 365$.

We throw the balls one after each other. Let A_i be the event that ball i lands in an empty box.

We are interested in the event $A = \bigcap_{i=1}^{m} A_i$. We know that

$$\Pr(A) = \Pr(A_1) \cdot \Pr(A_2 \mid A_1) \cdot \Pr(A_3 \mid A_1 \cap A_2) \cdot \dots \cdot \Pr(A_n \mid A_1 \cap \dots \cap A_{n-1}).$$

Conditional probabilities

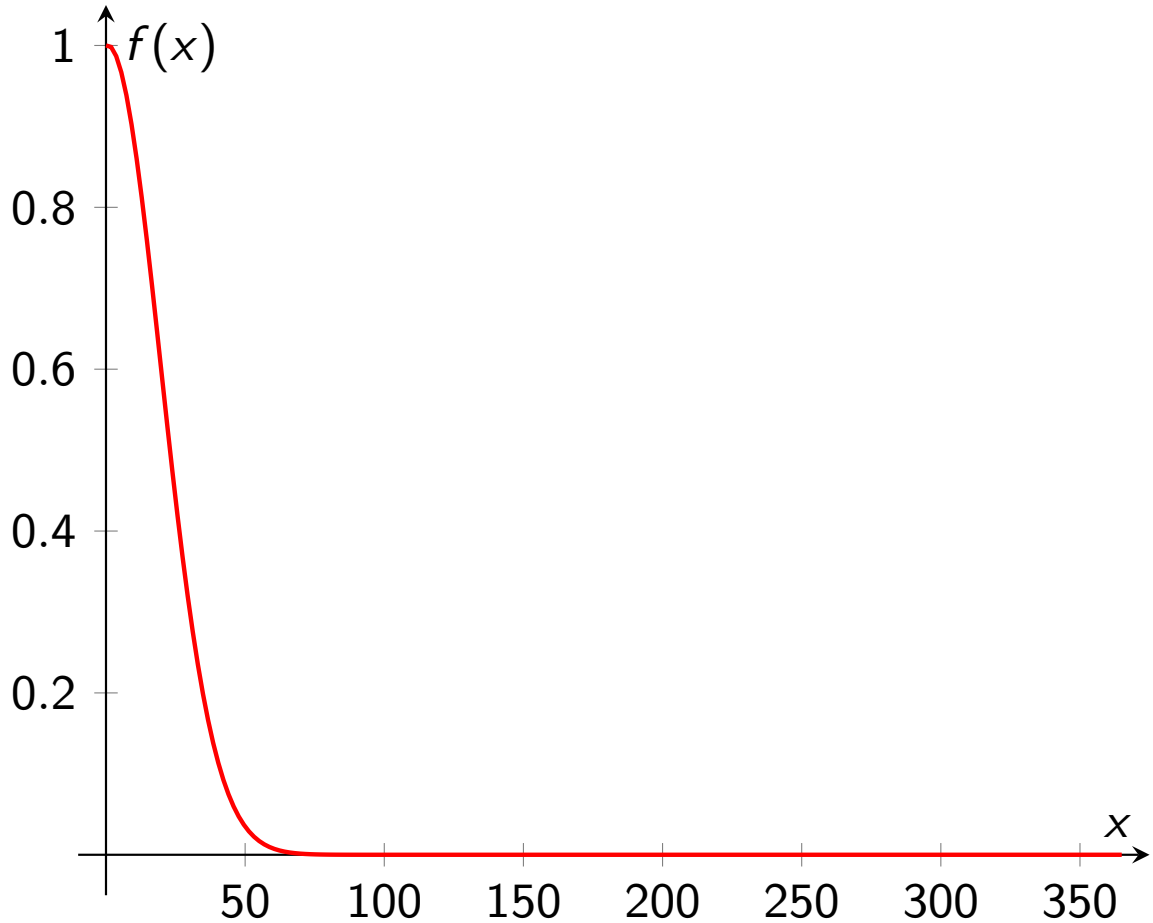
Now,

$$\Pr(A_i \mid A_1 \cap \cdots \cap A_{i-1}) = \frac{n-i+1}{n} = 1 - \frac{i-1}{n}.$$

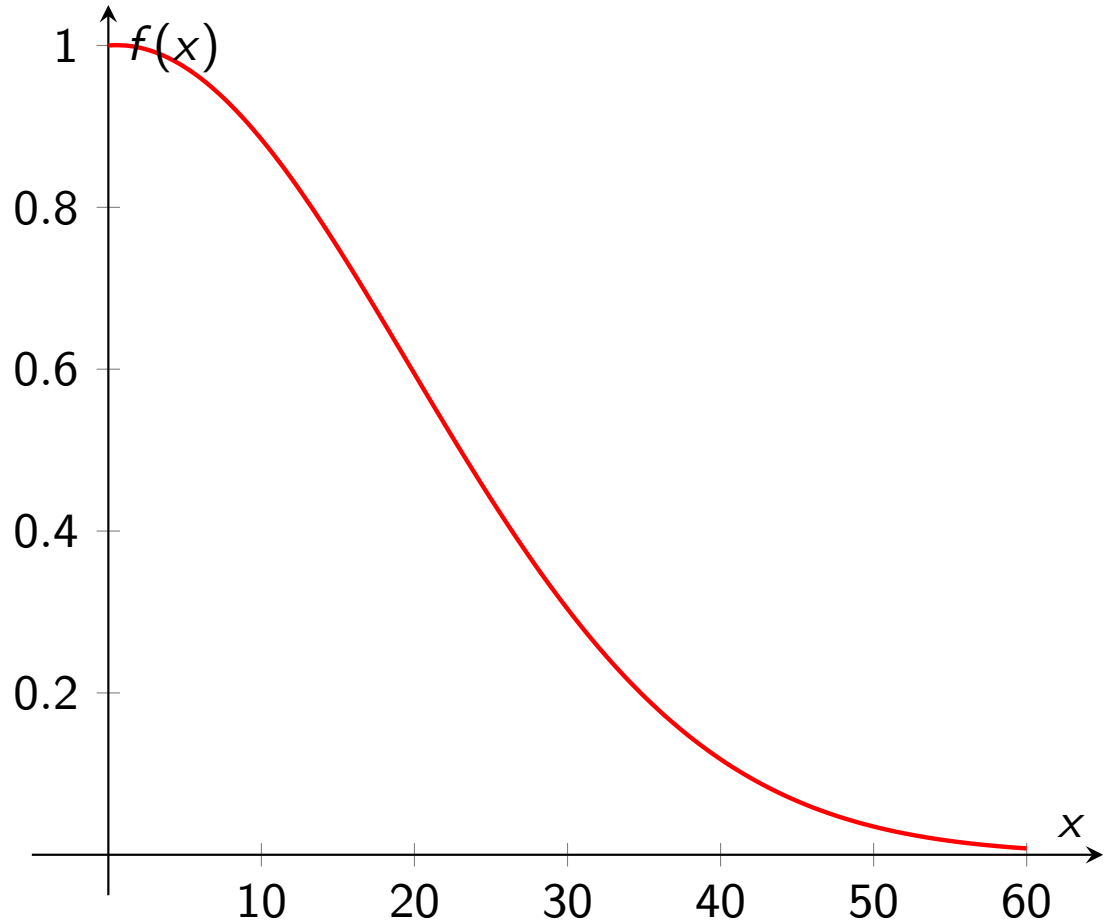
Using $1 - x \leq e^{-x}$, we obtain

$$\begin{aligned}
\Pr(A) &= \prod_{i=1}^{m} \left(1 - \frac{i-1}{n}\right) \\
&\leq \prod_{i=1}^{m} e^{-(i-1)/n} = e^{-(1/n)\sum_{j=1}^{m-1} j} \\
&= e^{-\frac{m(m-1)}{2n}} =: f(m).
\end{aligned}$$

Conditional probabilities



Conditional probabilities



Conditional probabilities

Theorem (Theorem of total probability)

Let $B_1, \dots B_k$ be events that partition the sample space (i.e., they are pairwise disjoint and $\bigcup_{i=1}^{k} B_i$ is the whole sample space) with $\Pr(B_i) > 0$ for $i = 1, \dots, k$.
For any event A,

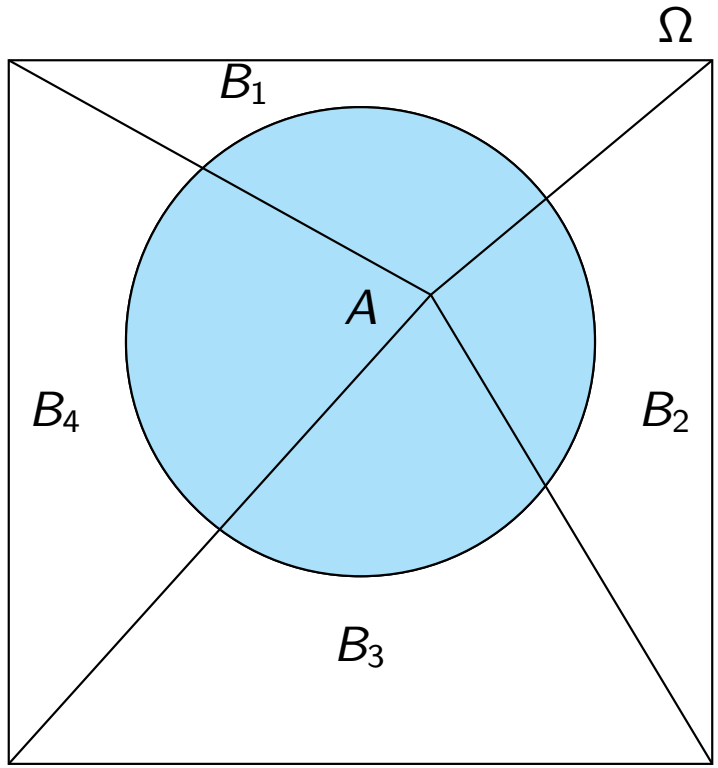
$$\Pr(A) = \sum_{i=1}^{k} \Pr(B_i) \cdot \Pr(A \mid B_i).$$

Proof.

The event A is the union of the pairwise disjoint events $A \cap B_i$ for $i = 1, \dots, k$. For each i, $\Pr(A \cap B_i) = \Pr(B_i) \cdot \Pr(A \mid B_i)$. So, indeed

$$\Pr(A) = \Pr\left(\bigcup_{i=1}^{k} (A \cap B_i)\right) = \sum_{i=1}^{k} \Pr(A \cap B_i) = \sum_{i=1}^{k} \Pr(B_i) \cdot \Pr(A \mid B_i) \qquad \square$$

Conditional probabilities



$$P(A) = \Pr(B_1) \cdot \Pr(A \mid B_1) + \Pr(B_2) \cdot \Pr(A \mid B_2) + \Pr(B_3) \cdot \Pr(A \mid B_3) + \Pr(B_4) \cdot \Pr(A \mid B_4)$$

Bayes' rule

Theorem (Bayes' rule)

Let $B_1, \dots, B_k$ be a partition of the sample space Ω such that $\Pr(B_i) \neq 0$ for $i = 1, \dots, k$. Then for any event $A \subseteq \Omega$ with $\Pr(A) \neq 0$ and any r,

$$\Pr(B_r \mid A) = \frac{\Pr(B_r)\Pr(A \mid B_r)}{\sum_{i=1}^{k} \Pr(B_i)\Pr(A \mid B_i)}.$$

Proof.

First, $\Pr(B_r)\Pr(A \mid B_r) = \Pr(B_r \cap A)$.

Second, $\sum_{i=1}^{k} \Pr(B_i)\Pr(A \mid B_i) = \Pr(A)$ by the theorem of total probability. □

Bayes' rule

Corollary

Let A and B be events with $\Pr(A), \Pr(B) > 0$. Then,

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}.$$

Example

A hit and run accident happened. A witness says that the car was black.

Let A be the event that the car was black.
Let B be the event that the witness says that the car is black.

It is known that a witness correctly identifies whether the color of a car is black or not in 70% of cases, i.e., $\Pr(B \mid A) = 0.7$ and $\Pr(B \mid A^c) = 0.3$.
Further, it is known that 10% of cars are black, i.e., $\Pr(A) = 0.1$.
We conclude $\Pr(B) = 0.07 + 0.27 = 0.34$.

What is the probability that the car was indeed black in our scenario?

$$\Pr(B \mid A) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)} = \frac{0.7 \cdot 0.1}{0.34} = \frac{7}{34} \approx 0.21.$$

Independence

Definition

If two events A and B are called *independent* if

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B).$$

Lemma

Let A and B be two events with $\Pr(B) > 0$. Then, A and B are independent if and only if

$$\Pr(A \mid B) = \Pr(A).$$

Proof.

$\frac{\Pr(A \cap B)}{\Pr(B)} = \Pr(A)$ implies $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$, and vice versa. □

Example

We draw a card from a deck of 52 cards. Let A be the event that the suit is hearts. Let B be the event that the card is a queen.

The sample space Ω contains the 52 cards and $\Pr(\omega) = 1/52$ for each $\omega \in \Omega$.

The event A contains the 13 cards of hearts, the event B the four queens. Now,

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(\{\text{queen of hearts}\})}{\Pr(B)} = \frac{1}{4} = \Pr(A).$$

So, A and B are independent.

Let C be the event that the card is red. Then,

$$\Pr(A \mid C) = \frac{\Pr(A \cap C)}{\Pr(C)} = \frac{\Pr(A)}{\Pr(C)} = \frac{1}{2} \neq \Pr(A).$$

So, A and C are not independent.

Example

We roll three dice in succession. Consider the events
A: the first roll is even.
B: the sum of all rolls is even.

Are A and B independent?

$\Pr(A) = 1/2$.

Define
C: the sum of the second and third roll is even.

$$\begin{aligned}\Pr(A \mid B) = \Pr(C) =& \Pr(C \mid \text{second roll even}) \cdot \Pr(\text{second roll even}) + \\ & \Pr(C \mid \text{second roll odd}) \cdot \Pr(\text{second roll odd}) \\ =& \Pr(\text{third roll even} \mid \text{second roll even}) \cdot \Pr(\text{second roll even}) + \\ & \Pr(\text{third roll odd} \mid \text{second roll odd}) \cdot \Pr(\text{second roll odd}) \\ =& 1/2 \cdot 1/2 + 1/2 \cdot 1/2 = 1/2.\end{aligned}$$

Independence

Definition

The events $A_1, \ldots, A_n$ are called *independent* if for all subsets $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ with $i_1 < \cdots < i_k$, we have

$$\Pr(A_{i_1} \cap \cdots \cap A_{i_k}) = \Pr(A_{i_1}) \cdot \ldots \cdot \Pr(A_{i_k}).$$

Lemma

The events $A_1, \ldots, A_n$ are independent if and only if for all $(s_1, \ldots, s_n) \in \{0,1\}^n$,

$$\Pr(A_1^{s_1} \cap \cdots \cap A_n^{s_n}) = \Pr(A_1^{s_1}) \cdot \ldots \cdot \Pr(A_n^{s_n}) \qquad (*)$$

where $A_i^1 = A_i$ and $A_i^0 = A_i^c$.

Proof.

First, we show that condition $(*)$ holds for independent events. We prove this by induction over the number of zeroes in $(s_1, \dots, s_n)$.

If there are no zeroes, there is nothing to show.

If there is at least one zero, we assume w.l.o.g. that $s_1 = 0$. In this case, we know

$$\Pr(A_1^0 \cap A_2^{s_2} \cap \dots \cap A_n^{s_n}) = \Pr(A_2^{s_2} \cap \dots \cap A_n^{s_n}) - \Pr(A_1^1 \cap A_2^{s_2} \cap \dots \cap A_n^{s_n}).$$

On the right hand side, the tuples $(s_2, \dots, s_n)$ and $(1, s_2, \dots, s_n)$ contain less zeroes than $(s_1, \dots, s_n)$. So, we can apply the induction hypothesis:

$$\begin{aligned}
&\Pr(A_1^0 \cap A_2^{s_2} \cap \dots \cap A_n^{s_n}) \\
=&\Pr(A_2^{s_2}) \cdot \dots \cdot \Pr(A_n^{s_n}) - \Pr(A_1) \cdot \Pr(A_2^{s_2}) \cdot \dots \cdot \Pr(A_n^{s_n}) \\
=&(1 - \Pr(A_1)) \cdot \Pr(A_2^{s_2}) \cdot \dots \cdot \Pr(A_n^{s_n}) \\
=&\Pr(A_1^0) \cdot \Pr(A_2^{s_2}) \cdot \dots \cdot \Pr(A_n^{s_n}).
\end{aligned}$$

Proof (continued).

For the other direction, let $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ with $i_1 < \cdots < i_k$. W.l.o.g., we show the claim for $i_1 = 1, \ldots, i_k = k$.

Now, by $(*)$ and dividing $\bigcap_{j=1}^{k} A_j$ into disjoint events,

$$\begin{aligned}
\Pr\left(\bigcap_{j=1}^{k} A_j\right) &= \sum_{s_{k+1},\ldots,s_n \in \{0,1\}} \Pr\left(\bigcap_{j=1}^{k} A_j \cap \bigcap_{j=k+1}^{n} A_j^{s_j}\right) \\
&= \sum_{s_{k+1},\ldots,s_n \in \{0,1\}} \Pr(A_1) \cdot \ldots \cdot \Pr(A_k) \cdot \Pr(A_{k+1}^{s_{k+1}}) \cdot \ldots \cdot \Pr(A_n^{s_n}) \\
&= \Pr(A_1) \cdot \ldots \cdot \Pr(A_k) \cdot \sum_{s_{k+1},\ldots,s_n \in \{0,1\}} \Pr(A_{k+1}^{s_{k+1}}) \cdot \ldots \cdot \Pr(A_n^{s_n}) \\
&= \Pr(A_1) \cdot \ldots \cdot \Pr(A_k)
\end{aligned}$$

□

Independence

From the previous lemma, we can conclude that for independent events A and B also A^c and B are independent.
(And analogously A and B^c as well as A^c and B^c are inedependent.)

Corollary

Let A, B, and C be independent events. Then, $A \cap B$ and C as well as $A \cup B$ and C are independent.

Proof.

The independence of $A \cap B$ and C follows immediately from the definition of independence.

For $A \cup B$ and C, we observe

$$\begin{aligned}
\Pr((A \cup B) \cap C) &= \Pr((A \cap C) \cup (B \cap C)) \\
&= \Pr(A \cap C) + \Pr(B \cap C) - \Pr(A \cap B \cap C) \\
&= \Pr(C) \cdot (\Pr(A) + \Pr(B) - \Pr(A \cap B)) \\
&= \Pr(C) \cdot \Pr(A \cup B)
\end{aligned}$$

□

Outline

Random Variable

Definition

Let Ω be a sample space. A *random variable* X is a function that associates a real number with each sample point, i.e.,

$$X : \Omega \to \mathbb{R}.$$

If Ω is finite or countably infinite, X is called a *discrete* random variable.

Example

Let $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ be the sample space of three coin tosses.
Let X be "the number of heads". So, X maps each sample point to a real number, e.g.

$$X(HHT) = 2 \text{ and } X(TTT) = 0.$$

Random Variable

Example

Statisticians use sampling plans to either accept or reject batches or lots of material. Suppose one of these sampling plans involves sampling independently 10 items from a lot of 100 items in which 12 are defective. Let X be the random variable defined as the number of items found defective in the sample of 10. In this case, the random variable takes on the values 0, 1, 2, . . . , 9, 10.

Random Variable

Let $(\Omega, \Pr)$ be a discrete probability space and $X : \Omega \to \mathbb{R}$ a discrete random variable.
We denote the *value range* of X by

$$W_X = X(\Omega) = \{r \in \mathbb{R} \mid \text{ there is an } \omega \in \Omega \text{ with } X(\omega) = r\}.$$

For $r \in \mathbb{R}$, we are interested in events such as

$$X^{-1}(r) = \{\omega \in \Omega \mid X(\omega) = r\}.$$

We denote the probability of this event also by

$$\Pr(X = r) = \Pr(X^{-1}(r)).$$

Similarly, we write

$$\Pr(X \leq r) = \Pr(\{\omega \in \Omega \mid X(\omega) \leq r\}).$$

Random Variable

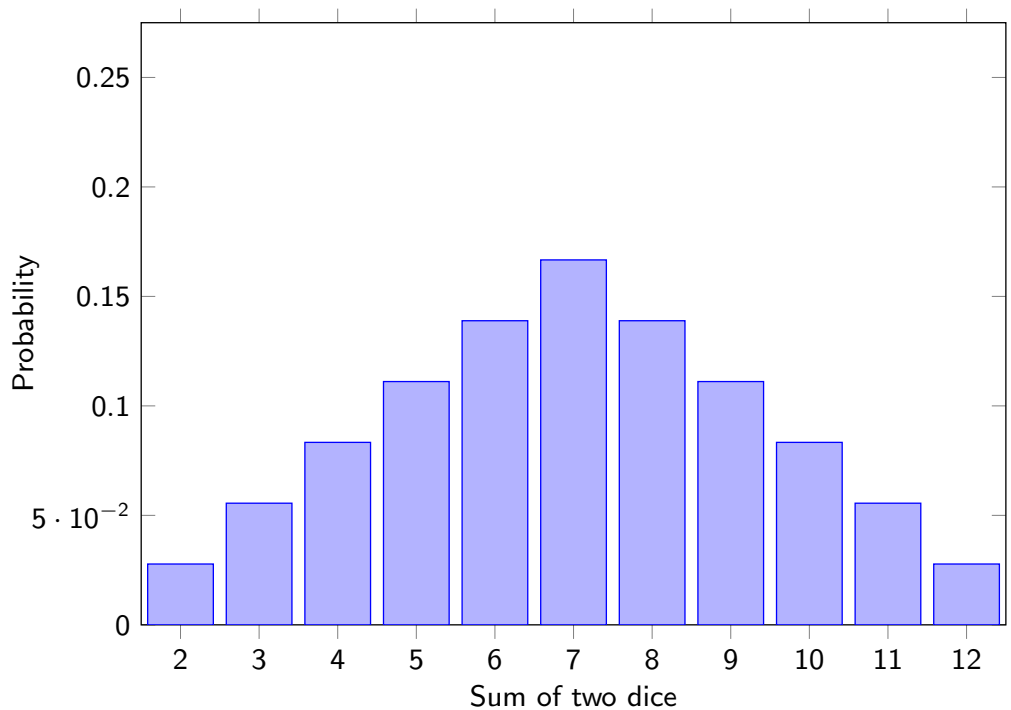
Example

Let $\Omega = \{1, \dots, 6\}^2$ with $\Pr(\omega) = 1/36$ for all $\omega \in \Omega$.

Let $X(i, j) = i + j$ for all $(i, j) \in \Omega$.

Then, $X^{-1}(4) = \{(1, 3), (2, 2), (3, 1)\}$ and $\Pr(X = 4) = \Pr(X^{-1}(4)) = 3/36$.

Likewise, $\Pr(X \leq 4) = \Pr(\{(1, 1), (1, 2), (2, 1), (1, 3), (2, 2), (3, 1)\} = 6/36$.



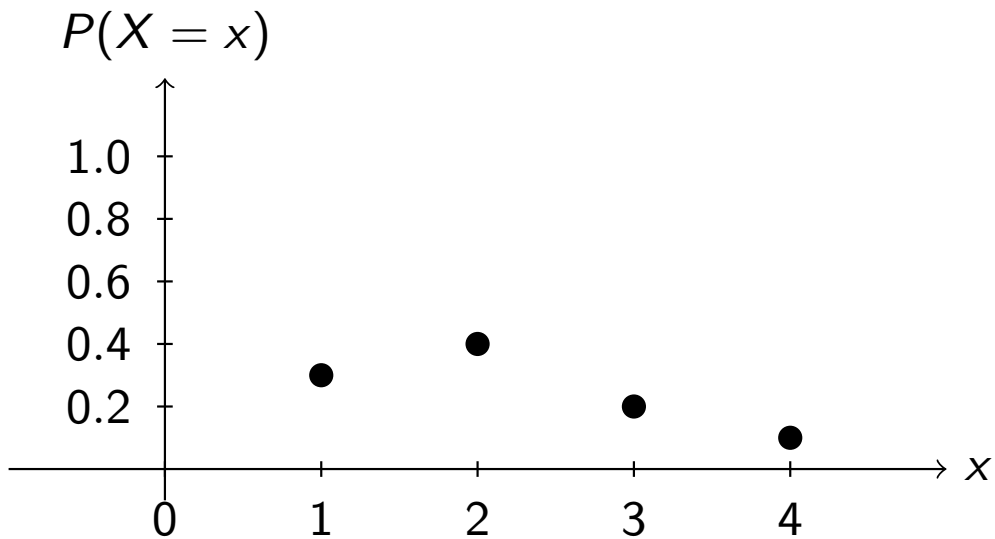
Random Variable

Definition

Given a discrete random variable X, the *probability mass function* (or *probability distribution*) of X is the function

$$f_X : \mathbb{R} \to [0, 1]$$
$$r \mapsto \Pr(X = r).$$

Example



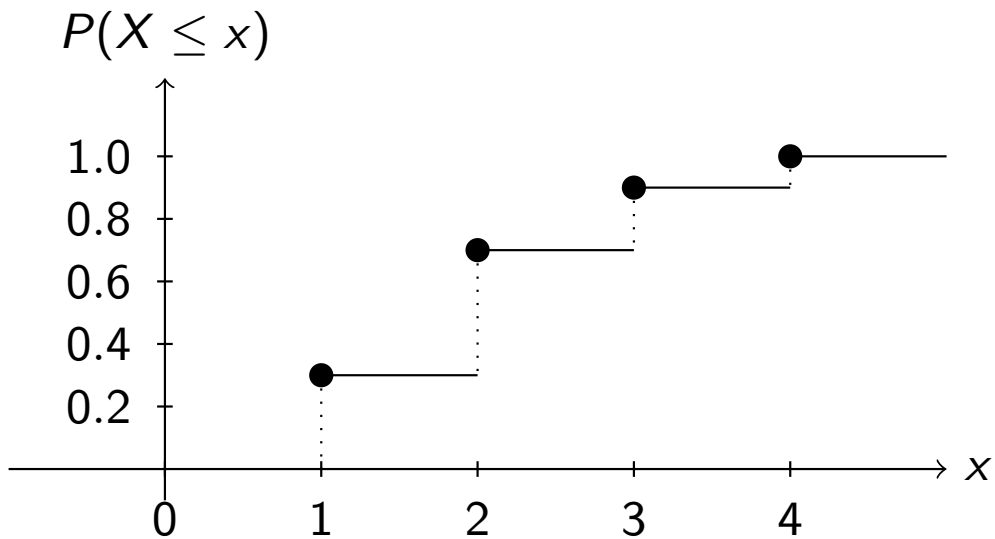
Random Variable

Definition

Given a discrete random variable X, the *cumulative distribution function* of X is the function

$$F_X : \mathbb{R} \to [0, 1]$$
$$r \mapsto \Pr(X \leq r).$$

Example



Expected Value

Definition

Let X be a discrete random variable. The *expected value* of X is

$$\mathbb{E}(X) := \sum_{x \in W_X} x \cdot \Pr(X = x) = \sum_{x \in W_X} x \cdot f_X(x)$$

if $\sum_{x \in W_X} |x| \cdot \Pr(X = x)$ converges.

Example

Let $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$ be the sample space of three coin tosses with $\Pr(\omega) = 1/8$ for all $\omega \in \Omega$ and let X be "the number of heads".

$$\mathbb{E}(X) = \sum_{x \in W_X} x \cdot \Pr(X = x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}.$$

Example

We throw a coin until "heads" comes up. Let k be the number of tosses needed. If k is odd we have to pay k EUR to the casino. If k is even, the casino pays us k EUR.
So, the sample space is $\Omega = \{T^i H \mid i \in \mathbb{N}\}$ and for each i,
$\Pr(T^i H) = (1/2)^{i+1}$.
We define

$$W(T^i H) = \begin{cases} -(i+1) & \text{if } i \text{ is even,} \\ i+1 & \text{if } i \text{ is odd.} \end{cases}$$

We obtain

$$\mathbb{E}(W) = \sum_{i=0}^{\infty} (-1)^{i+1} \cdot (i+1) \cdot (1/2)^{i+1}.$$

Expected Value

Example (continued)

Let us first check that the expected value exists. The following sum has to converge:

$$\sum_{i=0}^{\infty} |(-1)^{i+1} \cdot (i+1)| \cdot (1/2)^{i+1} = \sum_{i=0}^{\infty} (i+1) \cdot (1/2)^{i+1} = 2.$$

Now, we compute the expected value:

$$\begin{aligned}
\mathbb{E}(W) &= \sum_{i=0}^{\infty} (-1)^{i+1} \cdot (i+1) \cdot (1/2)^{i+1} \\
&= \sum_{j=0}^{\infty} -(2j+1) \cdot (1/2)^{2j+1} + (2j+2) \cdot (1/2)^{(2j+2)} \\
&= \frac{1}{2} \cdot \sum_{j=0}^{\infty} (1/4)^j \cdot (-2j - 1 + j + 1) \\
&= \frac{1}{2} \cdot \sum_{j=0}^{\infty} (1/4)^j \cdot (-j) = -\frac{2}{9}.
\end{aligned}$$

Example (continued)

To increase the risk, we change the amount to be paid from k to 2^k. So, we define

$$W'(T^i H) = \begin{cases} -2^{i+1} & \text{if } i \text{ is even,} \\ 2^{i+1} & \text{if } i \text{ is odd.} \end{cases}$$

This time,

$$\sum_{i=0}^{\infty} |(-1)^{i+1} \cdot 2^{i+1}| \cdot (1/2)^{i+1} = \sum_{i=0}^{\infty} 1$$

does not converge.

Note that if we try to compute the expected value, we get

$$\sum_{i=0}^{\infty} (-1)^{i+1} \cdot 2^{i+1} \cdot (1/2)^{i+1} = -1 + 1 - 1 + 1 - \dots$$

Expected Value

Lemma

Let X be a discrete random variable over the sample space Ω. Then,

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

Proof.

$$\mathbb{E}(X) = \sum_{r \in W_X} r \cdot \Pr(X = r) = \sum_{r \in W_X} \sum_{\substack{\omega \in \Omega, \\ X(\omega) = r}} X(\omega) \cdot \Pr(\omega) = \sum_{\omega \in \Omega} X(\omega) \cdot \Pr(\omega).$$

□

Theorem (Monotonicity of Expected Values)

Let X and Y be discrete random variables over the sample space Ω such that $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. Then,

$$\mathbb{E}(X) \leq \mathbb{E}(Y).$$

Proof.

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) \cdot \Pr(\omega) \leq \sum_{\omega \in \Omega} Y(\omega) \cdot \Pr(\omega) = \mathbb{E}(Y)$$

□

Corollary

If $a \leq X(\omega) \leq b$ for all $\omega \in \Omega$, then $a \leq \mathbb{E}(X) \leq b$.

Expected Value

Often, a function is applied to the result of a random variable.
E.g., if X is the termination time of a randomized algorithm, we might want to apply a cost function $f : \mathbb{R} \to \mathbb{R}$ to X.
We define

$$Y = f(X) = f \circ X.$$

This works for any function $f : D \to \mathbb{R}$ with $W_X \subseteq D$.
Now, Y is also a random variable.

Lemma

For X and f as above, $\mathbb{E}(f(X)) = \sum_{x \in W_X} f(x) \cdot \Pr(X = x)$.

Proof.

$$\begin{aligned}
\mathbb{E}(f(X)) &= \sum_{y \in W_{f(X)}} y \cdot \Pr(f(X) = y) \\
&= \sum_{y \in W_{f(X)}} y \cdot \sum_{x : f(x) = y} \Pr(X = x) \\
&= \sum_{x \in W_X} f(x) \cdot \Pr(X = x).
\end{aligned}$$

□

Expected Value

Theorem (Linearity of expected values, simple version)

Let X be a discrete random variable and $a, b \in \mathbb{R}$. Then,

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

Proof.

$$\begin{aligned}
\mathbb{E}(a \cdot X + b) &= \sum_{x \in W_x} (a \cdot x + b) \cdot \Pr(X = x) \\
&= a \cdot \sum_{x \in W_X} x \cdot \Pr(X = x) + b \cdot \sum_{x \in W_X} \Pr(X = x) \\
&= a \cdot \mathbb{E}(X) + b.
\end{aligned}$$

□

Expected Value

Theorem

Let X be a discrete random variable with $W_X \subseteq \mathbb{N}$. Then,

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} \Pr(X \geq i).$$

Proof.

$$\begin{aligned}\mathbb{E}(X) &= \sum_{i=0}^{\infty} i \cdot \Pr(X = i) = \sum_{i=0}^{\infty} \sum_{j=1}^{i} \Pr(X = i) \\ &= \sum_{1 \leq j \leq i} \Pr(X = i) = \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} \Pr(X = i) = \sum_{j=1}^{\infty} \Pr(X \geq j).\end{aligned}$$

□

Definition (Conditional random variable)

Let X be a random variable and A an event with $\Pr(A) > 0$. The *conditional random variable* $X|A$ is given by the probability mass function

$$f_{X|A}(x) := \Pr(X = x \mid A) = \frac{\Pr(X^{-1}(x) \cap A)}{\Pr(A)}.$$

This is a valid definition as

$$\sum_{x \in W_X} f_{X|A}(x) = \sum_{x \in W_X} \frac{\Pr(X^{-1}(x) \cap A)}{\Pr(A)} = \frac{\sum_{x \in W_X} \Pr(X^{-1}(x) \cap A)}{\Pr(A)} = 1.$$

The expected value of $X|A$ is given by

$$\mathbb{E}(X|A) = \sum_{x \in W_X} x \cdot f_{X|A}(x).$$

Expected Value

Theorem

Let X be a random variable over $(\Omega, \Pr)$ and let $A_1, \ldots, A_n$ be a partition of Ω. Then,

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X|A_i) \cdot \Pr(A_i).$$

Proof.

$$\mathbb{E}(X) = \sum_{x \in W_X} x \cdot \Pr(X = x) = \sum_{x \in W_X} x \cdot \sum_{i=1}^{n} \Pr(X = x \mid A_i) \cdot \Pr(A_i)$$

$$= \sum_{i=1}^{n} \Pr(A_i) \cdot \sum_{x \in W_X} x \cdot \Pr(X = x \mid A_i) = \sum_{i=1}^{n} \Pr(A_i) \cdot \mathbb{E}(X \mid A_i). \qquad \square$$

Example

Toss a coin until "heads" comes up and denote the number of tosses by the random variable X. Let's say that the tosses are independent and the chance of "heads" is p in each toss. For $k \geq 1$,

$$\Pr(X = k) = p(1-p)^{k-1}.$$

So,

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k \cdot p \cdot (1-p)^{k-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

Example (continued)

Let H_1 be the event that the first toss results in "heads". Clearly,

$$\mathbb{E}(X|H_1) = 1.$$

Now assume that the first toss is not "heads" and restart the experiment. Let X' be the number of tosses until "heads" comes up after the restart. So, $\mathbb{E}(X') = \mathbb{E}(X)$ and

$$\mathbb{E}(X|H_1^c) = 1 + \mathbb{E}(X') = 1 + \mathbb{E}(X).$$

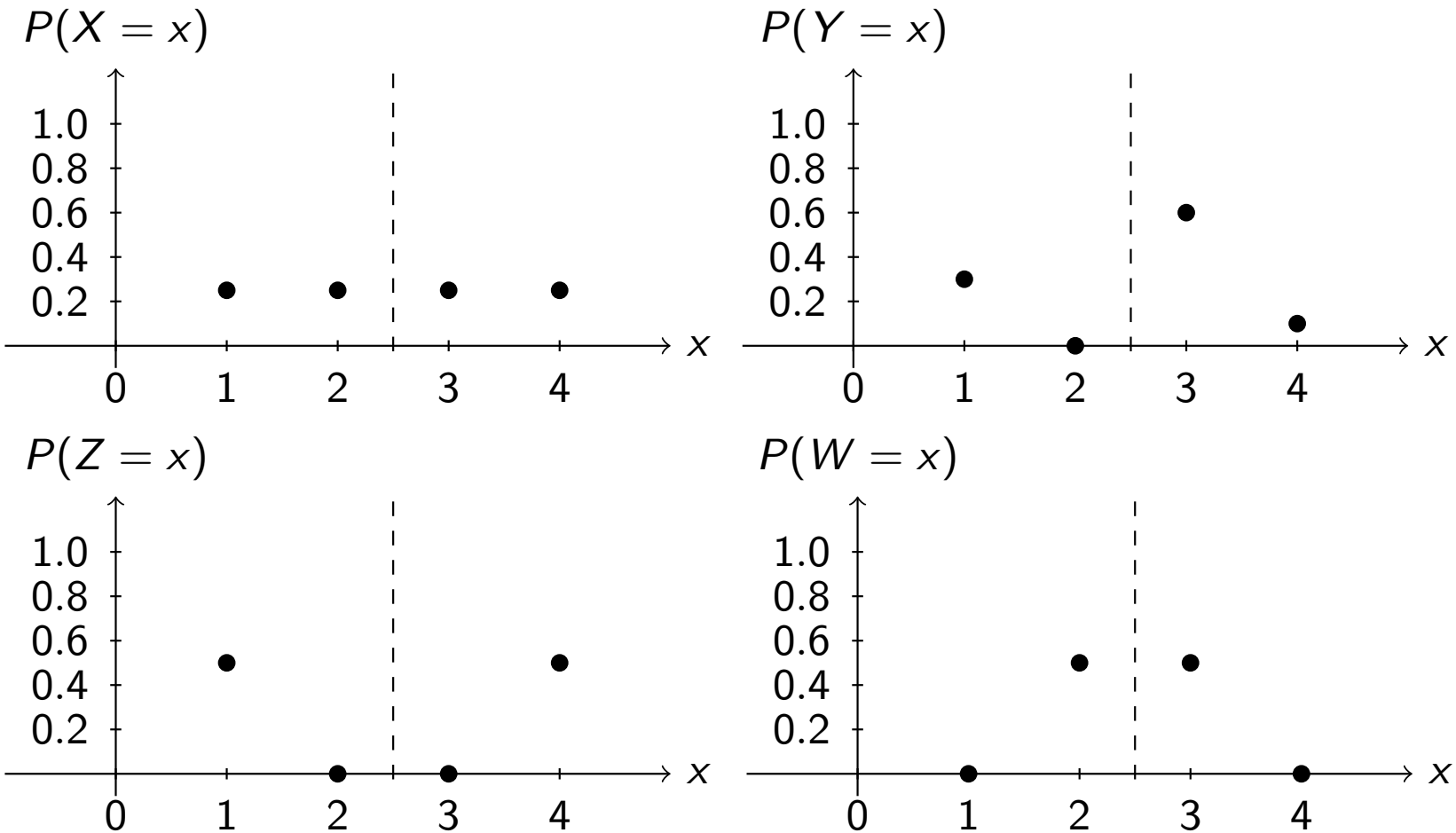
By the previous theorem,

$$\mathbb{E}(X) = \mathbb{E}(X|H_1) \cdot \Pr(H_1) + \mathbb{E}(X|H_1^c) \cdot \Pr(H_1^c) = 1 \cdot p + (1 + \mathbb{E}(X)) \cdot (1 - p).$$

We can solve now for $\mathbb{E}(X)$ and obtain $\mathbb{E}(X) = 1/p$.

Variance

Probability mass functions of two random variables with the same expected value might look quite different.



Variance

The deviations from the expected value can vary a lot.

If you want to invest 100 EUR, which of the following options would you prefer? After one year, you get

1. 100 EUR back with probability 0.5 and 104 EUR back with probability 0.5.
2. 50 EUR back with probability 0.5 and 154 EUR back with probability 0.5.
3. 0 EUR back with probability 0.5 and 205 EUR back with probability 0.5.

In option 1. and 2., the expected return is 102 EUR.
In option 3., the expected return is 102.5 EUR.

But is option 3. better than the other options?

Variance

Definition (Variance)

Let X be a random variable with expected value $\mu := \mathbb{E}(X)$. We define the *variance* of X as

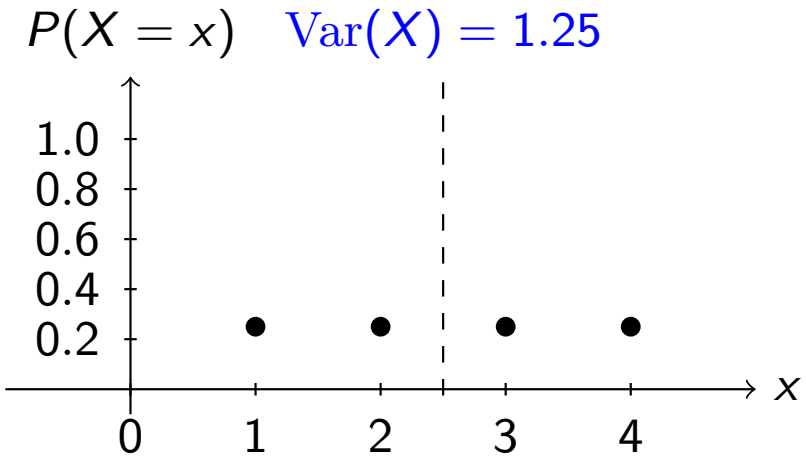
$$\operatorname{Var}(X) := \mathbb{E}((X - \mu)^2) = \sum_{x \in W_X} (x - \mu)^2 \cdot \Pr(X = x).$$

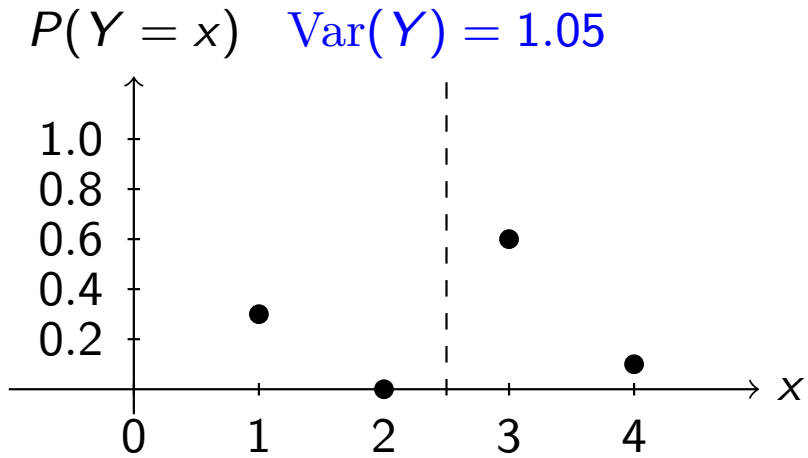
The value

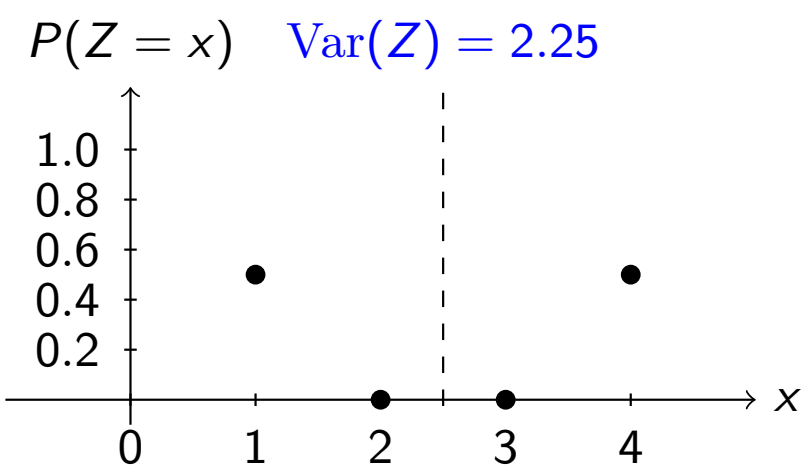
$$\sigma(X) := \sqrt{\operatorname{Var}(X)}$$

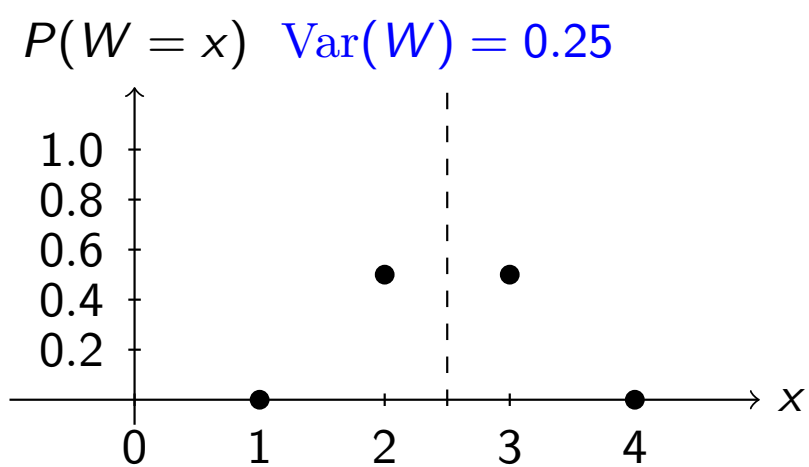
is called the *standard deviation* of X.

Variance


$P(X = x)$ $\mathrm{Var}(X) = 1.25$
1.0
0.8
0.6
0.4
0.2
0
1
2
3
4
x


$P(Y = x)$ $\mathrm{Var}(Y) = 1.05$
1.0
0.8
0.6
0.4
0.2
0
1
2
3
4
x


$P(Z = x)$ $\mathrm{Var}(Z) = 2.25$
1.0
0.8
0.6
0.4
0.2
0
1
2
3
4
x


$P(W = x)$ $\mathrm{Var}(W) = 0.25$
1.0
0.8
0.6
0.4
0.2
0
1
2
3
4
x

Variance

Theorem

For a random variable X,

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

Proof.

Let $\mu := \mathbb{E}(X)$. Then,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2 - 2\mu X + \mu^2) \\
&= \sum_{x \in W_X} (x^2 - 2\mu x + \mu^2) \cdot \mathrm{Pr}(X = x) \\
&= \sum_{x \in W_X} x^2 \cdot \mathrm{Pr}(X = x) - 2\mu x \cdot \mathrm{Pr}(X = x) + \mu^2 \cdot \mathrm{Pr}(X = x) \\
&= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \\
&= \mathbb{E}(X^2) - \mathbb{E}(X)^2.
\end{aligned}$$

□

Variance

Theorem

For a random variable X and $a, b \in \mathbb{R}$, we have

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

Proof.

We know that the expected value is linear and hence, $\mathbb{E}(Y + b) = \mathbb{E}(Y) + b$ for any random variable Y.
We obtain

$$\mathrm{Var}(Y + b) = \mathbb{E}((Y + b - \mathbb{E}(Y + b))^2) = \mathbb{E}((Y - \mathbb{E}(Y))^2) = \mathrm{Var}(Y).$$

By the previous theorem,

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

Put together, this shows the claim. □

Variance

Variance and expected value are part of the *moments* of a random variable:

Definition (Moments)

For a random variable X, the value $\mathbb{E}(X^k)$ is called the kth moment of X and the value $\mathbb{E}((X - \mathbb{E}(X))^k)$ is called the kth central moment of X.

So, the expected value is the first moment while the variance is the second central moment.

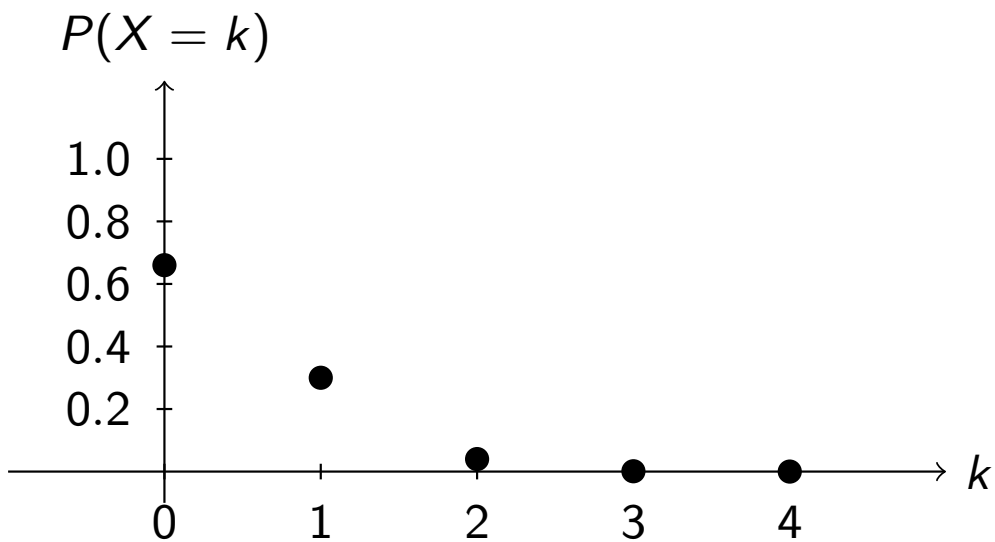
Joint probability distribution

We may consider several random variables on one discrete probability space.

Example

Two players are dealt five cards from a poker deck with 52 cards. Let X be the number of aces the first player holds and let Y be the number of aces the second player holds.
We can compute for $k = 0, \ldots, 4$ that

$$\Pr(X = k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}.$$



Joint probability distribution

For $\Pr(Y = k)$, we get the same distribution. However, clearly the value of X has an influence on the value of Y. E.g., $X = 3$ and $Y = 2$ is impossible.

We will be interested in probabilities like

$$\Pr(X = 3, Y = 1) = \Pr(\{\omega \mid X(\omega) = 3, Y(\omega) = 1\}).$$

Remark:
A probability distribution of this form $\Pr(X = k) = \frac{\binom{4}{k}\binom{48}{5-k}}{\binom{52}{5}}$ is also called *hyper-geometric distribution* with parameters $a = 48$, $b = 4$, and $r = 5$.

Joint probability distribution

Definition

Given two random variables X and Y on a discrete probability space $(\Omega, \Pr)$, the *joint probability mass function* (or *joint probability distribution*) of X and Y is the function

$$f_{X,Y}(x, y) := \Pr(X = x, Y = y).$$

The probability mass function f_X can be expressed as

$$f_X(x) = \sum_{y \in W_Y} f_{X,Y}(x, y)$$

and analogously for Y.
In this context, this probability mass function is also called the *marginal probability mass function* of X.

Joint probability distribution

Definition

Given two random variables X and Y on a discrete probability space $(\Omega, \Pr)$, the *joint cumulative distribution function* of X and Y is the function

$$F_{X,Y}(a, b) := \Pr(X \leq a, Y \leq b).$$

The cumulative distribution function of X can be obtained from the joint cumulative distribution function by

$$F_X(a) = \lim_{b \to \infty} F_{X,Y}(a, b)$$

and analogously for Y.

Joint probability distribution

Definition

The random variables $X_1, \ldots, X_n$ are called *independent* if, for all $(x_1, \ldots, x_n) \in W_{X_1} \times \cdots \times W_{X_n}$, we have

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_1 = x_1) \cdot \Pr(X_2 = x_2) \cdot \ldots \cdot \Pr(X_n = x_n).$$

This can be rephrased as

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n).$$

So, the joint probability mass function is the product of the respective marginal probability mass functions.

Joint probability distribution

Lemma

For independent random variables $X_1, \ldots, X_n$, for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, we have

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n).$$

Proof.

$$\begin{aligned}
F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= \Pr(X_1 \leq x_1, \ldots, X_n \leq x_n) \\
&= \sum_{\substack{y_1 \in W_{X_1} \\ y_1 \leq x_1}} \cdots \sum_{\substack{y_n \in W_{X_n} \\ y_n \leq x_n}} \Pr(X_1 = y_1, \ldots, X_n = y_n) \\
&= \sum_{\substack{y_1 \in W_{X_1} \\ y_1 \leq x_1}} \cdots \sum_{\substack{y_n \in W_{X_n} \\ y_n \leq x_n}} \Pr(X_1 = y_1) \cdot \ldots \cdot \Pr(X_n = y_n) \\
&= \left(\sum_{\substack{y_1 \in W_{X_1} \\ y_1 \leq x_1}} \Pr(X_1 = y_1) \right) \cdot \ldots \cdot \left(\sum_{\substack{y_n \in W_{X_n} \\ y_n \leq x_n}} \Pr(X_n = y_n) \right) \\
&= F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n).
\end{aligned}$$

□

Joint probability distribution

Theorem

Let $X_1, \ldots, X_n$ be independent random variables and let $S_1, \ldots, S_n \subseteq \mathbb{R}$. Then, the events "$X_1 \in S_1$", ..., "$X_n \in S_n$" are independent.

Proof.

Analogous to the proof on the previous slide. □

Joint probability distribution

Theorem

Let $f_1, \ldots, f_n$ be functions from $\mathbb{R}$ to $\mathbb{R}$. If the random variables $X_1, \ldots, X_n$ are independent, then so are the random variables $f_1(X_1), \ldots, f_n(X_n)$.

Proof.

For $i = 1, \ldots, n$, let $z_i \in W_{f_i(X_i)}$ and $S_i = \{x \mid f_i(x) = z_i\} = f_i^{-1}(z_i)$.
Now,

$$\begin{aligned}
&\Pr(f_1(X_1) = z_1, \ldots, f_n(X_n) = z_n) \\
=&\Pr(X_1 \in S_1, \ldots, X_n \in S_n) \\
=&\Pr(X_1 \in S_1) \cdot \ldots \cdot \Pr(X_n \in S_n) \\
=&\Pr(f_1(X_1) = z_1) \cdot \ldots \cdot \Pr(f_n(X_n) = z_n).
\end{aligned}$$

□

Composite random variables

Theorem

Let X and Y be independent random variables. Let $Z := X + Y$. Then,

$$f_Z(z) = \sum_{x \in W_X} f_X(x) \cdot f_Y(z - x).$$

Proof.

By the theorem of total probability, we have

$$\begin{aligned} f_Z(z) = \Pr(Z = z) &= \sum_{x \in W_x} \Pr(X + Y = z \mid X + x) \cdot \Pr(X = x) \\ &= \sum_{x \in W_X} \Pr(Y = z - x) \cdot \Pr(X = x) \\ &= \sum_{x \in W_X} f_X(x) \cdot f_Y(z - x). \end{aligned}$$

□

The expression

$$\sum_{x \in W_X} f_X(x) \cdot f_Y(z - x)$$

is also called the *convolution* of the probability mass functions f_X and f_Y.

Composite random variables

Example

Let X and Y be the results of the rolls of two independent dice. So, $f_X(k) = f_Y(k) = 1/6$ for $k \in \{1, \ldots, 6\}$.

Let $Z = X + Y$. Then,

$$\begin{aligned}\Pr(Z = 4) &= \sum_{x \in \{1,\ldots,6\}} f_X(k) \cdot f_Y(4-k) \\ &= f_X(1) \cdot f_Y(3) + f_X(2) \cdot f_Y(2) + f_X(3) \cdot f_Y(1) \\ &= 3/36.\end{aligned}$$

Composite random variables

Theorem (Linearity of expected values, full version)

Let $X_1, \ldots, X_n$ be random variables and let $a_1, \ldots, a_n \in \mathbb{R}$. Define $X := a_1 X_1 + \cdots + a_n X_n$. We have

$$\mathbb{E}(X) = a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n).$$

Proof.

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{\omega \in \Omega} (a_1 X_1(\omega) + \ldots a_n X_n(\omega)) \cdot \Pr(\omega) \\
&= a_1 \cdot \sum_{\omega \in \Omega} X_1(\omega) \cdot \Pr(\omega) + \cdots + a_n \cdot \sum_{\omega \in \Omega} X_n(\omega) \cdot \Pr(\omega) \\
&= a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n).
\end{aligned}$$

□

Composite random variables

Example

On a ship, n drunk sailors go to sleep. As they lost their orientation, it is completely random which sailor chooses which of the n beds. I.e., each outcome with exactly one sailor per bed is equally likely. How many sailors will lie in their own bed on average?

Let X be the number of sailors in the correct bed. We have $X = X_1 + \ldots + X_n$ where

$$X_i = \begin{cases} 1 & \text{if sailor } i \text{ lies in the correct bed,} \\ 0 & \text{otherwise.} \end{cases}$$

We observe that $\Pr(X_1 = 1) = \frac{1}{n}$ as every sailor is equally likely to end up in each of the beds. Hence,

$$\mathbb{E}(X_i) = 0 \cdot \Pr(X_i = 0) + 1 \cdot \Pr(X_i = 1) = \frac{1}{n}.$$

We conclude

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = n \cdot \frac{1}{n} = 1.$$

Theorem (Multiplicativity of expected values)

For **independent** *random variables* $X_1, \ldots, X_n$*, we have*

$$\mathbb{E}(X_1 \cdot \ldots \cdot X_n) = \mathbb{E}(X_1) \cdot \ldots \cdot \mathbb{E}(X_n).$$

Proof.

We prove the statement in the case $n = 2$. The general case can be shown analogously.

$$\begin{aligned}
\mathbb{E}(X \cdot Y) &= \sum_{x \in W_X} \sum_{y \in W_Y} x \cdot y \cdot \Pr(X = x, Y = y) \\
&\overset{\text{ind.}}{=} \sum_{x \in W_X} \sum_{y \in W_Y} x \cdot y \cdot \Pr(X = x) \cdot \Pr(Y = y) \\
&= \sum_{x \in W_X} x \cdot \Pr(X = x) \sum_{y \in W_Y} y \cdot \Pr(Y = y) \\
&= \mathbb{E}(X) \cdot \mathbb{E}(Y).
\end{aligned}$$

□

Composite random variables

Theorem

For independent random variables $X_1, \dots, X_n$ and $X := X_1 + \cdots + X_n$,

$$\operatorname{Var}(X) = \operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n).$$

Proof.

We prove the claim for $n = 2$. The general case follows easily.

$$\mathbb{E}((X+Y)^2) = \mathbb{E}(X^2 + 2XY + Y^2) = \mathbb{E}(X^2) + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y^2)$$
$$\mathbb{E}(X+Y)^2 = (\mathbb{E}(X) + \mathbb{E}(Y))^2 = \mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2.$$

Subtracting the second equation from the first, we obtain

$$\mathbb{E}((X+Y)^2) - \mathbb{E}(X+Y)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 + \mathbb{E}(Y^2) - \mathbb{E}(Y)^2.$$

This finishes the proof.

□

Composite random variables

Example

If random variables are not independent, the previous theorem does not hold in general.

E.g., let $X = -Y$ and assume $\mathrm{Var}(X) > 0$. Then,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(0) = 0 \neq \mathrm{Var}(X) + \mathrm{Var}(Y) = 2\mathrm{Var}(X).$$

Outline

Introduction

We will take a look at some important discrete probability distributions. These typically depend on some *parameters* such as the number of trials and the probability that a single trial is successful.

So, we will look at *families* of discrete probability distributions

Bernoulli distribution

Definition

A random variable X with $W_X = \{0, 1\}$ and the probability mass function

$$f_X(x) = \begin{cases} p & \text{for } x = 1, \\ 1 - p & \text{for } x = 0 \end{cases}$$

is called *Bernoulli-distributed*.
The parameter p is called the *success probability*.

Theorem

For a Bernoulli-distributed random variable X with success probability p,

$$\mathbb{E}(X) = p \text{ and } \operatorname{Var}(X) = p(1 - p).$$

Proof.

Clearly, $\mathbb{E}(X) = p$ and hence $\mathbb{E}(X)^2 = p^2$. Furthermore, $\mathbb{E}(X^2) = p$ and so

$$\operatorname{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = p(1 - p). \qquad \square$$

Binomial distribution

Often, we iterated a Bernoulli-distributed trial n times. If we count the number of successful outcomes, we obtain a binomial random variable.

So, a binomial random variable counts, e.g., the number of heads if we toss a coin n times.

Definition

A random variable X is called *binomial random variable* with parameters n and p, if

$$X := X_1 + \cdots + X_n$$

is the sum of n independent Bernoulli-distributed random variables $X_1, \ldots, X_n$ with success probability p.

In this case, we also write

$$X \sim \mathrm{Bin}(n, p).$$

Binomial distribution

Definition

For $X \sim \mathrm{Bin}(n, p)$, we have $W_X = \{0, \ldots, n\}$ and the probability mass function of X is given by

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

We define $b(x; n, p) := f_X(x)$.

Theorem

For $X \sim \mathrm{Bin}(n, p)$,

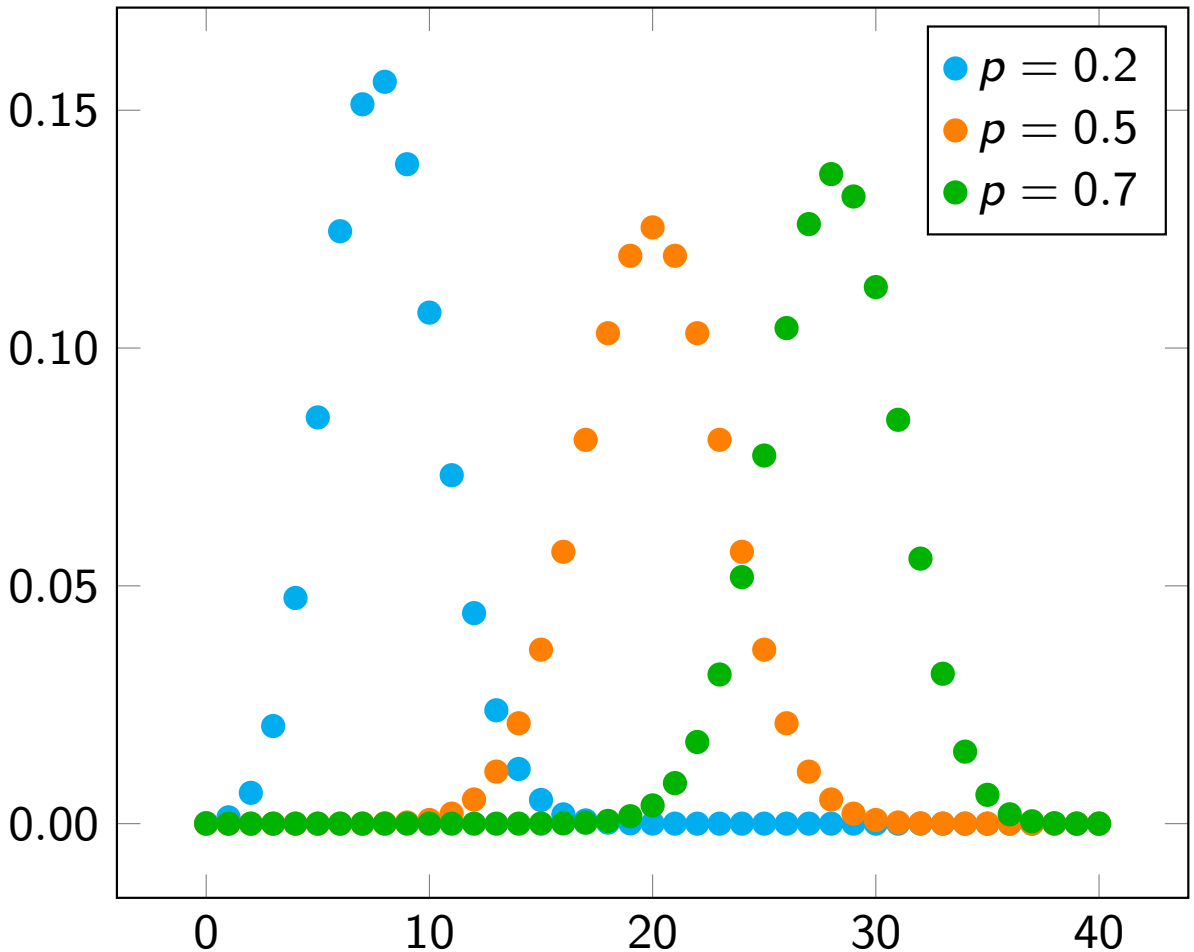
$$\mathbb{E}(X) = np \text{ and } \mathrm{Var}(X) = np(1-p).$$

Proof.

Follows immediately from the results for sums of independent random variables and the fact that $X = X_1 + \ldots X_n$ for independent Bernoulli-distributed random variables $X_1, \ldots, X_n$ with success probability p. □

Binomial distribution

Plot of the probability mass function of $X \sim \mathrm{Bin}(40, p)$:



Binomial distribution

Theorem

If $X \sim \mathrm{Bin}(n_X, p)$ *and* $Y \sim \mathrm{Bin}(n_Y, p)$ *are independent, then*

$$X + Y \sim \mathrm{Bin}(n_X + n_Y, p).$$

Proof.

We know that we can represent X as sum of n_X independent Bernoulli-distributed random variables with success probability p and, likewise, Y as sum of n_Y independent such random variables.
Then, $X + Y$ is the sum of $n_X + n_Y$ such independent random variables. □

Geometric distribution

Suppose that independent trials, each having a probability p, $0 < p < 1$, of being a success, are performed until a success occurs. Let X be the number of trials until the first success.

Definition

A *geometric random variable* X with success probability p has the probability mass function

$$f_X(i) = \Pr(X = i) = (1-p)^{i-1}p \quad \text{for } i \in \mathbb{N}, i \geq 1.$$

We check

$$\sum_{i=1}^{\infty} f_X(i) = \sum_{i=1}^{\infty}(1-p)^{i-1}p = p \cdot \sum_{j=0}^{\infty}(1-p)^j = p\frac{1}{1-(1-p)} = 1.$$

Geometric distribution

Theorem

For a geometric random variable X with success probability p,

$$\mathbb{E}(X) = \frac{1}{p} \quad \textit{and} \quad \operatorname{Var}(X) = \frac{1-p}{p^2}.$$

Geometric distribution

Proof.

First,

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1} p = p \cdot \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1} = p \cdot \frac{1}{(1-(1-p))^2} = \frac{p}{p^2} = \frac{1}{p}.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{i=1}^{\infty} i^2 \cdot (1-p)^{i-1} p \\
&= p \cdot \left(\sum_{i=1}^{\infty} i(i+1)(1-p)^{i-1} - \sum_{i=1}^{\infty} i \cdot (1-p)^{i-1} \right) \\
&= p \cdot \left(\frac{2}{p^3} - \frac{1}{p^2} \right) = \frac{2-p}{p^2}.
\end{aligned}$$

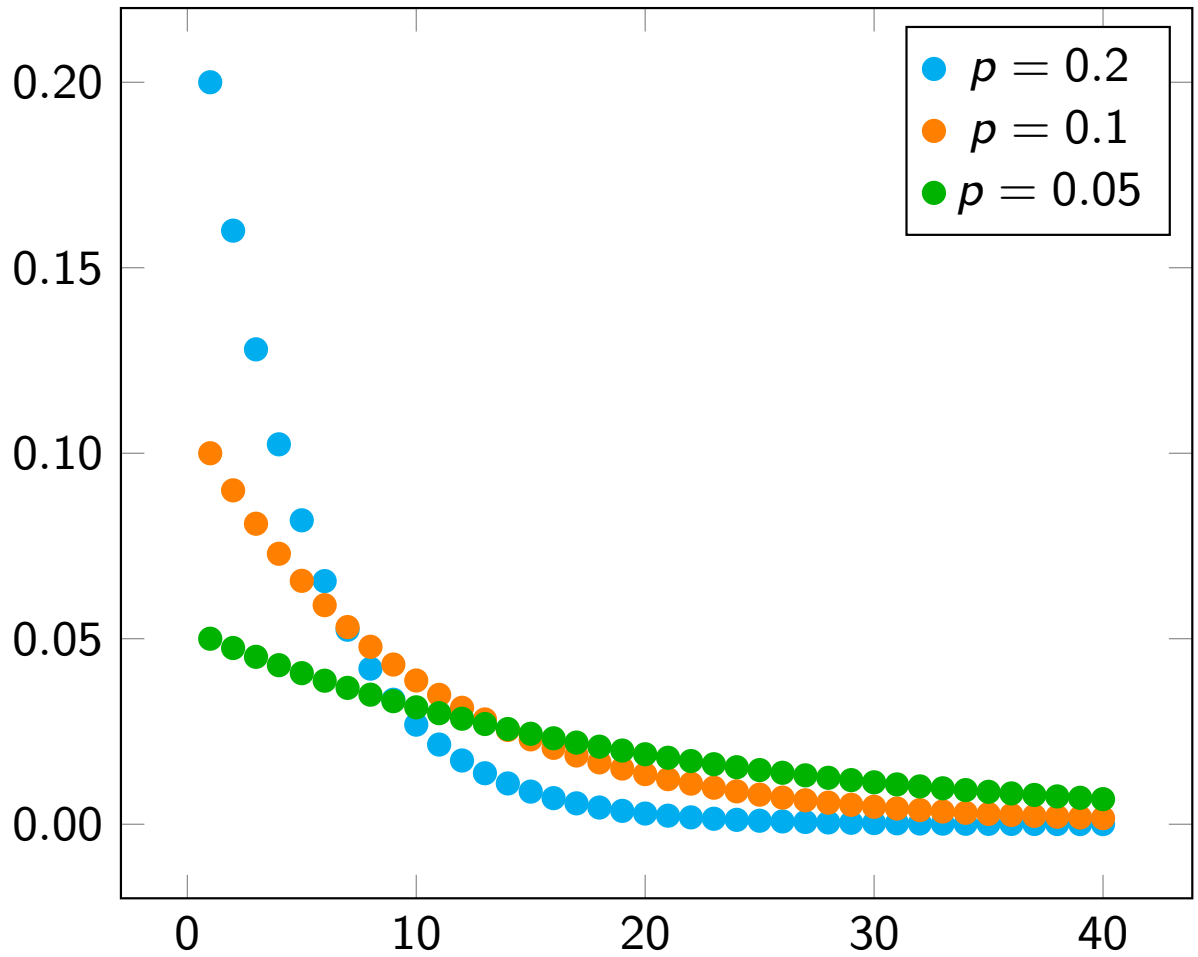
So,

$$\operatorname{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

□

Geometric distribution

Plot of the probability mass function of geometric random variables with different parameters p:



Geometric distribution

Let X be a geometric random variable with success probability p.

For $x, y \in \mathbb{N}$, what is

$$\Pr(X > x + y \mid X > x)?$$

The event "$X > x$" means that the first x trials were not successful.
Let's consider the trials starting with the $x + 1$st trial as a new experiment and let X' be the number of trials until the first success in this new experiment.
Then X' is also a geometric random variable with success probability p.
For $X > x + y$, given that we know that $X > x$, we need that $X' > y$. So,

$$\Pr(X > x + y \mid X > x) = \Pr(X' > y).$$

Geometric distribution

Let us prove this formally.

Theorem

Let X be a geometric random variable with success probability p. Then for $x, y \in \mathbb{N}$,

$$\Pr(X > x + y \mid X > x) = \Pr(X > y).$$

Proof.

$$\Pr(X > x) = \sum_{i=x+1}^{\infty} (1-p)^{i-1} p = (1-p)^x p \cdot \sum_{j=0}^{\infty} (1-p)^j$$

$$= (1-p)^x p \cdot \frac{1}{1-(1-p)} = (1-p)^x.$$

Using this, we obtain

$$\Pr(X > x + y \mid X > x) = \frac{\Pr(X > x + y, X > x)}{\Pr(X > x)} = \frac{\Pr(X > x + y)}{\Pr(X > x)}$$

$$= \frac{(1-p)^{x+y}}{(1-p)^x} = (1-p)^y = \Pr(X > y). \qquad \square$$

Geometric distribution

Due to the result that

$$\Pr(X > x + y \mid X > x) = \Pr(X > y),$$

geometric random variables are called *memoryless*. After x unsuccessful trials, the random variable behaves exactly as before the first trial. The probability to need at least $y + 1$ more trials is the same initially as after x unsuccessful trials.

Negative binomial distribution

Again, we consider a sequence of independent trials that lead to a success with probability p.
Now, we are interested in the time until n trials were successful. Let X be the number of trials until the nth successful trial.
Observe that X is the sum of n independent geometric random variables $X_1, \ldots, X_n$ with success probability p.

Definition

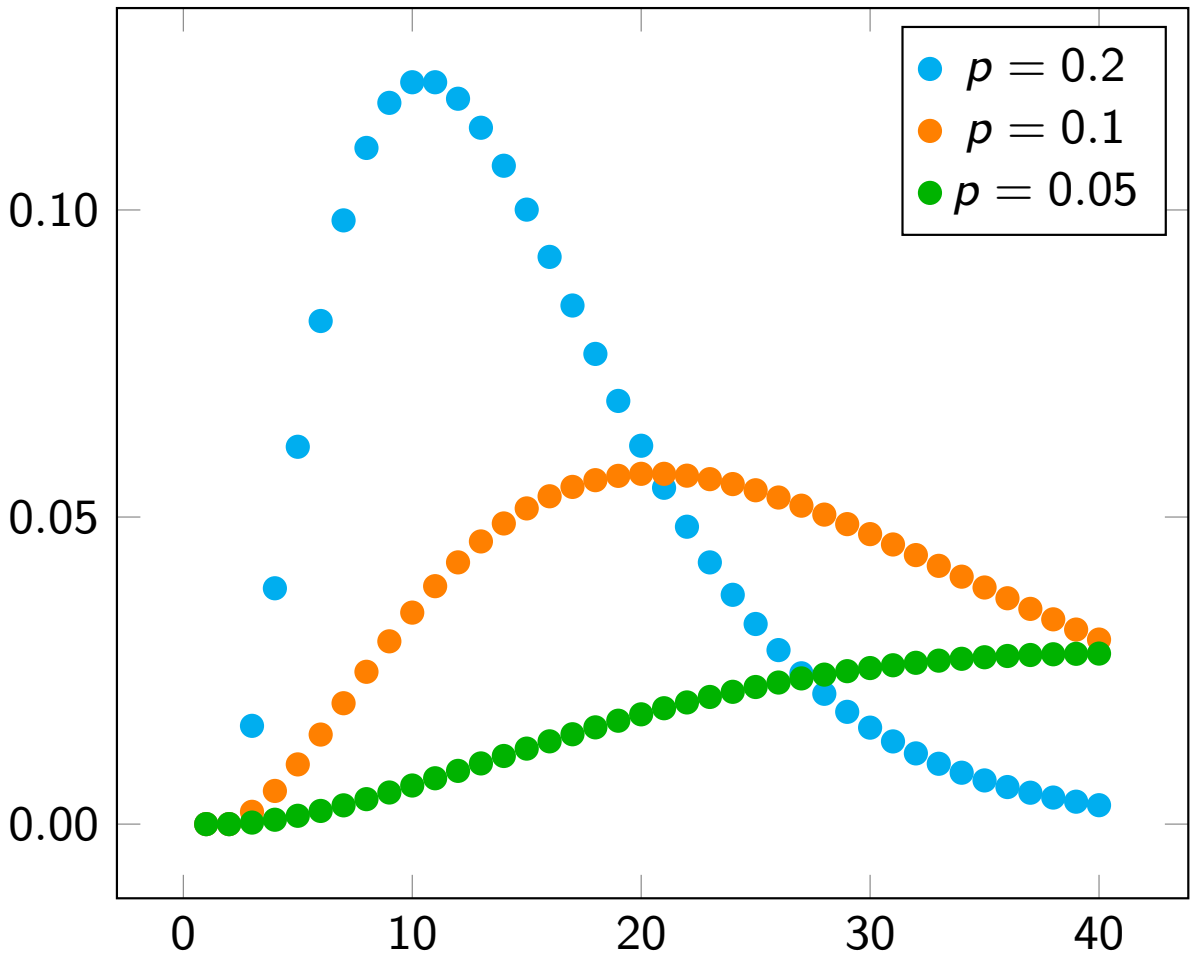
A random variable X with probability mass function

$$f_X(i) = \binom{i-1}{n-1} \cdot p^n (1-p)^{i-n} \text{ for } i \in \mathbb{N}$$

is called *negative binomial random variable* with parameters (n, p).

Negative binomial distribution

Plot of the probability mass function of negative binomial random variables with parameter $n = 3$ and different parameters p:



Negative binomial distribution

Theorem

For a negative binomial random variable X with parameters (n, p),

$$\mathbb{E}(X) = \frac{n}{p} \text{ and } \operatorname{Var}(X) = \frac{n(1-p)}{p^2}.$$

Proof.

This follows immediately from the fact that X is the sum of n independent geometric random variables X_i with success probability p and that

$$\mathbb{E}(X_i) = \frac{1}{p} \text{ and } \operatorname{Var}(X_i) = \frac{1-p}{p^2}.$$

□

Coupon collector's problem

If each box of a brand of cereals contains a coupon, and there are n different types of coupons, how many boxes of cereals do you have to buy in expectation to collect all types of coupons?

We assume that each box contains all types of coupons with equal probability.

Let X be the number of boxes until we find all n types of coupons.
Let X_i, for $i = 1, \ldots, n$, be the number of boxes we have to buy until we find the ith coupon after we have already found $i - 1$ different coupons.
So,

$$X = X_1 + \cdots + X_n.$$

Furthermore, $X_1, \ldots, X_n$ are independent.

E.g., let $n = 4$ and assume we draw coupons $1, \ldots, 4$ in the following order:

$$\underbrace{2}_{X_1}, \underbrace{2, 2, 3}_{X_2}, \underbrace{1}_{X_4}, \underbrace{1, 3, 2, 2, 4}_{X_4}.$$

Then, $X = X_1 + \cdots + X_4 = 1 + 3 + 1 + 5 = 10$.

Coupon collector's problem

Note that X_i counts the number of trials until we find one of the $n - i + 1$ coupons we still miss.
So, X_i is a geometric random variable with success probability $p_i = \frac{n-i+1}{n}$.
Hence,

$$\mathbb{E}(X_i) = \frac{n}{n - i + 1}.$$

We conclude

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} \frac{n}{n - i + 1} = n \cdot \sum_{i=1}^{n} \frac{1}{i}.$$

It is known that

$$\sum_{i=1}^{n} \frac{1}{i} = \log n + \mathcal{O}(1).$$

So,

$$\mathbb{E}(X) = n \log n + \mathcal{O}(n).$$

Coupon collector's problem

Number of Coupons	Expected Number of Boxes
1	1.00
2	3.00
3	5.50
4	8.33
5	11.42
6	14.70
7	18.15
8	21.76
9	25.51
10	29.41

Poisson distribution

Consider the following random variables:

1. The number of misprints on a page (or a group of pages) of a book
2. The number of wrong telephone numbers that are dialled in a day
3. The number of customers entering a post office on a given day
4. The number of α-particles discharged in a fixed period of time from some radioactive material

All of these random variables count the number of occurrences of events that approximately happen with a constant "rate" during a certain time period (or within a certain area). They satisfy:

- The number of outcomes occurring in one time interval or specified region of space is independent of the number that occur in any other disjoint time interval or region.
- The probability that a single outcome will occur during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of outcomes occurring outside this time interval or region.
- The probability that more than one outcome will occur in such a short time interval or fall in such a small region is negligible.

Poisson distribution

Such random variables can be modelled as Poisson random variables:

Definition

A *Poisson random variable* X with parameter $\lambda \geq 0$ satisfies $W_X = \mathbb{N}$ and has the probability mass function

$$f_X(i) = \frac{e^{-\lambda}\lambda^i}{i!} \text{ for } i \in \mathbb{N}.$$

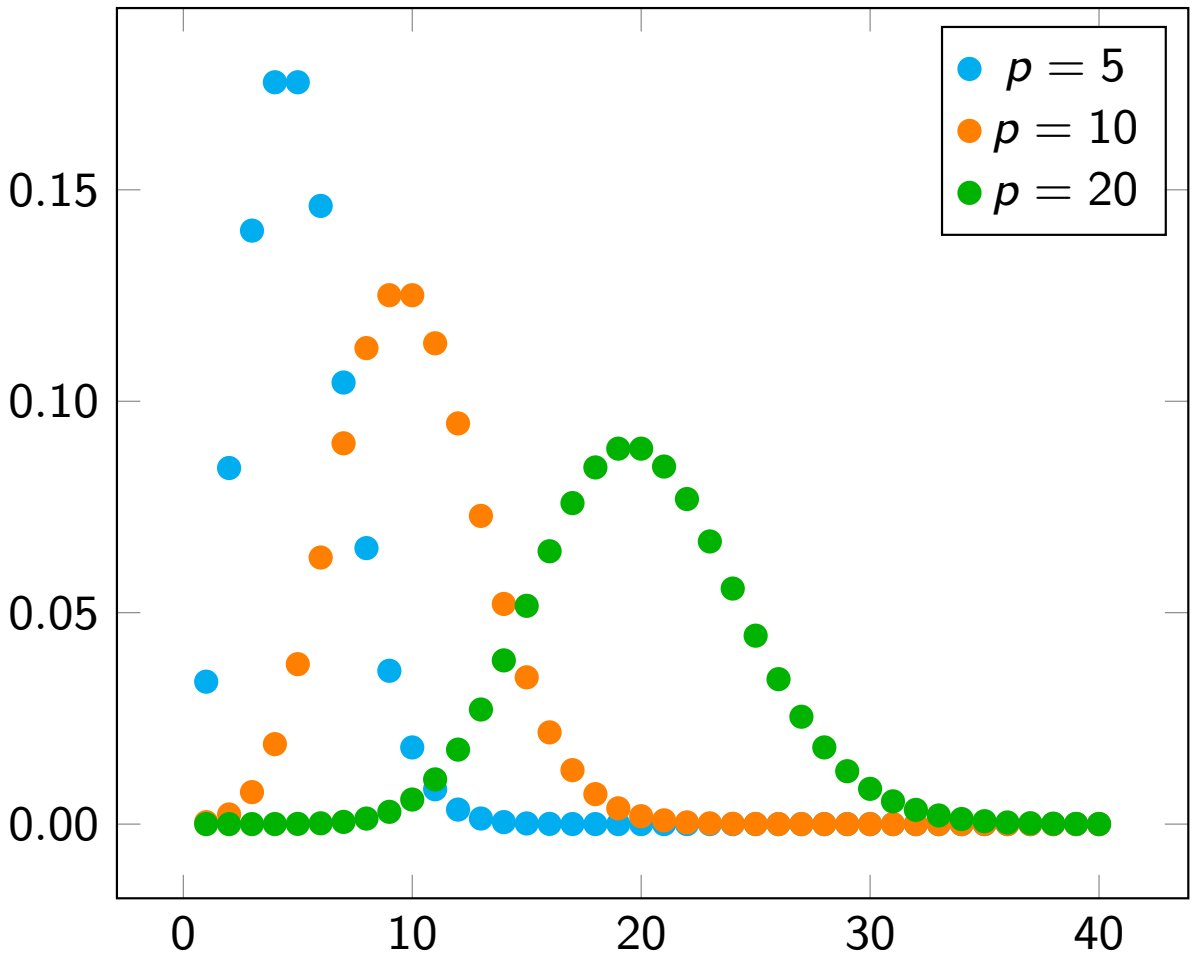
We write $X \sim \mathrm{Po}(\lambda)$ in this case.

This is an admissible probability mass functions as

$$\sum_{i=o}^{\infty} f_X(i) = \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Poisson distribution

Plot of the probability mass function of Poisson random variables with different parameters λ:



Poisson distribution

Example

Assume, we now that a certain region is hit by an earthquake (of at least a certain magnitude) on average once every 10^3 years.

This seems to satisfy all conditions for a Poisson distribution.

So, we assume that the number of earthquakes in the next 100 years is $X \sim \mathrm{Po}(\lambda)$ with $\lambda = 1/10$.

What is the probability of more than one earthquake in the next 100 years?

$$\Pr(X \geq 2) = 1 - \Pr(X = 0) - \Pr(X = 1) = 1 - e^{-\lambda} - \lambda e^{-\lambda} \approx 0.0047.$$

Poisson distribution

Theorem

For a Poisson random variable X with parameter $\lambda \geq 0$,

$$\mathbb{E}(X) = \lambda \text{ and } \operatorname{Var}(X) = \lambda.$$

Poisson distribution

Proof.

$$\mathbb{E}(X) = \sum_{i=0}^{\infty} i \cdot \frac{e^{-\lambda}\lambda^i}{i!} = \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}$$
$$= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

To compute the variance, we first compute

$$\mathbb{E}(X(X-1)) = \sum_{i=0}^{\infty} i(i-1) \cdot \frac{e^{-\lambda}\lambda^i}{i!} = \lambda^2 e^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!}$$
$$= \lambda^2 e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda^2.$$

Now

$$\mathrm{Var}(X) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \qquad \square$$

Poisson distribution

Fix a $\lambda > 0$ and, for $n > \lambda$, define $p_n := \lambda/n$. Let now

$$X_n \sim \mathrm{Bin}(n, p_n)$$

be a binomial random variable with parameters (n, p_n). Then, for $0 \leq k \leq n$,

$$\begin{aligned}
\Pr(X_n = k) = b(k; n, p_n) &= \binom{n}{k} \cdot p_n^k \cdot (1-p_n)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \cdot p_n^k \cdot (1-p_n)^{-k} \cdot (1-p_n)^n \\
&= \frac{(n \cdot p_n)^k}{k!} \cdot \frac{n!/(n-k)!}{n^k} \cdot (1-p_n)^{-k} \cdot (1-p_n)^n \\
&= \frac{\lambda^k}{k!} \cdot \frac{n!/(n-k)!}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \left(1 - \frac{\lambda}{n}\right)^n .
\end{aligned}$$

Poisson distribution

$$\Pr(X_n = k) = \frac{\lambda^k}{k!} \cdot \frac{n!/(n-k)!}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \left(1 - \frac{\lambda}{n}\right)^{n}.$$

We now let $n \to \infty$. Observe:

- $\lim_{n\to\infty} \frac{n!/(n-k)!}{n^k} = 1$.
- $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1$.
- $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n} = e^{-\lambda}$.

So,

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

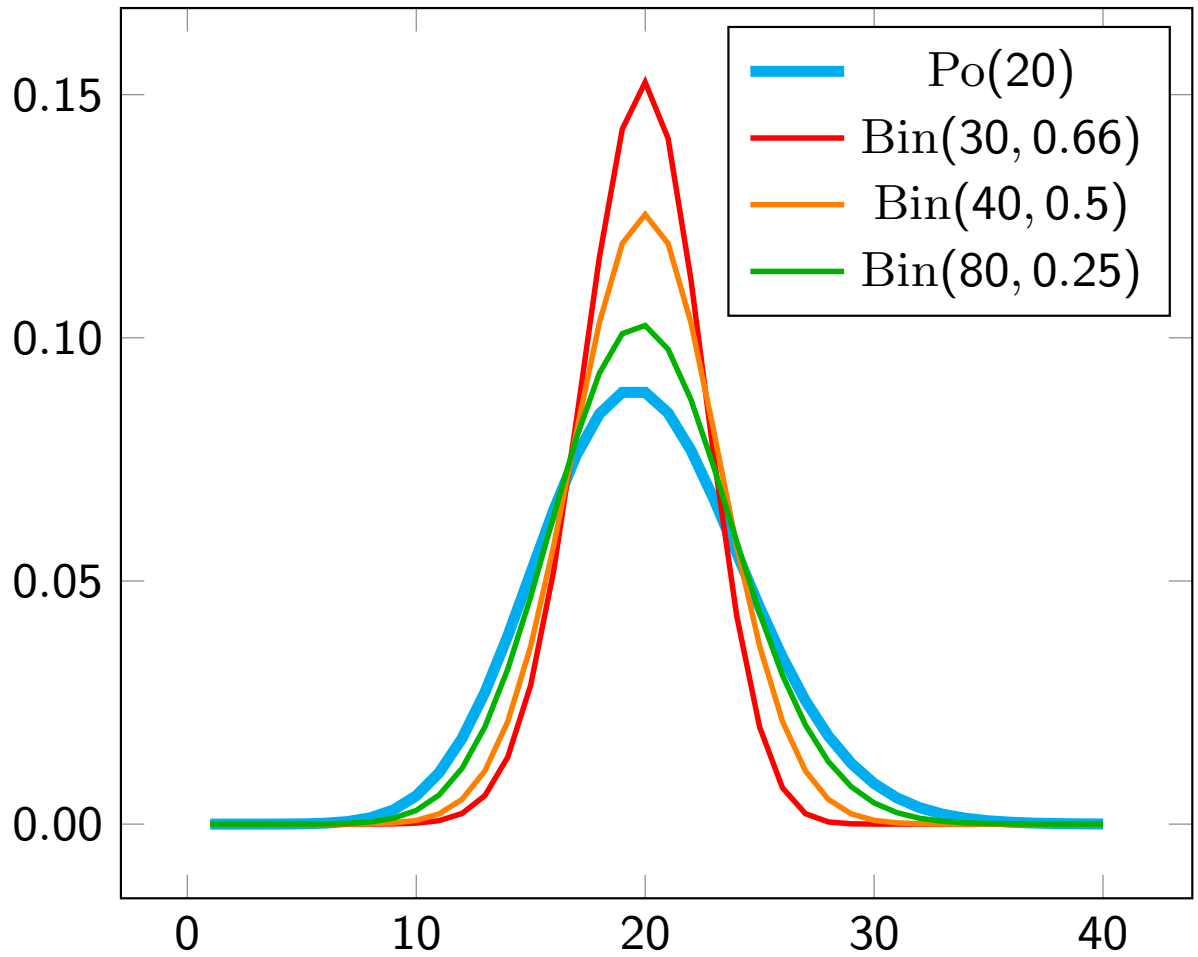
Note that for the Poisson random variable $Y \sim \mathrm{Po}(\lambda)$, we have

$$\Pr(Y = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

So, we can view a Poisson random variable with parameter λ as the limit of a sequence of binomial random variables with parameters $(n, \lambda/n)$.

Poisson distribution

Plot of the probability mass function of Poisson random variables with different parameters λ:



Poisson distribution

Theorem

Let X and Y be independent random variables with $X \sim \mathrm{Po}(\lambda)$ and $Y \sim \mathrm{Po}(\mu)$. Then,

$$Z := X + Y \sim \mathrm{Po}(\lambda + \mu).$$

Proof.

$$\begin{aligned} f_Z(z) &= \sum_{x=0}^{\infty} f_X(x) f_Y(z-x) = \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!} \cdot \frac{e^{-\mu}\mu^{z-x}}{(z-x)!} \\ &= e^{-(\lambda+\mu)} \cdot \frac{(\lambda+\mu)^z}{z!} \cdot \sum_{x=0}^{z} \frac{z!}{x!(z-x)!} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^x \cdot \left(\frac{\mu}{\lambda+\mu}\right)^{z-x} \\ &= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^z}{z!} \cdot \sum_{x=0}^{z} \binom{z}{x} p^x (1-p)^{z-x} \end{aligned}$$

where $p := \frac{\lambda}{\lambda+\mu}$. Note that

$$\sum_{x=0}^{z} \binom{z}{x} p^x (1-p)^{z-x} = 1.$$

□

Outline

Markov inequality

Theorem (Markov Inequality)

Let X be a random variable that takes only **non-negative** *values in $\mathbb{R}$. Let $a > 0$. Then,*

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Note that we can equivalently say that for all $t > 0$,

$$\Pr(X \geq t \cdot \mathbb{E}(X)) \leq \frac{1}{t}.$$

Markov inequality

Theorem (Markov Inequality)

Let X be a random variable that takes only non-negative values in $\mathbb{R}$. Let $a > 0$. Then,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof.

We show that $a \cdot \Pr(X \geq a) \leq \mathbb{E}(X)$:

$$\begin{aligned} a \cdot \Pr(X \geq a) &= \sum_{x \in W_X, x \geq a} t \cdot \Pr(X = x) \\ &\leq \sum_{x \in W_X, x \geq a} x \cdot \Pr(X = x) \\ &\leq \sum_{x \in W_X} x \cdot \Pr(X = x) = \mathbb{E}(X). \end{aligned}$$

□

Markov inequality

Theorem (Markov Inequality)

Let X be a random variable that takes only non-negative values in $\mathbb{R}$. Let $a > 0$. Then,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Alternative proof.

We know that

$$\mathbb{E}(X) = \mathbb{E}(X \mid X < a) \cdot \Pr(X < a) + \mathbb{E}(X \mid X \geq a) \cdot \Pr(X \geq a).$$

But now, $\mathbb{E}(X \mid X < a) \cdot \Pr(X < a) \geq 0$ and $\mathbb{E}(X \mid X \geq a) \geq a$ and we can conclude the claim immediately. □

Chebyshev Inequality

Theorem (Chebyshev Inequality)

Let X be a random variable and $a \in \mathbb{R}$ with $a > 0$. Then,

$$\Pr(|X - \mathbb{E}(X)| \geq a) \leq \frac{\operatorname{Var}(X)}{a^2}.$$

Equivalently, for all $t > 0$,

$$\Pr\left(|X - \mathbb{E}(X)| \geq t\sqrt{\operatorname{Var}(X)}\right) \leq \frac{1}{t^2}.$$

Recall that $\sqrt{\operatorname{Var}(X)}$ is the standard deviation of X.

Chebyshev Inequality

Theorem (Chebyshev Inequality)

Let X be a random variable and $a \in \mathbb{R}$ with $a > 0$. Then,

$$\Pr(|X - \mathbb{E}(X)| \geq a) \leq \frac{\operatorname{Var}(X)}{a^2}.$$

Proof.

First, observe that

$$\Pr\left(|X - \mathbb{E}(X)| \geq a\right) = \Pr\left(\left(X - \mathbb{E}(X)\right)^2 \geq a^2\right).$$

We define $Y := (X - \mathbb{E}(X))^2$.
Using the Markov inequality for Y and a^2, we obtain

$$\Pr(Y \geq a^2) \leq \frac{\mathbb{E}(Y)}{a^2}.$$

Now, $\Pr(Y \geq a^2) = \Pr\left(|X - \mathbb{E}(X)| \geq a\right)$ and $\mathbb{E}(Y) = \operatorname{Var}(X)$. □

Chebyshev Inequality

Example

We toss a fair coin 1000 times. Let X be the number of heads.

So, $X \sim \mathrm{Bin}(1000, 0.5)$ and hence

$$\mathbb{E}(X) = \frac{1}{2} \cdot 1000 = 500 \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{1}{2} \cdot \frac{1}{2} \cdot 1000 = 250.$$

What is the probability that there are more than 550 heads? The Chebyshev

Inequality tells us

$$\Pr(X \geq 550) \leq \Pr(|X - 500| \geq 50) \leq \frac{\mathrm{Var}(X)}{50^2} = 0.1.$$

For $Y \sim \mathrm{Bin}(10000, 0.5)$ we get

$$\Pr(Y \geq 5500) \leq \Pr(|Y - 5000| \geq 500) \leq \frac{\mathrm{Var}(Y)}{500^2} = \frac{2500}{250000} = 0.01.$$

The Weak Law of Large Numbers

Theorem (Weak law of large numbers)

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with expected value $\mu := \mathbb{E}(X_i)$ and finite variance $\sigma^2 := \operatorname{Var}(X_i)$. Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

Proof.

We have

$$\mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu$$

and

$$\operatorname{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\operatorname{Var}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

So, by the Chebyshev inequality,

$$\Pr\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

□

The Weak Law of Large Numbers

Note that the weak law of large numbers bounds the relative deviation

$$\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right|$$

and not the absolute deviation

$$|X_1 + \cdots + X_n - n\mu| \,.$$

Chernoff bounds

The Chernoff bounds are named after Herman Chernoff (*1923).

Theorem (Chernoff bounds)

Let $X_1, \dots, X_n$ be independent Bernoulli random variables with parameters $p_1, \dots, p_n$, respectively. So, $\Pr(X_i = 1) = p_i$ and $\Pr(X_i = 0) = 1 - p_i$ for $1 \leq i \leq n$.
Define $X := X_1 + \cdots + X_n$ and $\mu := \mathbb{E}(X) = \sum_{i=1}^{n} p_i$. For every $\delta > 0$,

$$\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu}.$$

Proof.

For any $t > 0$, we have

$$\Pr(X \geq (1+\delta)\mu) = \Pr(e^{tX} \geq e^{t(1+\delta)\mu}).$$

By the Markov Inequality, we conclude

$$\Pr(X \geq (1+\delta)\mu) = \Pr(e^{tX} \geq e^{t(1+\delta)\mu}) \leq \frac{\mathbb{E}(e^{tX})}{e^{t(1+\delta)\mu}}.$$

As $X_1, \ldots, X_n$ are independent,

$$\mathbb{E}(e^{tX}) = \mathbb{E}\left(\exp\left(\sum_{i=1}^{n} tX_i\right)\right) = \mathbb{E}\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} \mathbb{E}(e^{tX_i}).$$

Further, for $1 \leq i \leq n$, we can rewrite

$$\mathbb{E}(e^{tX_i}) = e^{t\cdot 1} p_i + e^{t\cdot 0}(1-p_i) = 1 + p_i(e^t - 1).$$

Proof continued.

Put together, we obtain

$$\begin{aligned}\Pr(X \geq (1+\delta)\mu) &\leq \frac{\prod_{i=1}^{n}(1+p_i(e^t-1))}{e^{t(1+\delta)\mu}} \\ &\leq \frac{\prod_{i=1}^{n}\exp(p_i(e^t-1))}{e^{t(1+\delta)\mu}} \\ &= \frac{\exp\left(\sum_{i=1}^{n}p_i(e^t-1)\right)}{e^{t(1+\delta)\mu}} \\ &= \frac{\exp\left(\mu(e^t-1)\right)}{e^{t(1+\delta)\mu}} =: f(t)\end{aligned}$$

As this holds for any $t > 0$, we can choose $t = \log(1+\delta)$ and obtain

$$f(t) = \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} = \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}.$$

□

Example

Again, we toss a fair coin n-times and want to provide an upper bound for the probability that heads occurs at least

$$\frac{n}{2}(1 + 10\%)\text{-times.}$$

We get

n	Chebyshev	Chernoff
1000	0.1	0.0889
10000	0.01	$0.308 \cdot 10^{-10}$
n	$\frac{100}{n}$	$\left(\frac{e^{0.1}}{1.1^{1.1}}\right)^{n/2} \approx 0.9976^n$

For n larger than 1000, the Chernoff bounds quickly become much better than the bounds obtained from the Chebyshev Inequality.

Definition (Probability-generating function)

For a random variable X with $W_X \subseteq \mathbb{N}$, the *probability-generating function* is the function

$$G_X(s) := \sum_{k=0}^{\infty} \Pr(X = k) \cdot s^k = \mathbb{E}(s^X).$$

We will focus on $s \in [-1, 1]$. In this interval, convergence is not an issue as

$$\begin{aligned} |G_X(s)| &= \left| \sum_{k=0}^{\infty} \Pr(X = k) \cdot s^k \right| \\ &\leq \sum_{k=0}^{\infty} \Pr(X = k) \cdot |s^k| \leq \sum_{k=0}^{\infty} \Pr(X = k) = 1. \end{aligned}$$

Probability- and moment-generating functions

Theorem

The probability mass function f_X and the cumulative distribution function F_X of a random variable X are uniquely determined by the probability-generating function G_X.

Proof.

This follows from the fact that two power series that are equal (as functions) are equal term by term. □

Probability- and moment-generating functions

Observe the following:

- Let $Y := X + t$ for $t \in \mathbb{N}$. Then,

$$G_Y(s) = \mathbb{E}(s^Y) = \mathbb{E}(s^{X+t}) = \mathbb{E}(s^t \cdot s^X) = s^t \cdot \mathbb{E}(s^X) = s^t \cdot G_X(s).$$

- $G'_X(s) = \sum_{k=1}^{\infty} k \cdot \Pr(X = k) \cdot s^{k-1}$ and hence $G'_X(0) = \Pr(X = 1)$.
- Analogously, $G_X^{(i)}(0) = \Pr(X = i) \cdot i!$, or equivalently $G_X^{(i)}(0)/i! = \Pr(X = i)$.

Probability- and moment-generating functions

Let us determine the probability-generating functions for some probability distributions we know.

Bernoulli distribution. Let X be a Bernoulli random variable with success probability p, i.e., $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$. Then,

$$G_X(s) = \mathbb{E}(s^X) = (1-p)s^0 + ps^1 = 1 - p + ps.$$

Binomial distribution. Let $Y \sim \mathrm{Bin}(n, p)$. Then,

$$G_Y(s) = \mathbb{E}(s^Y) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot s^k = (1 - p + ps)^n.$$

Note that Y can be seen as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$ with success probability p and that $G_Y(s) = G_{X_1}(s) \cdot \cdots \cdot G_{X_n}(s)$.

Theorem (Probability-generating function of sums)

Let $X_1, \dots, X_n$ be independent random variables. Let $Z := X_1 + \cdots + X_n$. Then,

$$G_Z(s) = G_{X_1}(s) \cdot \dots \cdot G_{X_n}(s).$$

Proof.

As $X_1, \dots, X_n$ are independent,

$$\begin{aligned} G_Z(s) &= \mathbb{E}\left(s^{X_1+\cdots+X_n}\right) \\ &= \mathbb{E}\left(s^{X_1} \cdot \dots \cdot s^{X_n}\right) \\ &= \mathbb{E}(s^{X_1}) \cdot \dots \cdot \mathbb{E}(s^{X_n}) \\ &= G_{X_1}(s) \cdot \dots \cdot G_{X_n}(s). \end{aligned}$$

□

Poisson distribution. Let $X \sim \mathrm{Po}(\lambda)$ for $\lambda \geq 0$. Then,

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} s^k = e^{-\lambda+\lambda s} = e^{\lambda(s-1)}.$$

Let $Y \sim \mathrm{Bin}(n, \lambda/n)$. For $n \to \infty$, we get

$$G_Y(s) = \left(1 - \frac{\lambda}{n} + \frac{\lambda s}{n}\right)^n = \left(1 + \frac{\lambda(s-1)}{n}\right)^n \to e^{\lambda(s-1)} = G_X(s).$$

One can show in general that for a sequence of random variables $X_1, X_2, \ldots$ and a random variable X, it follows from $G_{X_n} \underset{n\to\infty}{\longrightarrow} G_X$ that the probability mass functions of $X_1, X_2, \ldots$ converge to the probability mass function of X.

Uniform distribution on $\{0, \dots, n\}$. Let X be uniformly distributed on $\{0, \dots, n\}$, i.e., for $0 \leq k \leq n$, we have $\Pr(X = k) = 1/(n+1)$. Then,

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{n} \frac{1}{n+1} s^k = \frac{s^{n+1} - 1}{(n+1)(s-1)}.$$

Geometric distribution. Let X be a geometric random variable with success probability p. Then,

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=1}^{\infty} p(1-p)^{k-1} s^k = ps \sum_{k=1}^{\infty} ((1-p)s)^{k-1} = \frac{ps}{1-(1-p)s}.$$

Probability- and moment-generating functions

Theorem

For a random variable X with $W_X \subseteq \mathbb{N}$ and probability-generating function $G_X(s)$,

$$G_X'(1) = \mathbb{E}(X).$$

Proof.

We know that

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} \Pr(X = k) \cdot s^k.$$

So,

$$G_X'(s) = \sum_{k=1}^{\infty} \Pr(X = k) \cdot k \cdot s^{k-1}.$$

For $s = 1$, we obtain

$$G_X'(1) = \sum_{k=1}^{\infty} \Pr(X = k) \cdot k = \mathbb{E}(X).$$

□

Example

Let $X \sim \mathrm{Bin}(n, p)$. So,

$$G_X(s) = (1 - p + ps)^n.$$

We obtain

$$G'_X(s) = n \cdot (1 - p + ps)^{n-1} \cdot p$$

and hence

$$\mathbb{E}(X) = G'_X(1) = np.$$

Example

If we take the second derivative of

$$G_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} \Pr(X = k) \cdot s^k,$$

we obtain

$$G_X''(s) = \sum_{k=2}^{\infty} \Pr(X = k) \cdot k \cdot (k-1) \cdot s^{k-2}.$$

So,

$$G_X''(1) = \sum_{k=2}^{\infty} \Pr(X = k) \cdot k \cdot (k-1) = \mathbb{E}(X(X-1)).$$

This allows us to compute

$$\mathrm{Var}(X) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) - \mathbb{E}(X)^2 = G_X''(1) + G_X'(1) - (G_X'(1))^2.$$

Other moments of X can be computed similarly.

Example

Let $X_i \sim \mathrm{Bin}(n_i, p)$ for $i = 1, \ldots, k$ be independent random variables and let $Z := X_1 + \cdots + X_k$. Then,

$$G_Z(s) = \prod_{i=1}^{k} (1 - p + ps)^{n_i} = (1 - p + ps)^{\sum_{i=1}^{k} n_i}$$

and hence

$$Z \sim \mathrm{Bin}(\sum_{i=1}^{k} n_i, p).$$

Let $X_1, \ldots, X_k \sim \mathrm{Po}(\lambda)$ be independent random variables and let $Z := X_1 + \cdots + X_k$. Then,

$$G_Z(s) = \prod_{i=1}^{k} e^{\lambda(s-1)} = e^{k\lambda(s-1)}$$

and hence

$$Z \sim \mathrm{Po}(k\lambda).$$

Probability- and moment-generating functions

We now consider sums of a random number of random variables:

Theorem

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $W_{X_i} \subseteq \mathbb{N}$ and probability-generating function $G_X(s)$. Let N be a random variable independent from $X_1, X_2, \ldots$ with $W_N \subseteq \mathbb{N}$ and probability-generating function $G_N(s)$.
Let $Z := X_1 + \cdots + X_N$. Then,

$$G_Z(s) = G_N(G_X(s)).$$

Proof.

$$
\begin{aligned}
G_Z(s) &= \mathbb{E}(s^Z) \\
&= \sum_{n=0}^{\infty} \mathbb{E}(s^Z \mid N = n) \cdot \Pr(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}(s^{X_1 + \cdots + X_n}) \cdot \Pr(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}(s^{X_1}) \cdot \ldots \cdot \mathbb{E}(s^{X_n}) \cdot \Pr(N = n) \\
&= \sum_{n=0}^{\infty} (G_X(s))^n \cdot \Pr(N = n) \\
&= \mathbb{E}\left((G_X(s))^N \right) \\
&= G_N(G_X(s)).
\end{aligned}
$$

□

Definition (Moment-generating function)

Let X be a random variable. The *moment-generating function* of X is defined as

$$M_X(s) := \mathbb{E}\left(e^{sX}\right).$$

We can write

$$M_X(s) = \mathbb{E}(e^{sX}) = \mathbb{E}\left(\sum_{i=0}^{\infty} \frac{(sX)^i}{i!}\right) = \sum_{i=0}^{\infty} \frac{s^i \mathbb{E}(X^i)}{i!}.$$

Furthermore, if $W_X \subseteq \mathbb{N}$,

$$M_X(s) = \mathbb{E}(e^{sX}) = \mathbb{E}((e^s)^X) = G_X(e^s).$$

Theorem (Moment-generating function of sums)

Let $X_1, \dots, X_n$ be independent random variables. Let $Z := X_1 + \cdots + X_n$. Then,

$$M_Z(s) = M_{X_1}(s) \cdot \dots \cdot M_{X_n}(s).$$

Proof.

As $X_1, \dots, X_n$ are independent,

$$\begin{aligned} M_Z(s) &= \mathbb{E}\left(e^{s(X_1+\cdots+X_n)}\right) \\ &= \mathbb{E}\left(e^{sX_1} \cdot \dots \cdot e^{sX_n}\right) \\ &= \mathbb{E}(e^{sX_1}) \cdot \dots \cdot \mathbb{E}(e^{sX_n}) \\ &= M_{X_1}(s) \cdot \dots \cdot M_{X_n}(s). \end{aligned}$$

□

Outline

Motivation

Many quantities do not take a discrete set of values, but arbitrary real numbers:

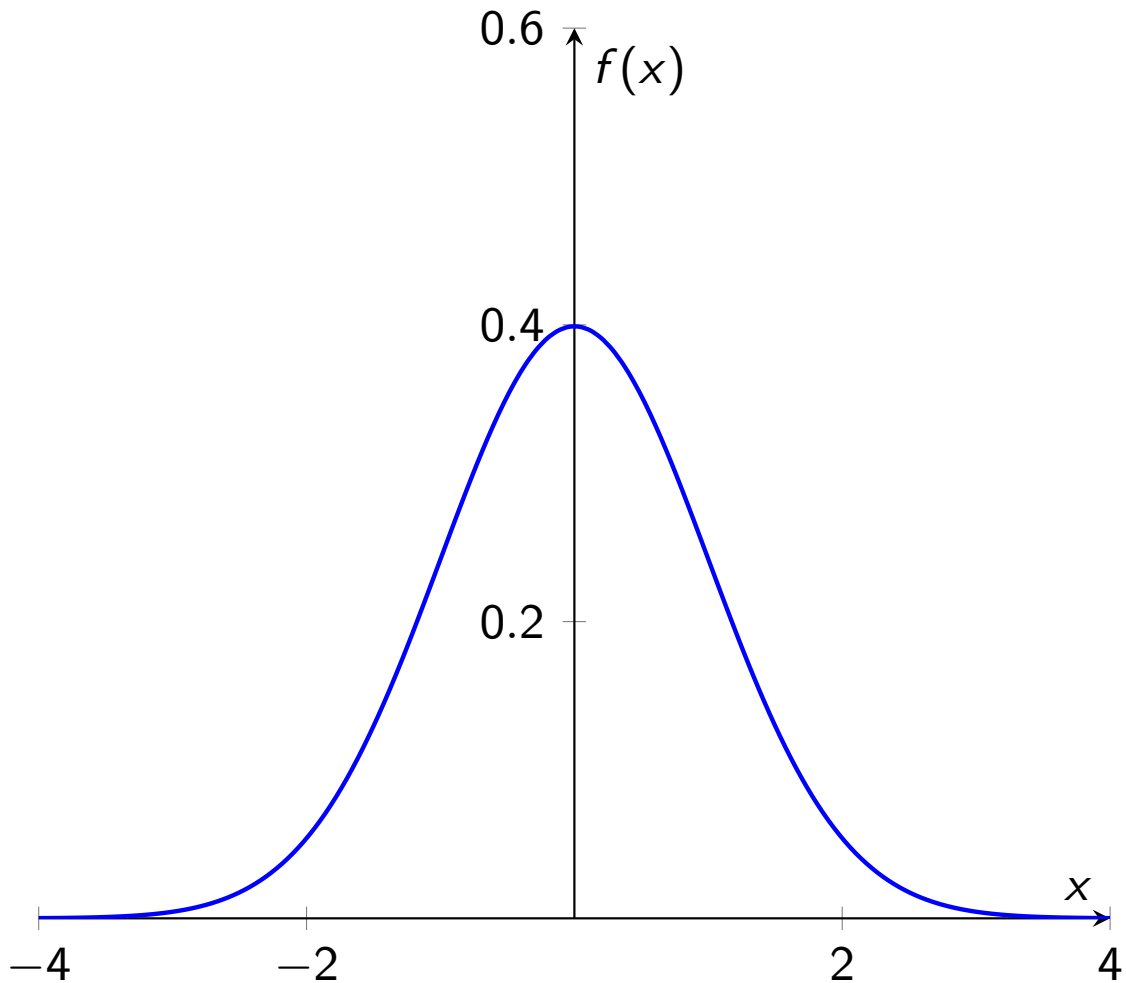
- the time until the next request is sent to the server
- the height of an individual from some population
- the exchange ratio between two currencies at a certain point in time

Example

For a different perspective, assume a server checks new requests every $1/n$ time units. A random variable T measuring the time until the next request then takes values $W_T = \{1/n, 2/n, 3/n, \dots\}$.

If we let $n \to \infty$, i.e., let the time interval between to checks tend to 0, a discrete set of values W_T is not sufficient anymore. Instead T can take any value in $\mathbb{R}_{\geq 0}$.

Motivation



density: how likely is it to land in a short interval around a point (relative to the width of an interval)

Continuous random variable

Definition (Continuous random variable)

A *continuous random variable* X and the underlying *real probability space* are given by a *probability density function* $f_X : \mathbb{R} \to \mathbb{R}_0^+$ with the property that

$$\int_{-\infty}^{+\infty} f_X(x)\mathrm{d}x = 1.$$

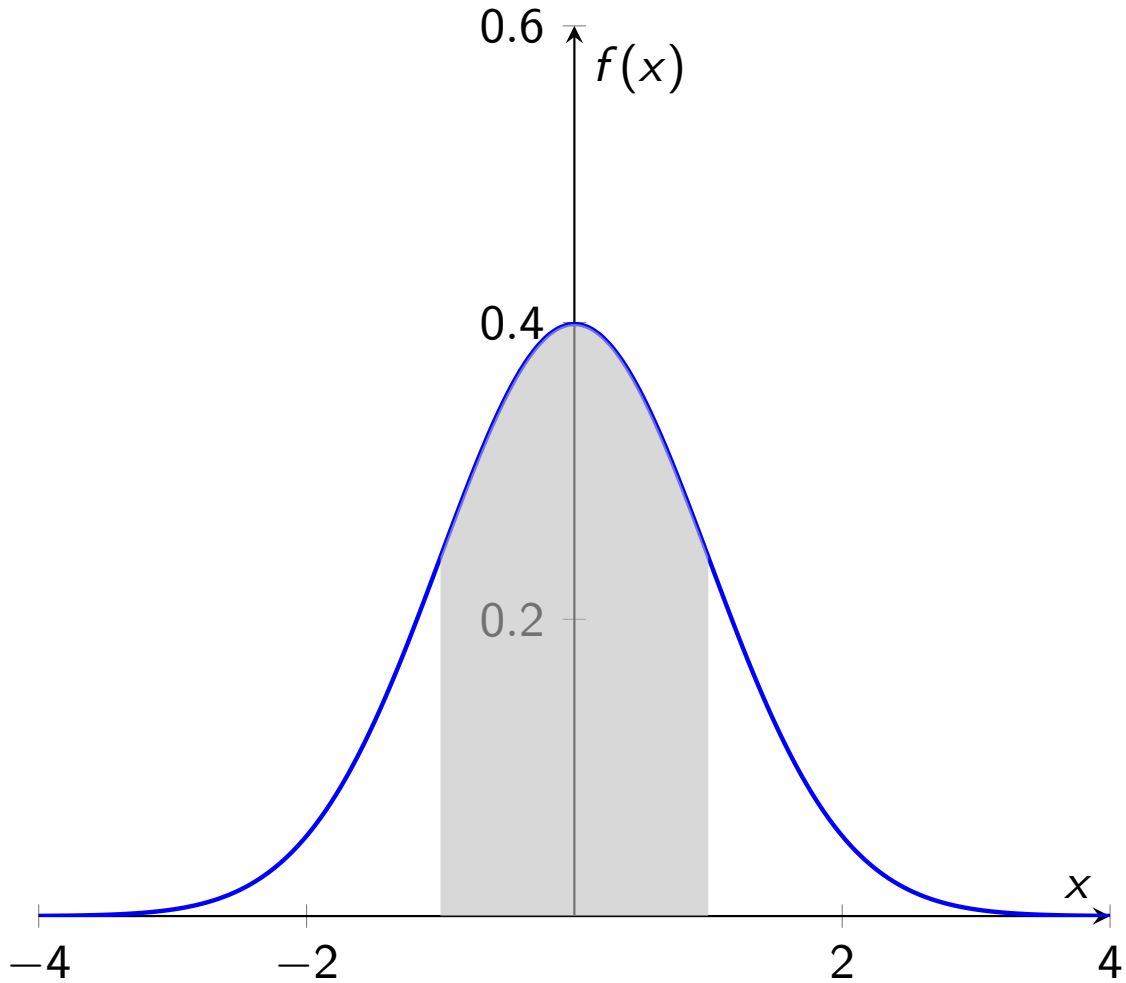
An *event* is a set $A \subseteq \mathbb{R}$ that is the union $\bigcup_{i \in I} A_i$ of at most countably many intervals (of any type; bounded, unbounded, open closed, half-open) A_i, $i \in I$.

The event A occurs if X takes a value in A. The probability of A is

$$\Pr(A) = \int_A f_X(x)\mathrm{d}x = \sum_{i \in I} \int_{A_i} f_X(x)\mathrm{d}x$$

Continuous random variable

Example



For the event $A = [-1.1]$, the probability is $\Pr(A) = \int_{-1}^{1} f(x)\mathrm{d}x$.

Example

A simple probability density function is the *uniform distribution* on a real interval $[a, b]$. It is defined as

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

For any interval $A = [c, d] \subseteq [a, b]$, we now have

$$\Pr(A) = \int_c^d f(x)\mathrm{d}x = \frac{d-c}{b-a}.$$

Continuous random variable

Definition

Given a continuous random variable X with probability density function f_X, we define the *cumulative distribution function* F_X as

$$F_X(c) = \Pr(X \leq c) = \Pr(\{t \in \mathbb{R} \mid t \leq c\}) = \int_{-\infty}^{c} f_X(t)\mathrm{d}t.$$

Example

For the uniform distribution on $[a, b]$, with probability density function

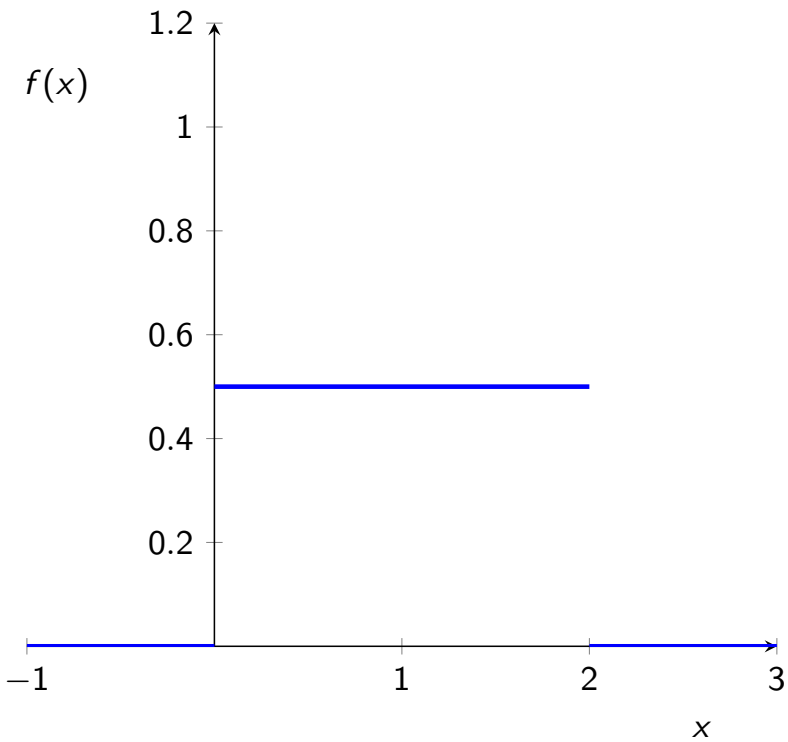
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b], \\ 0 & \text{otherwise,} \end{cases}$$

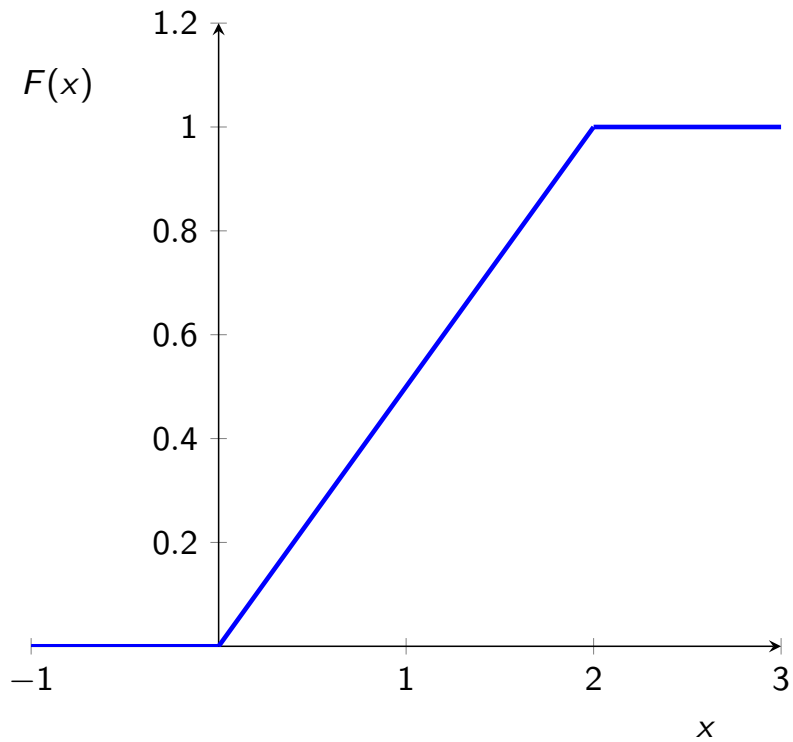
we obtain the cumulative distribution function

$$F(x) = \int_{-\infty}^{x} f(x)\mathrm{d}x = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b, \\ 1 & \text{for } x > b. \end{cases}$$

Example

The uniform distribution on $[0, 2]$ has the following probability density function and cumulative distribution function:





Continuous random variable

Observations on the cumulative distribution function:

- F_X is monotonically increasing.
- F_X is continuous. That is why these random variables are called continuous.
- We have $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.
- Every differentiable function F with these properties is the cumulative distribution function of a continuous random variable with probability density function $f(x) = F'(x)$.
- We have $\Pr(a < X < b) = F_X(b) - F_X(a)$.
- In the continuous case, we do not have to distinguish between the events "$a \leq X \leq b$", "$a < X \leq b$", "$a \leq X < b$", and "$a < X < b$".

General probability spaces

Before we take a closer look at continuous random variables, we line out the mathematical foundations of general probability spaces.

In these probability spaces, we want to allow arbitrary sample spaces.

The challenge is that we cannot assign probabilities to single sample points anymore.

σ-algebra

If a sample space Ω is uncountable, e.g., a real interval $[a, b]$, we cannot allow all arbitrary subsets of Ω as events.

Definition (σ-algebra)

Let Ω be a set. A set $\mathcal{A} \subseteq 2^\Omega$ is called *σ-algebra* over Ω if the following conditions hold:

1. $\Omega \in \mathcal{A}$,
2. if $A \in \mathcal{A}$, then also $\Omega \setminus A \in \mathcal{A}$,
3. if $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$, then also $\bigcup_{n=0}^{\infty} A_n \in \mathcal{A}$.

For any set Ω, the power set 2^Ω is a σ-algebra.

For $\Omega = \mathbb{R}$, the *Borel σ-algebra* consists of all sets $A \subseteq \mathbb{R}$ that can be constructed from intervals (open and closed) by taking *countable* intersections and unions. Such sets are called *Borel sets*.

Kolmogorov axioms

Definition (Probability space)

Let Ω be an arbitrary set and $\mathcal{A}$ a σ-algebra over Ω. A map

$$\Pr\colon \mathcal{A} \to [0,1]$$

is called *probability measure* on $\mathcal{A}$ if:

1. $\Pr(\Omega) = 1$,
2. for countably many pairwise disjoint sets $A_1, A_2, \dots \in \mathcal{A}$,

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

The tuple $(\Omega, \mathcal{A}, \Pr)$ is then called a *probability space*.

The elements of $\mathcal{A}$ are called *events* and $\Pr(A)$ is the *probability* of an event A.

Probability spaces

Let $(\Omega, \mathcal{A}, \Pr)$ be a probability space. For events A, B, we have

1. $\Pr(\emptyset) = 0$ and $\Pr(\Omega) = 1$.
2. $0 \leq \Pr(A) \leq 1$.
3. $\Pr(\Omega \setminus A) = 1 - \Pr(A)$.
4. If $A \subseteq B$, then $\Pr(A) \leq \Pr(B)$.

Lebesgue integral

A function $f : \mathbb{R} \to \mathbb{R}$ is called *measurable* if the pre-image of any Borel set is a Borel set.

Every continuous function is measurable. Products and sums of measurable functions are measurable.

For each Borel set A, the indicator function

$$I_A : x \mapsto \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise} \end{cases}$$

is measurable.

To every measurable function, we can assign the *Lebesgue integral*

$$\int f \, \mathrm{d}\lambda.$$

(Here the λ denotes the Lebesgue measure that assigns a "volume" to each Borel set. You can think of $\int_{-\infty}^{\infty} f(x) \, \mathrm{d}x$ instead.)

Lebesgue integral

Let $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a measurable function with $\int f \, \mathrm{d}\lambda = 1$. Then, we can define the following map on the Borel σ-algebra $\mathcal{A}$:

$$\begin{aligned} \Pr : \mathcal{A} &\to [0, 1] \\ A &\mapsto \int f \cdot I_A \, \mathrm{d}\lambda. \end{aligned}$$

This results in the probability space $(\mathbb{R}, \mathcal{A}, \Pr)$.

We cannot extend such a probability space to the σ-algebra $2^{\mathbb{R}}$.

Computations with continuous random variables

Let X be a continuous random variable with probability density function f_X.
Let $Y = g(X)$ for a function $g\colon \mathbb{R} \to \mathbb{R}$.

The cumulative distribution function of Y is given by

$$F_Y(x) = \Pr(Y \leq x) = \Pr(g(X) \leq x) = \int_C f_X(t)\mathrm{d}t$$

where $C = \{t \in \mathbb{R} \mid g(t) \leq y\}$.

Note that this integral is only defined if C is an admissible event. This is the case if g is a measurable function.

From F_Y, we can determine f_Y by taking the derivative.

Computations with continuous random variables

Example

Let X be uniformly distributed on $(0, 1)$. For a constant $\lambda > 0$, we define $Y = -1/\lambda \cdot \log(X)$.

$$\begin{aligned} F_Y(x) &= \Pr(-(1/\lambda)\log(X) \leq x) = \Pr(\log(X) \geq -\lambda x) \\ &= \Pr(X \geq e^{-\lambda x}) = 1 - F_X(e^{-\lambda x}) \\ &= \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

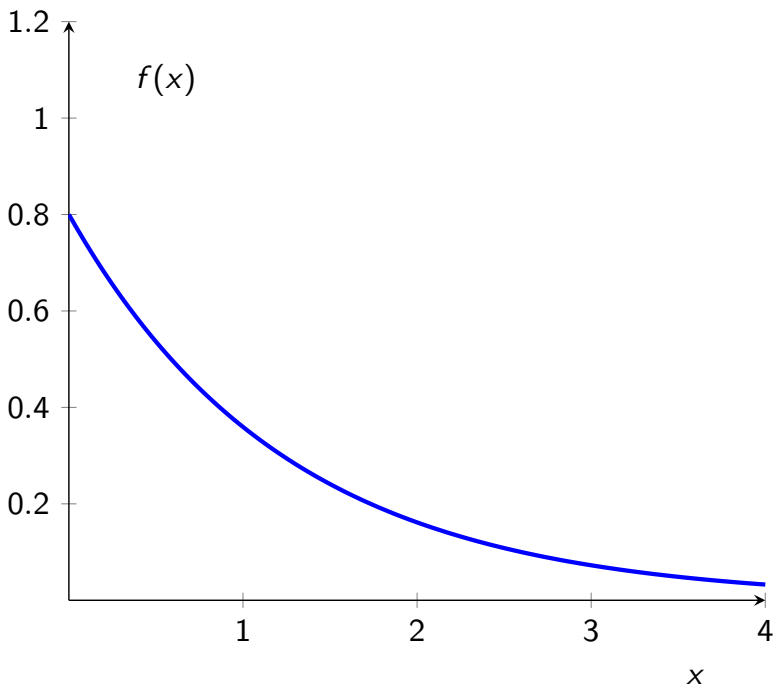
For the probability density function $f_Y(x) = F_Y'(x)$, we obtain

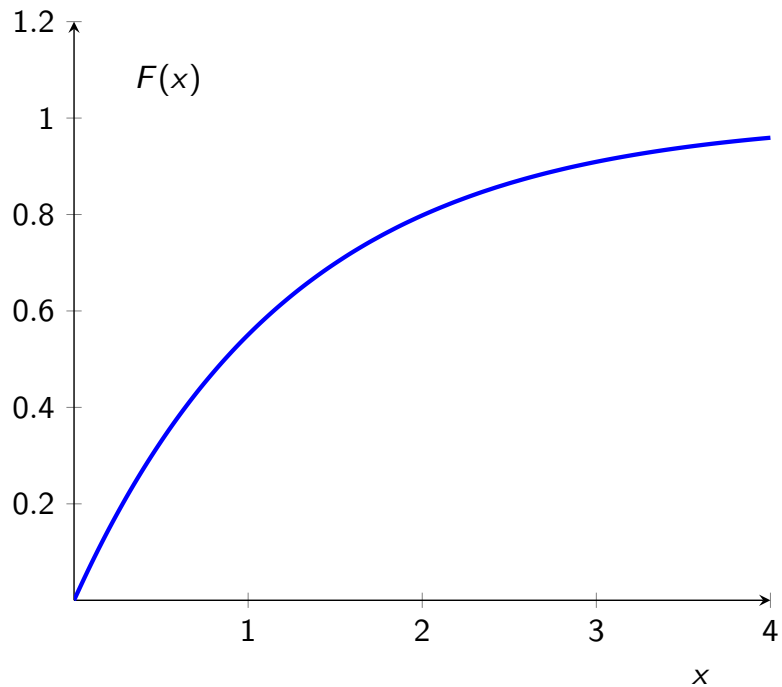
$$f_Y(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

A random variable with such a probability density function is called *exponentially distributed*.

Example

The exponential distribution with parameter $\lambda = 0.8$ has the following probability density function and cumulative distribution function:





Computations with continuous random variables

Example

Let X be a continuous random variable with probability density function f_X.
Let $Y = a \cdot X + b$ for $a, b \in \mathbb{R}$ with $a > 0$.

$$F_Y(x) = \Pr(aX + b \leq x) = \Pr\left(X \leq \frac{x-b}{a}\right) = F_X\left(\frac{x-b}{a}\right)$$

Hence

$$f_Y(x) = F_Y'(x) = \frac{\mathrm{d}(x-b)/a}{\mathrm{d}x} = f_X\left(\frac{x-b}{a}\right) \cdot \frac{1}{a},$$

Simulation of continuous random variables

Simulating a continuous random variable X with probability density function f_X means algorithmically generating random values whose distribution follows the distribution of X.

We assume that X has a *strictly* increasing, continuous cumulative distribution function F_X with value range $(0, 1)$.

At the same time, we assume that we can simulate a uniformly distributed random variable U on $(0, 1)$.

By our assumptions on F_X, there is a unique inverse function F_X^{-1} such that, for all $x \in (0, 1)$,

$$F_X(F_X^{-1}(x)) = x.$$

Now, we define

$$\tilde{X} = F_X^{-1}(U).$$

Then,

$$\Pr(\tilde{X} \leq t) = \Pr(F_X^{-1}(U) \leq t) = \Pr(U \leq F_X(t)) = F_U(F_X(t)) = F_X(t).$$

Example

The example above was obtained in this way: For the exponential distribution with $\lambda > 0$, we have $F_X(t) = 1 - e^{-\lambda t}$ for $t \geq 0$.

On $(0, 1)$, we obtain the inverse function $F_X^{-1}(x) = -(1/\lambda)\log(1 - x)$. So, we choose $\tilde{X} = F_X^{-1}(U) = -(1/\lambda)\log(1 - U)$.

Clearly, U and $1 - U$ have the same distribution. Hence, we actually chose $\tilde{X} = -(1/\lambda)\log(U)$ instead.

Continuous distributions as limits of discrete distributions

Let X be a continuous random variable. We construct the following discrete random variables from X:

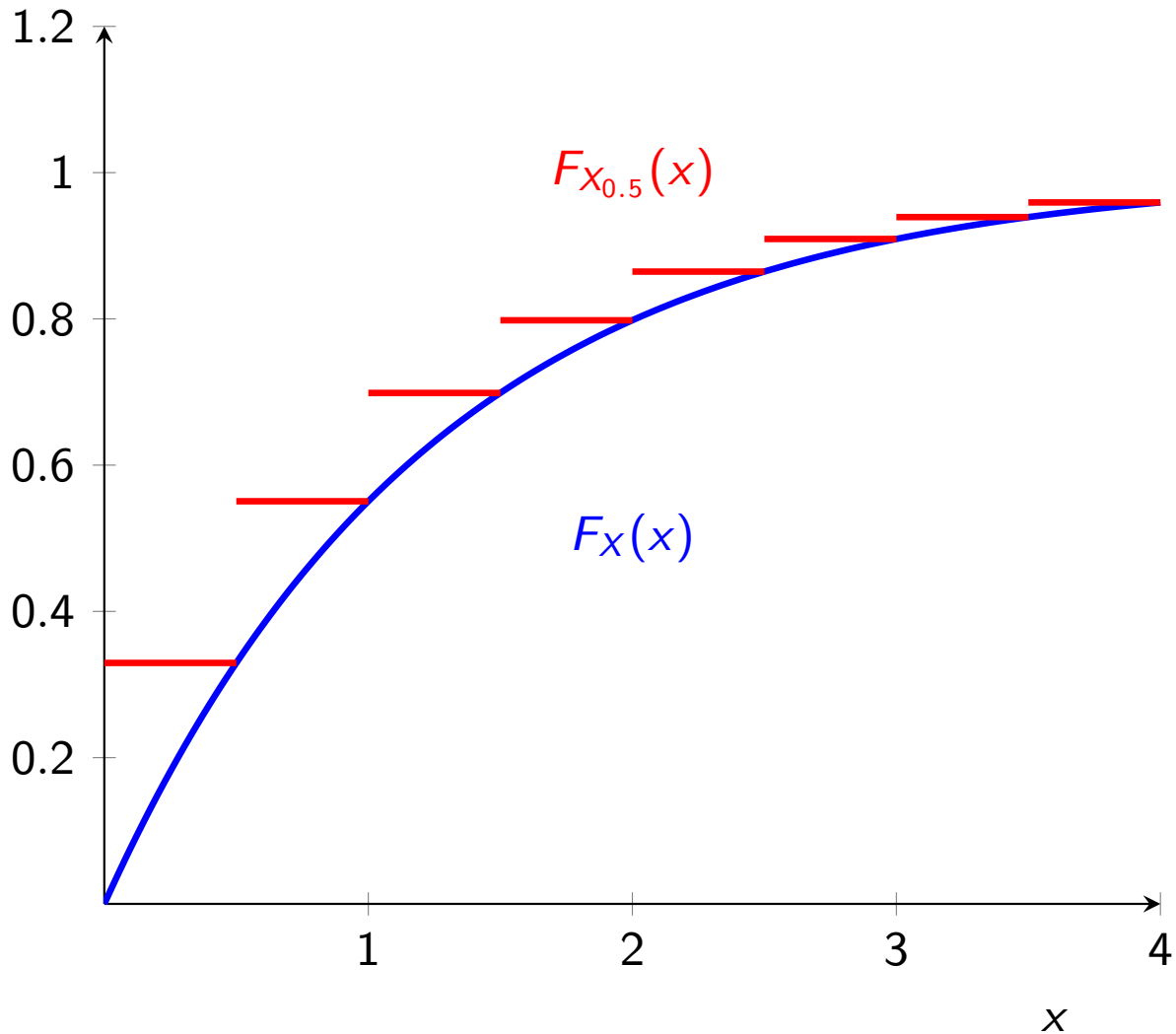
For $\delta > 0$, we define

$$X_\delta = n\delta \text{ iff } X \in [n\delta, (n+1)\delta) \text{ for } n \in \mathbb{Z}.$$

Then,

$$\Pr(X_\delta = n\delta) = F_X((n+1)\delta) - F_X(n\delta).$$

Continuous distributions as limits of discrete distributions



For $\delta \to 0$, the cumulative distribution functions F_{X_δ} converge to F_X.

Expected value and variance

Definition

Let X be a continuous random variable with probability density function f_X. The expected value of X is defined as

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} t \cdot f_X(t) \mathrm{d}t$$

if the integral

$$\int_{-\infty}^{\infty} |t| \cdot f_X(t) \mathrm{d}t$$

is finite.

Note that for discrete random variables Y with probability *mass* function f_Y, we had the analogue requirement for the existence of the expected value that

$$\sum_{y \in W_Y} |y| \cdot f_Y(y)$$

converges (i.e., is finite).

Expected value and variance

Definition

Let X be a continuous random variable with probability density function f_X.
The variance of X is defined as

$$\operatorname{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \int_{-\infty}^{\infty} (t - \mathbb{E}(X))^2 \cdot f_X(t) \mathrm{d}t$$

if the expected value $\mathbb{E}\left((X - \mathbb{E}(X))^2\right)$ exists.

As in the discrete setting, we can compute the variance also as

$$\operatorname{Var}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}\left(X\right)^2.$$

Example

Consider a uniformly distributed random variable X on the interval $[a, b]$. So, the probability density function is

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

For the expected value, we obtain

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} t \cdot f_X(t) \mathrm{d}t = \frac{1}{b-a} \int_a^b t \mathrm{d}t \\ &= \frac{1}{2(b-a)} \cdot \left[t^2 \right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

Example (Example continued)

Furthermore,

$$\mathbb{E}\left(X^2\right) = \int_{-\infty}^{\infty} t^2 \cdot f_X(t)\mathrm{d}t = \frac{1}{b-a}\int_a^b t^2\mathrm{d}t$$
$$= \frac{b^3 - a^3}{3(b-1)} = \frac{b^2 + ab + a^2}{3}.$$

So, we obtain for the variance

$$\mathrm{Var}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}\left(X\right)^2 = \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4} = \frac{(a-b)^2}{12}.$$

Expected value and variance

We already used the following statement

Lemma

Let X be a continuous random variable and let $Y = g(X)$ for a function $g : \mathbb{R} \to \mathbb{R}$. Then,

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(t) \cdot f_X(t) \mathrm{d}t.$$

Proof.

We only prove the statement for the case that g is linear, i.e., $Y = aX + b$ for $a, b \in \mathbb{R}$ with $a \neq 0$.

We have already computed the probability density function

$$f_Y(x) = f_X\left(\frac{x-b}{a}\right) \cdot \frac{1}{a}.$$

So,

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} t \cdot f_X\left(\frac{x-b}{a}\right) \cdot \frac{1}{a} \mathrm{d}t.$$

Letting $u = (t-b)/a$ and accordingly $\mathrm{d}u = (1/a)\mathrm{d}t$, we obtain

$$\mathbb{E}(aX+b) = \int_{-\infty}^{\infty} (au+b) f_X(u) \mathrm{d}u.$$

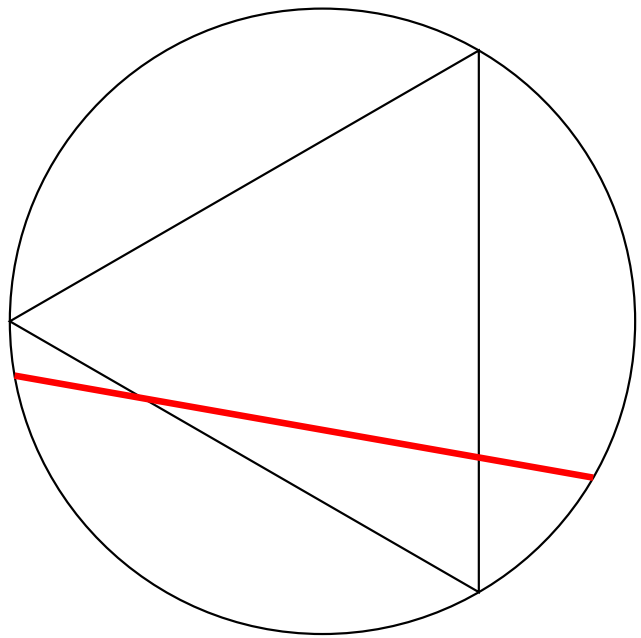
□

Principle of Indifference

In the discrete setting, the principle of indifference stated that all sample points should have equal probability if there is no reason to believe otherwise.

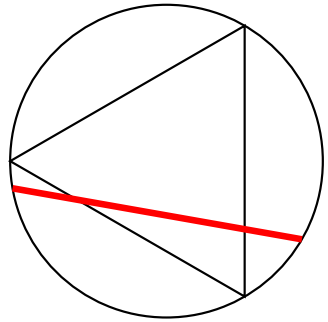
Example (Bertrand paradox)

Consider a circle with an equilateral triangle inscribed.



What is the probability that a randomly chosen chord of the circle is longer that a side of the triangle (event A)?

Principle of Indifference



We want to choose chords of the circle uniformly. But what does this mean?

The length of a chord of the circle depends only on one of the following parameters:

- the distance d to the center of the circle.
- the angle φ with the center of the circle.

Note that the sides of the triangle have distance $r/2$ from the center of the circle with radius r and angle $120°$.

If we assume a uniform distribution of one of the parameters, we obtain:

- If the distance $d \in [0, r]$ is uniformly distributed, $\Pr(A) = \Pr(d < r/2) = 1/2$.
- If $\phi \in [0°, 180°]$ is uniformly distributed, $\Pr(A) = \Pr(\phi > 120°) = 1/3$.

Outline

Uniform distribution

Probability density function:

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

Cumulative distribution function:

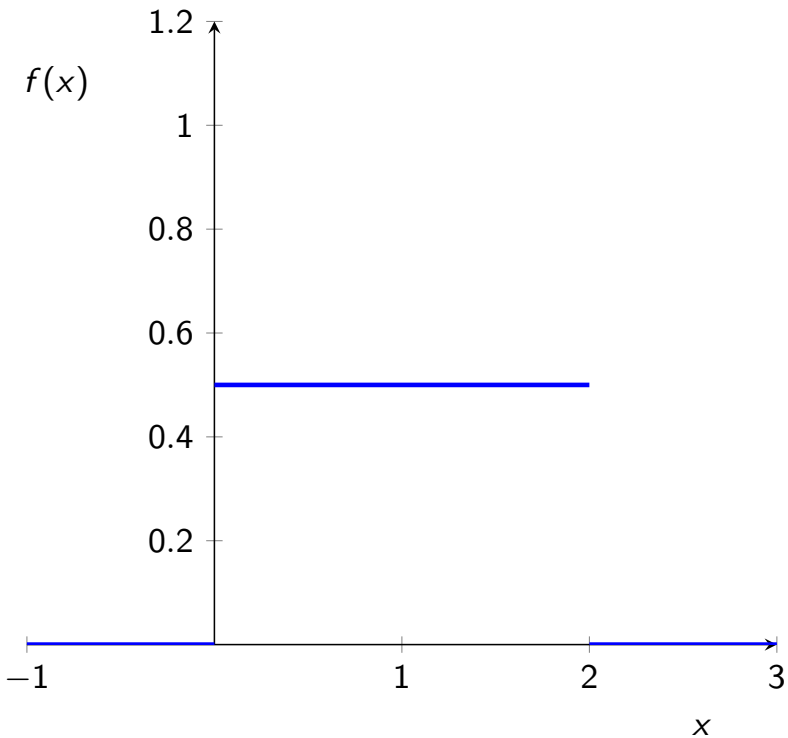
$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b, \\ 1 & \text{for } x > b. \end{cases}$$

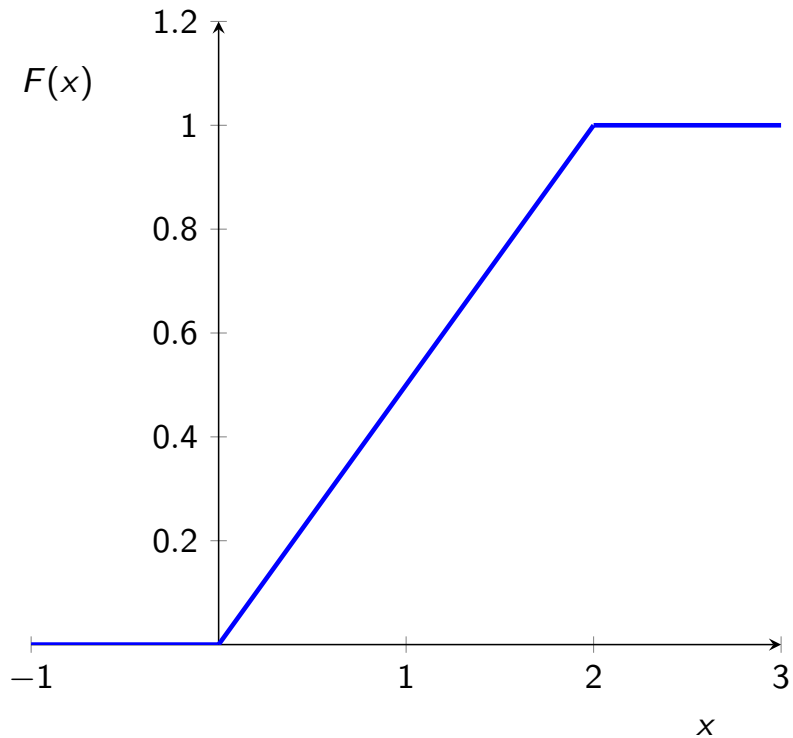
Expected value and variance:

$$\mathbb{E}(X) = \frac{a+b}{2} \qquad \text{and} \qquad \operatorname{Var}(X) = \frac{(a-b)^2}{12}.$$

Uniform distribution

The uniform distribution on $[0, 2]$ has the following probability density function and cumulative distribution function:





Normal distribution

The normal distribution is arguably the most important continuous distribution.

Definition

A continuous random variable X with $W_X = \mathbb{R}$ is called *normally distributed* with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ if it has the probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) =: \varphi(x; \mu, \sigma).$$

We write $X \sim \mathcal{N}(\mu, \sigma^2)$ in this case.

$\mathcal{N}(0, 1)$ is called the standard normal distribution. Its probability density function $\varphi(x; 0, 1)$ is also abbreviated by $\varphi(x)$.

Normal distribution

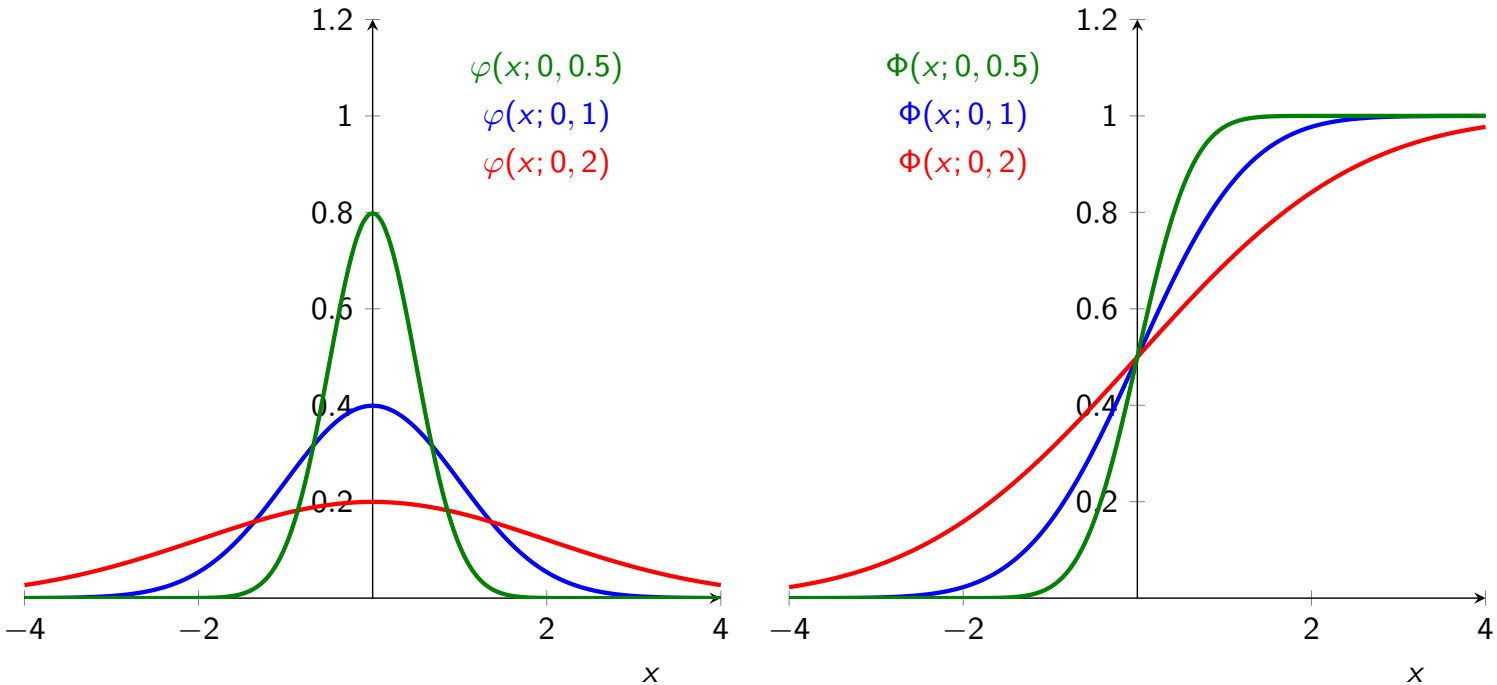
The cumulative distribution function of a $\mathcal{N}(\mu, \sigma^2)$-distributed random variable is

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \mathrm{d}t =: \Phi(x; \mu, \sigma).$$

There is no closed form for the function Φ.

Normal distribution

Probability density function and cumulative distribution function of $\mathcal{N}(0, \sigma^2)$:



Normal distribution

Lemma

$$I := \int_{-\infty}^{\infty} e^{-x^2/2} \mathrm{d}x = \sqrt{2\pi}.$$

Proof.

We compute I^2:

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2/2} \mathrm{d}x \right) \left(\int_{-\infty}^{\infty} e^{-y^2/2} \mathrm{d}y \right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} \mathrm{d}x \mathrm{d}y.$$

We compute this integral by switching to polar coordinates r and ϕ. So, $x = r\cos(\phi)$ and $y = r\sin(\phi)$ and $\mathrm{d}x\mathrm{d}y = r\mathrm{d}r\mathrm{d}\phi$. (We do not go into the details here.)

$$I^2 = \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r \mathrm{d}r \mathrm{d}\phi = \int_0^{2\pi} \left[-e^{-r^2/2} \right]_0^{\infty} \mathrm{d}\phi = \int_0^{2\pi} 1 \mathrm{d}\phi = 2\pi. \qquad \square$$

Normal distribution

Theorem (Linear transformation of the normal distribution)

Let X be a normally distributed random variable with $X \sim \mathcal{N}(\mu, \sigma^2)$. For $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$, define $Y = aX + b$. Then, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Proof.

We show the result for $a > 0$. The case $a < 0$ works analogously.

$$\begin{aligned}\Pr(Y \leq y) &= \Pr(aX + b \leq y) = \Pr\left(X \leq \frac{y-b}{a}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{(y-b)/a} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) \mathrm{d}u.\end{aligned}$$

We substitute $u = (v - b)/a$ and $\mathrm{d}u = (1/a) \cdot \mathrm{d}v$. We obtain

$$\Pr(Y \leq y) = \frac{1}{\sqrt{2\pi}a\sigma} \int_{-\infty}^{y} \exp\left(-\frac{(v - a\mu - b)^2}{2a^2\sigma^2}\right) \mathrm{d}v.$$

So, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. □

Normal distribution

By the theorem, for each $X \sim \mathcal{N}(\mu, \sigma^2)$, the random variable $Y := \frac{X-\mu}{\sigma}$ has a standard normal distribution, i.e., $Y \sim \mathcal{N}(0, 1)$.

In particular, this means that

$$\Pr(a \leq X \leq b) = \Pr\left(\frac{a-\mu}{\sigma} \leq Y \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Normal distribution

Theorem

Let $X \sim \mathcal{N}(0,1)$. Then $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) = 1$.

Proof.

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

As the function under the integral is odd (for all x, $f(x) = -f(-x)$), the integral is 0. So, $\mathbb{E}(X) = 0$.

For the variance, we use integration by parts to observe

$$\sqrt{2\pi} = \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \left[x \cdot \exp\left(-\frac{x^2}{2}\right)\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} x^2 \cdot \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

As $\left[x \cdot \exp\left(-\frac{x^2}{2}\right)\right]_{-\infty}^{\infty} = 0$, we obtain $\mathbb{E}(X^2) = 1$ and hence $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1$. □

Normal distribution

Theorem
Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof.
We know that $Y := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. So, $\mathbb{E}(Y) = 0$ and $\mathrm{Var}(Y) = 1$.

So, we conclude

$$\mathbb{E}(X) = \mathbb{E}(\sigma Y + \mu) = \sigma \cdot \mathbb{E}(Y) + \mu = \mu$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(\sigma Y + \mu) = \sigma^2 \mathrm{Var}(Y) = \sigma^2. \qquad \square$$

Exponential distribution

We have already seen the exponential distribution.

Definition

A random variable X is *exponentially distributed* with parameter $\lambda > 0$, if it has the probability density function

$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The cumulative distribution function for $x \geq 0$ is then

$$F_X(x) = \int_0^x \lambda \cdot e^{-\lambda t} \mathrm{d}t = \left[-e^{-\lambda t} \right]_0^x = 1 - e^{-\lambda x}.$$

For $x < 0$, $F_X(x) = 0$.

In a way, the exponential distribution is the continuous analogue of the geometric distribution. It is particularly useful to model "waiting times".

Exponential distribution

Theorem

Let X be exponentially distributed with parameter $\lambda > 0$. Then,

$$\mathbb{E}(X) = \frac{1}{\lambda} \qquad \textit{and} \qquad \operatorname{Var}(X) = \frac{1}{\lambda^2}.$$

Proof.

We compute, using integration by parts,

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty t \cdot \lambda \cdot e^{-\lambda t} \mathrm{d}t = \left[t \cdot (-e^{-\lambda t})\right]_0^\infty + \int_0^\infty e^{-\lambda t} \mathrm{d}t \\ &= 0 + \left[-\frac{1}{\lambda} \cdot (-e^{-\lambda t})\right]_0^\infty = \frac{1}{\lambda}.\end{aligned}$$

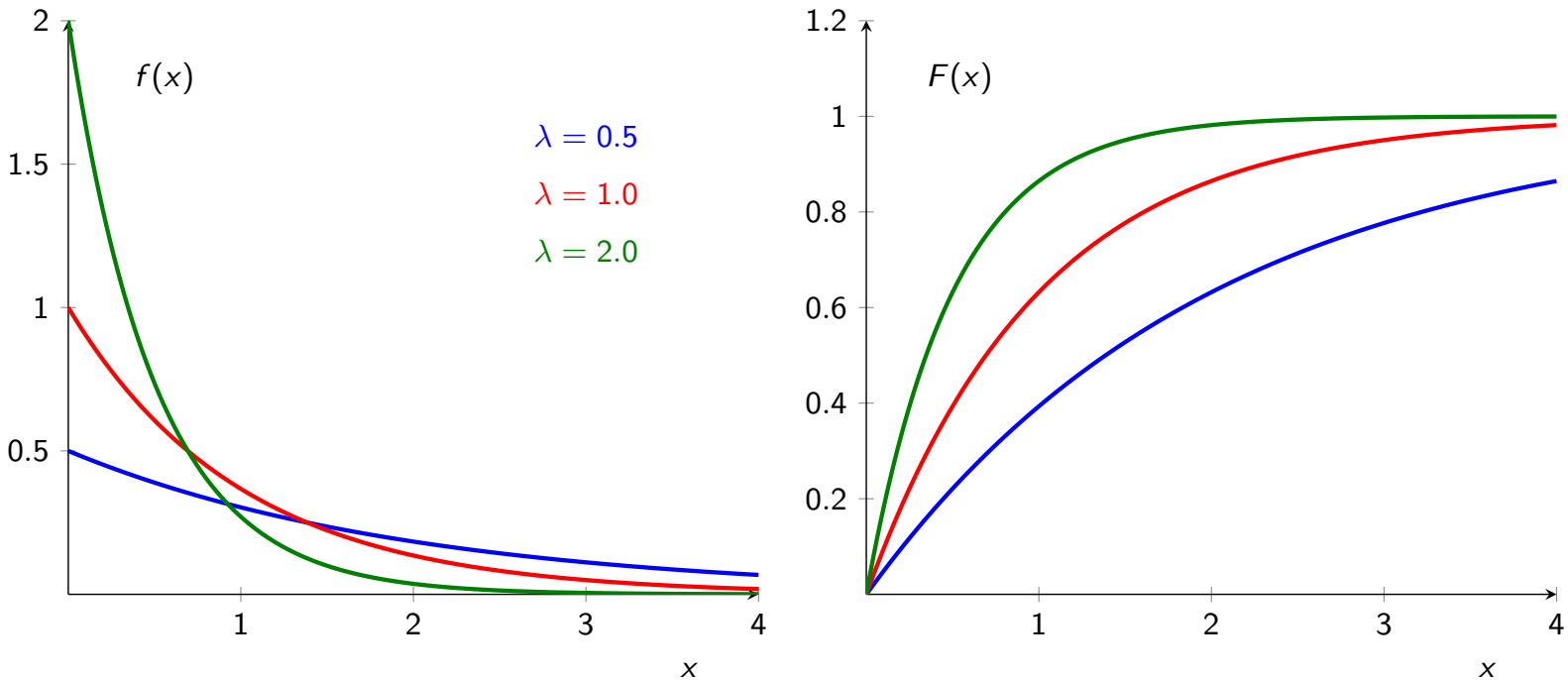
Similarly,

$$\begin{aligned}\mathbb{E}(X^2) &= \int_0^\infty t^2 \cdot \lambda \cdot e^{-\lambda t} \mathrm{d}t = \left[t^2 \cdot (-e^{-\lambda t})\right]_0^\infty + \int_0^\infty 2t \cdot e^{-\lambda t} \mathrm{d}t \\ &= 0 + \frac{2}{\lambda} \cdot \mathbb{E}(X) = \frac{2}{\lambda^2}\end{aligned}$$

and so $\operatorname{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{1}{\lambda^2}$.

Exponential distribution

The probability density function and cumulative distribution function of exponential distributions with different parameters λ:



Exponential distribution

Theorem

Let X be an exponentially distributed random variable with parameter $\lambda > 0$. Let $\alpha > 0$ and define $Y = \alpha X$. Then, Y is exponentially distributed with parameter λ/α.

Proof.

$$\begin{aligned} F_Y(x) &= \Pr(Y \leq x) = \Pr(aX \leq x) \\ &= \Pr\left(X \leq \frac{x}{a}\right) = F_X\left(\frac{x}{a}\right) \\ &= 1 - e^{-\frac{\lambda}{\alpha}x}. \end{aligned}$$

□

Exponential distribution

Like the geometric distribution in the discrete setting, the exponential distribution is *memoryless*.

Theorem

A continuous random variable with value range $\mathbb{R}_{>0}$ is exponentially distributed if and only if for all $x, y > 0$, we have

$$\Pr(X > x + y \mid X > y) = \Pr(X > x). \qquad (*)$$

Proof.

First, we show that $(*)$ holds for exponentially distributed random variables.

Let X be exponentially distributed with parameter $\lambda > 0$. Then,

$$\begin{aligned}\Pr(X > x + y \mid X > y) &= \frac{\Pr(X > x + y, X > y)}{\Pr(X > y)} = \frac{\Pr(X > x + y)}{\Pr(X > y)} \\ &= \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x} = \Pr(X > x).\end{aligned}$$

Exponential distribution

Proof continued.

For the other direction, let X be a continuous random variable with value range $\mathbb{R}_{>0}$ that satsifies $(*)$.

We define $g(x) = \Pr(X > x)$ for $x > 0$. For $x, y > 0$, we obtain that

$$\begin{aligned} g(x+y) &= \Pr(X > x+y) = \Pr(X > x+y \mid X > y) \cdot \Pr(X > y) \\ &= \Pr(X > x) \cdot \Pr(X > y) = g(x)g(y). \end{aligned} \tag{$**$}$$

From this functional equation, we want to conclude that, for some $0 < \gamma < 1$, we have that $g(x) = \gamma^x$ for all $x > 0$. From this, we can then conclude that there is a $\lambda > 0$ such that $\gamma = e^{-\lambda}$ and hence, for all $x > 0$,

$$F_X(x) = \Pr(X \leq x) = 1 - g(x) = 1 - e^{-\lambda x}.$$

Exponential distribution

Proof continued.

We can use that $g(x) = 1 - F_X(x)$ is continuous. Furthermore, as X only takes positive values, there is an $n \in \mathbb{N}$ such that $g(1/n) = \Pr(X > 1/n) > 0$.

Now, we define $\gamma = g(1)$ and observe by $(**)$ that

$$\gamma = g(1) = g(1/n)^n > 0.$$

Now, for every rational p/q with $p, q \in \mathbb{N}$, we have by $(**)$

$$g(p/q) = g(1/q)^p.$$

And as $g(1/q)^q = g(1)$ by $(**)$, we conclude

$$g(p/q) = \gamma^{p/q}.$$

As g is continuous, we obtain that $g(x) = \gamma^x$ for all $x > 0$. □

Exponential distribution

Example

Radioactive decay is random and memoryless. So, we can model the time X until a single particle decays as an exponential distribution.

The half-life t of a radioactive substance describes the time span until half of the particles of a sample decay. In other words, $\Pr(X > t) = \frac{1}{2}$.
The parameter λ of the exponential distribution of X then satisfies

$$\frac{1}{2} = \Pr(X > x) = e^{-\lambda t}.$$

So, $\lambda = \frac{\log(2)}{t}$. For example, the half-life of carbon-14 is 5730 years. So, the time until the decay of a carbon-14 atom is exponentially distributed with parameter

$$\lambda = \frac{\log(2)}{5730 \text{ years}} \approx 1.2 \cdot 10^{-4} \text{ years}^{-1}.$$

The parameter λ can be seen as the *rate* of decay. (Hence, it has the unit years^{-1}.)

Exponential distribution as limit of geometric distributions

Recall that a Poisson distributed random variable X with parameter λ can be seen as the limit for $n \to \infty$ of binomial random variables $Y_n \sim \mathrm{Bin}(n, \lambda/n)$.

Now, let $\lambda > 0$ and consider a sequence of geometric random variables X_n with parameter $p_n := \lambda/n$. Define $Y_n = \frac{1}{n} X_n$.

This can be understood as dividing time into smaller and smaller intervals of length $1/n$. X_n then counts how many intervals pass until the first occurrence of an event that happens in each interval with probability λ/n. The random variable Y_n the returns the point in time at which this interval ends.

For any $k \in \mathbb{N}$, we have

$$\begin{aligned}\Pr(X_n \leq k \cdot n) &= \sum_{i=1}^{kn} (1-p_n)^{i-1} \cdot p_n = p_n \cdot \frac{1-(1-p_n)^{kn}}{p_n} \\ &= 1 - \left(1 - \frac{\lambda}{n}\right)^{kn}.\end{aligned}$$

Exponential distribution as limit of geometric distributions

We obtained

$$\Pr(X_n \leq k \cdot n) = 1 - \left(1 - \frac{\lambda}{n}\right)^{kn}.$$

As $\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$, we obtain for any $t > 0$,

$$\begin{aligned}\lim_{n\to\infty} \Pr(Y_n \leq t) &= \lim_{n\to\infty} \Pr(X_n \leq t \cdot n) \\ &= \lim_{n\to\infty} 1 - \left(1 - \frac{\lambda}{n}\right)^{tn} = 1 - e^{-\lambda t}\end{aligned}$$

So, the sequence of scaled geometric random variables Y_n converges to an exponentially distributed random variable with parameter λ.

Exponential distribution as limit of geometric distributions

	time until next success	# successes in fixed interval
discrete time	geometric X_n/n for X_n geoemtric with parameter λ/n	binomial $Y_n \sim \mathrm{Bin}(n, \lambda/n)$
	$\downarrow n \to \infty$	$\downarrow n \to \infty$
continuous time	exponential X exponentially distributed with parameter λ	Poisson $Y \sim \mathrm{Po}(\lambda)$

Multiple continuous random variables

Definition

For two continuous random variables X and Y, the underlying common probability space on $\mathbb{R}^2$ is given by a *joint probability density function* $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ with

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1.$$

An event $A \subseteq \mathbb{R}^2$ is a set that can be constructed from countably many rectangles of the form $I \times J$ where $I, J \subseteq \mathbb{R}$ are intervals (open, closed, half-open, or unbounded) by taking unions and intersections.

The probability of A is given by

$$\Pr(A) = \int_A f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Multiple continuous random variables

As in the one-dimensional case, we also define a joint cumulative distribution function.

Definition

For two continuous random variables X and Y with joint probability density function $f_{X,Y}$, we define the *joint cumulative distribution function* $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ as

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) \,\mathrm{d}u \,\mathrm{d}v.$$

In the other direction, we can obtain the joint probability density function from the joint cumulative distribution functions by

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

Definition

Let $f_{X,Y}$ be the joint probability density function of two random variables X and Y. The *marginal cumulative distribution function* of X is given by

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} f_{X,Y}(u, v) \, \mathrm{d}v \right) \mathrm{d}u.$$

The *marginal probability density function* of X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, v) \, \mathrm{d}v.$$

For Y, we define these notions analogously.

Multiple continuous random variables

Definition

Two continuous random variables X and Y are called *independent* if

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x) \cdot \Pr(Y \leq y)$$

for all $x, y \in \mathbb{R}$.

This is equivalent to

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y).$$

Taking the derivative, we obtain

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_X(x) \cdot F_Y(y) \\ &= \frac{\partial}{\partial y} \left(\frac{\partial}{\partial x} F_X(x) \cdot F_Y(y) \right) = \frac{\partial}{\partial y} f_X(x) \cdot F_Y(y) \\ &= f_X(x) \cdot f_Y(y). \end{aligned}$$

Multiple continuous random variables

For multiple random variables $X_1, \ldots, X_n$, we say that $X_1, \ldots, X_n$ are independent if

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n)$$

or

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n),$$

respectively, for all $x_1, \ldots, x_n$.

Waiting times with exponential distribution

Theorem

Let $X_1, \ldots, X_n$ be independent, exponentially distributed random variables with parameters $\lambda_1, \ldots, \lambda_n$. Then, also $X := \min\{X_1, \ldots, X_n\}$ is exponentially distributed with parameter $\lambda_1 + \ldots + \lambda_n$.

Proof.

The general statement follows from the case $n = 2$ by induction. So, we prove the statement for $n = 2$:

For the cumulative distribution function F_X, we have

$$\begin{aligned} 1 - F_X(t) &= \Pr(X > t) = \Pr(\min(X_1, X_2) > t) = \Pr(X_1 > t, X_2 > t) \\ &= \Pr(X_1 > t) \cdot \Pr(X_2 > t) = e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} \\ &= e^{-(\lambda_1 + \lambda_2)t}. \end{aligned}$$

□

Waiting times with exponential distribution

Intuitively, the theorem can be understood as follows:

If we wait for one of n events that occur independently with rates $\lambda_1, \ldots, \lambda_n$, we simply have to add the rates to obtain the rate for waiting for the first event.

E.g., if an atom has a decay rate of λ (with unit s^{-1}), then in a sample of n atoms, the rate with which one of them will decay is $n\lambda$.

Poisson process

Recall that if the time between two successes of some repeated trial is geometrically distributed, then the number of successes in a fixed time interval is binomially distributed.

Now, we take a fixed time interval of one time unit and divide into n discrete time steps (for large n).

If we expect λ successes in this time interval, we can say that the number of discrete time steps until the next time success X_n is a geometric random variable with success probability λ/n. The time until a success is then X_n/n.

The number of successes Y_n is now $\mathrm{Bin}(n, \lambda/n)$-distributed.

For $n \to \infty$, we know that $X_n/n \to X$ where X is exponentially distributed with parameter λ and $Y_n \to Y \sim \mathrm{Po}(\lambda)$.

Poisson process

	time until next success	# successes in fixed interval
discrete time	geometric X_n/n for X_n geoemtric with parameter λ/n	binomial $Y_n \sim \mathrm{Bin}(n, \lambda/n)$
	$\downarrow n \to \infty$	$\downarrow n \to \infty$
continuous time	exponential X exponentially distributed with parameter λ	Poisson $Y \sim \mathrm{Po}(\lambda)$

Poisson process

Assume that we observe some experiment in which "successes" appear from time to time.

Let $T_1, T_2, \ldots$ be positive random variables. We assume that T_i models the time between the $(i-1)$th and the ith success in some experiment.

For a time $t > 0$, we now define the random variable

$$X_t := \max\{n \in \mathbb{N} \mid T_1 + \ldots + T_n \leq t\}.$$

So, X_t counts the number of successes until time t.

In this way, we obtain a family of random variables $(X_t)_{t>0}$. In general (no matter how the random variables X_t are defined), such a family is called a *stochastic process*.

Poisson process

The process $(X_t)_{t>0}$ as defined above is called a *Poisson process* if the random variables $T_1, T_2, \ldots$ are independent exponentially distributed random variables. In fact, one can show

Theorem (without proof)

Let $T_1, T_2, \ldots$ be independent positive random variables and let $(X_t)_{t>0}$ be defined as above. Then, X_t is a Poisson random variable with parameter λt for all $t > 0$ if and only if $T_1, T_2, \ldots$ are exponentially distributed with parameter λ.

Example

Consider a set of tasks that are processed sequentially. The runtime of each task is exponentially distributed with parameter $\lambda = 1/30\text{s}^{-1}$. So each task requires 30s in expectation.

Then, the number of tasks that are completed within the first minute is Poisson-distributed with parameter $t\lambda = 60\text{s} \cdot 1/30\text{s}^{-1} = 2$.

Hence, we can, e.g., compute the probability that at most one task is completed within the first minute:

$$e^{-t\lambda} + t\lambda e^{-t\lambda} \approx 0.41.$$

Sums of random variables

Theorem

Let X and Y be independent continuous random variables and let $Z := X + Y$. The probability density function of Z is

$$f_Z(t) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(t-x) \mathrm{d}x.$$

Proof.

For the cumulative distribution function, we have

$$F_Z(t) = \Pr(Z \leq t) = \Pr(X + Y \leq t) = \int_{A(t)} f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y$$

where $A(t) = \{(x, y) \in \mathbb{R}^2 \mid x + y \leq t\}$.

Sums of random variables

Proof continued.

By the independence of X and Y, we obtain

$$F_Z(t) = \int_{A(t)} f_X(x) \cdot f_Y(y) \mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty} f_X(x) \cdot \left(\int_{-\infty}^{t-x} f_Y(y) \mathrm{d}y \right) \mathrm{d}x.$$

By substituting $z := x + y$ and hence $\mathrm{d}z = \mathrm{d}y$, we get

$$\int_{-\infty}^{t-x} f_Y(y) \mathrm{d}y = \int_{-\infty}^{t} f_Y(z-x) \mathrm{d}z.$$

So,

$$F_Z(t) = \int_{-\infty}^{\infty} f_X(x) \cdot \left(\int_{-\infty}^{t} f_Y(z-x) \mathrm{d}z \right) \mathrm{d}x = \int_{-\infty}^{t} \left(\int_{-\infty}^{\infty} f_X(x) \cdot f_Y(z-x) \mathrm{d}x \right) \mathrm{d}z$$

Taking the derivative, we obtain

$$f_Z(t) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(t-x) \mathrm{d}x.$$ □

Sums of random variables

Theorem (Additivity of the normal distribution)

Let $X_1, \ldots, X_n$ be independent normally distributed random variables with parameters μ_i and σ_i, $1 \leq i \leq n$. Then, for $a_1, \ldots, a_n > 0$,

$$Z := a_1 X_1 + \ldots + a_n X_n$$

is normally distributed with parameters μ and σ, which are given by

$$\mu = \sum_{i=1}^{n} a_i \cdot \mu_i \qquad \text{and} \qquad \sigma^2 = \sum_{i=1}^{n} a_i^2 \cdot \sigma_i^2.$$

Proof.

First, recall that we have already shown that for a normally distributed Y with parameters μ_0 and σ_0 and $a \in \mathbb{R}$, the random variable aY is normally distributed with parameters μ_0 and $a\sigma_0$.

So, for each i, we have that $a_i X_i$ is normally distributed with parameters μ_i and $\sigma_i' = a_i \sigma_i$. Hence $(\sigma_i')^2 = a_i^2 \sigma_i^2$. Therefore, it is sufficient to now prove the statement for $a_1 = \cdots = a_n = 1$.

Proof continued.

Furthermore, we present the proof for $n = 2$. The general case then follows easily by induction.

For $Z = X_1 + X_2$, the previous theorem states that

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X_1}(z-y) \cdot f_{X_2}(y) \mathrm{d}y \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right) \mathrm{d}y \end{aligned}$$

We define

$$v := \frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}$$

and let $\mu := \mu_1 + \mu_2$, $\sigma^2 := \sigma_1^2 + \sigma_2^2$, $v_1 := (z-\mu)/\sigma$, and $v_2^2 := v - v_1^2$.

In this way, we obtain

$$v_2^2 = \frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{(z-\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}.$$

Proof continued.

We conclude

$$v_2 = \frac{y\sigma_1^2 - \mu_2\sigma_1^2 + y\sigma_2^2 - z\sigma_2^2 + \mu_1\sigma_2^2}{\sigma_1\sigma_2\sigma}.$$

For the probability density function, we obtain

$$f_Z(z) = \frac{1}{2\pi\sigma_1\sigma_2} \cdot \exp\left(-\frac{v_1^2}{2}\right) \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{v_2^2}{2}\right) \mathrm{d}y.$$

We substitute $t := v_2$ and $\mathrm{d}t = \frac{\sigma}{\sigma_1\sigma_2}\mathrm{d}y$.
Then,

$$f_Z(z) = \frac{1}{2\pi\sigma} \cdot \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t.$$

As we have shown that the integral at the end evaluates to $\sqrt{2\pi}$, we conclude that $f_Z(z) = \varphi(z; \mu, \sigma)$. □

Moment-generating functions

In the discrete setting, we defined the moment-generating function of a random variable X as

$$M_X(s) = \mathbb{E}(e^{sX}).$$

We can directly transfer this definition to the continuous setting:

Definition (Moment-generating function)

Let X be a continuous random variable. The moment-generating function M_X of X is defined by

$$M_X(s) = \mathbb{E}(e^{sX}).$$

As in the discrete setting, it holds that for independent random variables $X_1, \ldots, X_n$ and $Z := X_1 + \ldots + X_n$ we have

$$M_Z(s) = M_{X_1}(s) \cdot \ldots \cdot M_{X_n}(s).$$

Example

Let U be a uniformly distributed random variable on the interval $[a, b]$. Then

$$\begin{aligned} M_U(s) = \mathbb{E}(e^{sU}) &= \int_a^b e^{tx} \cdot \frac{1}{b-a} \mathrm{d}x \\ &= \left[\frac{e^{tx}}{t(b-1)} \right]_a^b = \frac{e^{tb} - e^{ta}}{t(b-a)}. \end{aligned}$$

Moment-generating functions

Example

Let $X \sim \mathcal{N}(0,1)$ be a standard normally distributed random variable. Then,

$$\begin{aligned}
M_X(s) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx - x^2/2} \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{s^2/2 - s^2/2 + sx - x^2/2} \mathrm{d}x \\
&= e^{s^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(s-x)^2/2} \mathrm{d}x \\
&= e^{s^2/2}
\end{aligned}$$

Moment-generating functions

Example

For $Y \sim \mathcal{N}(\mu, \sigma^2)$, we obtain the following using that $\frac{Y-\mu}{\sigma} \sim \mathcal{N}(0,1)$:

$$\begin{aligned} M_Y(s) &= \mathbb{E}(e^{sY}) = \mathbb{E}(e^{s\mu} \cdot e^{s\sigma \frac{Y-\mu}{\sigma}}) \\ &= e^{s\mu} \cdot \mathbb{E}\left(e^{s\sigma \frac{Y-\mu}{\sigma}}\right) \\ &= e^{s\mu} \cdot M_X(s\sigma) = e^{s\mu + (s\sigma)^2/2}. \end{aligned}$$

Moment-generating functions

We can use moment-generating functions to provide an alternative proof for the following theorem we showed earlier:

Theorem

Additivity of the normal distribution Let $X_1, \ldots, X_n$ be independent normally distributed random variables with parameters μ_i and σ_i, $1 \leq i \leq n$. Then

$$Z := a_1 X_1 + \ldots + a_n X_n$$

is normally distributed with parameters μ and σ, for which we have

$$\mu = \sum_{i=1}^{n} a_i \cdot \mu_i \qquad \text{and} \qquad \sigma^2 = \sum_{i=1}^{n} a_i^2 \cdot \sigma_i^2.$$

Proof.

As we have just seen, for $1 \leq i \leq n$, we have

$$M_{X_i}(s) = e^{s\mu_i + (s\sigma_i)^2/2}.$$

As the random variables are independent, we get

$$\begin{aligned} M_Z(s) &= \mathbb{E}\left(e^{s(a_1X_1+\ldots+a_nX_n)}\right) = \prod_{i=1}^{n} \mathbb{E}\left(e^{(a_it)X}\right) \\ &= \prod_{i=1}^{n} M_{X_i}(a_it) = \prod_{i=1}^{n} e^{a_it\mu_i + (a_it\sigma_i)^2/2} \\ &= e^{s\mu + (s\sigma)^2/2} \end{aligned}$$

where $\mu = a_1\mu_1 + \ldots + a_n\mu_n$ and $\sigma^2 = a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2$. □

Outline

Central limit theorem

The central limit theorem illuminates the exceptional role that the normal distribution plays in statistics.

Informally speaking, it states that the sum (or alternatively average) of more and more independent, identically distributed random variables converges to a normal distribution no matter which distribution these random variables have.

(One restriction is that the random variables are required to have finite expected value and variance.)

Central limit theorem

Theorem (Central limit theorem)

Let $X_1, X_2, \ldots$ be independent, identically distributed random variables. Assume that the expected value and the variance of the X_i exists (is finite) and let these values be μ and $\sigma^2 > 0$.

Let $Y_n := X_1 + \ldots + X_n$ for $n \geq 1$. Then, the random variables

$$Z_n := \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

are **asymptotically standard normally distributed**, *i.e.*, $\lim_{n\to\infty} Z_n \sim \mathcal{N}(0,1)$.

More formally, this statement means that for the cumulative distribution functions F_{Z_n} of Z_n, we have

$$\lim_{n\to\infty} F_{Z_n}(x) = \Phi(x) \qquad \text{for all } x \in \mathbb{R}.$$

Central limit theorem

Proof.

Define $X_i^* := (X_i - \mu)/\sigma$ for $i \in \mathbb{N}$. So, $\mathbb{E}(X_i^*) = 0$ and $\mathrm{Var}(X_i^*) = 1$.

So, for the moment-generating function of Z_n, we can write

$$\begin{aligned} M_Z(t) = \mathbb{E}(e^{tZ}) &= \mathbb{E}\left(e^{t(X_1+\cdots+X_n-n\mu)/\sigma\sqrt{n}}\right) \\ &= \mathbb{E}\left(e^{t(X_1^*+\ldots+X_n^*)/\sqrt{n}}\right) \\ &= \mathbb{E}\left(e^{t\cdot X_1^*/\sqrt{n}} \cdot \ldots \cdot e^{tX_n^*/\sqrt{n}}\right) \\ &= M_{X_1^*}(t/\sqrt{n}) \cdot \ldots \cdot M_{X_n^*}(t/\sqrt{n}). \end{aligned}$$

Now, we consider the Taylor expansion of $h(t) := M_{X_i^*}(t)$ at $t = 0$:

$$h(t) = h(0) + h'(0) \cdot t + h''(0)/2 \cdot t^2 + \mathcal{O}(t^3).$$

As the expected value is a linear function, deriving $\mathbb{E}(e^{tX_i^*})$ is not hard:

$$h'(t) = \mathbb{E}(e^{tX_i^*} \cdot X_i^*) \qquad \text{and} \qquad h''(t) = \mathbb{E}(e^{tX_i^*} \cdot (X_i^*)^2).$$

Proof continued.

So,

$$h'(0) = \mathbb{E}(X_i^*) = 0 \qquad \text{and} \qquad h''(0) = \mathbb{E}((X_i^*)^2) = 1.$$

Hence, the Taylor expansion is $h(t) = 1 + t^2/2 + \mathcal{O}(t^3)$. Therefore,

$$M_Z(t) = h(t\sqrt{n})^n = \left(1 + \frac{t^2}{2n} + \frac{\mathcal{O}(t^3)}{n^{3/2}}\right)^n \to e^{t^2/2} \qquad \text{for } n \to \infty.$$

Note that $e^{t^2/2}$ is the moment-generating function of a standard normally distributed random variable.

One can show that the convergence of the cumulative distribution functions follows from the convergence of the moment-generating functions. (We will, however, not do that here.) So, Z_n converges to a standard normally distributed random variable.

Proof continued.

There is one restriction to the proof presented here: The moment-generating function does not exist for all random variables. Therefore, this proof is incomplete.

For a complete proof, one can consider the *characteristic function* $\tilde{M}_X(t) := \mathbb{E}(e^{itX})$ instead.

□

Central limit theorem

To summarize:

The central limit theorem tells us that the linear combination of many independent, identically distributed random variables is approximately normally distributed.

Central limit theorem

Whenever many independent identically distributed random variables are added, we can approximate the probability of an interval in terms of the cumulative distribution function Φ of the standard normal distribution.

Table 5.1 Area $\Phi(x)$ Under the Standard Normal Curve to the Left of X.

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Picture: Sheldon Ross: A first course in Probability

Example

Assume that the number X of students that enrol in a certain course is Poisson distributed with expected value 100.

There is space for 120 students in the lecture rooms. What is the probability that this is sufficient?

The exact solution is

$$e^{-100} \sum_{i=0}^{120} \frac{100^i}{i!}.$$

We are, however, not able to evaluate this expression.

Example (continued)

Instead, we recall that a random variable $X \sim \mathrm{Po}(100)$ can be seen as the sum of n independent random variables that are Poisson distributed with parameter $100/n$.

We choose $n = 100$. For a random variable $Y \sim \mathrm{Po}(1)$, we know $\mathbb{E}(Y) = 1$ and $\mathrm{Var}(Y) = 1$.

Applying the central limit theorem, we get

$$\Pr(X \leq 120.5) = \Pr\left(\frac{X - 100 \cdot 1}{\sqrt{100} \cdot 1} \leq 2.05\right) \approx \Phi(2.05) \approx 0.9798.$$

Central limit theorem

An important special case of the central limit theorem is the following theorem:

Theorem (De Moivre's limit theorem)

Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with success probability p. Define, for $n \geq 1$,

$$H_n := X_1 + \ldots + X_n.$$

The distribution of the random variables

$$H_n^* := \frac{H_n - np}{\sqrt{np(1-p)}}$$

converges to the standard normal distribution for $n \to \infty$.

Proof.

The claim follows from the central limit theorem as $\mathbb{E}(X_i) = p$ and $\mathrm{Var}(X_i) = p(1-p)$ for all i. □

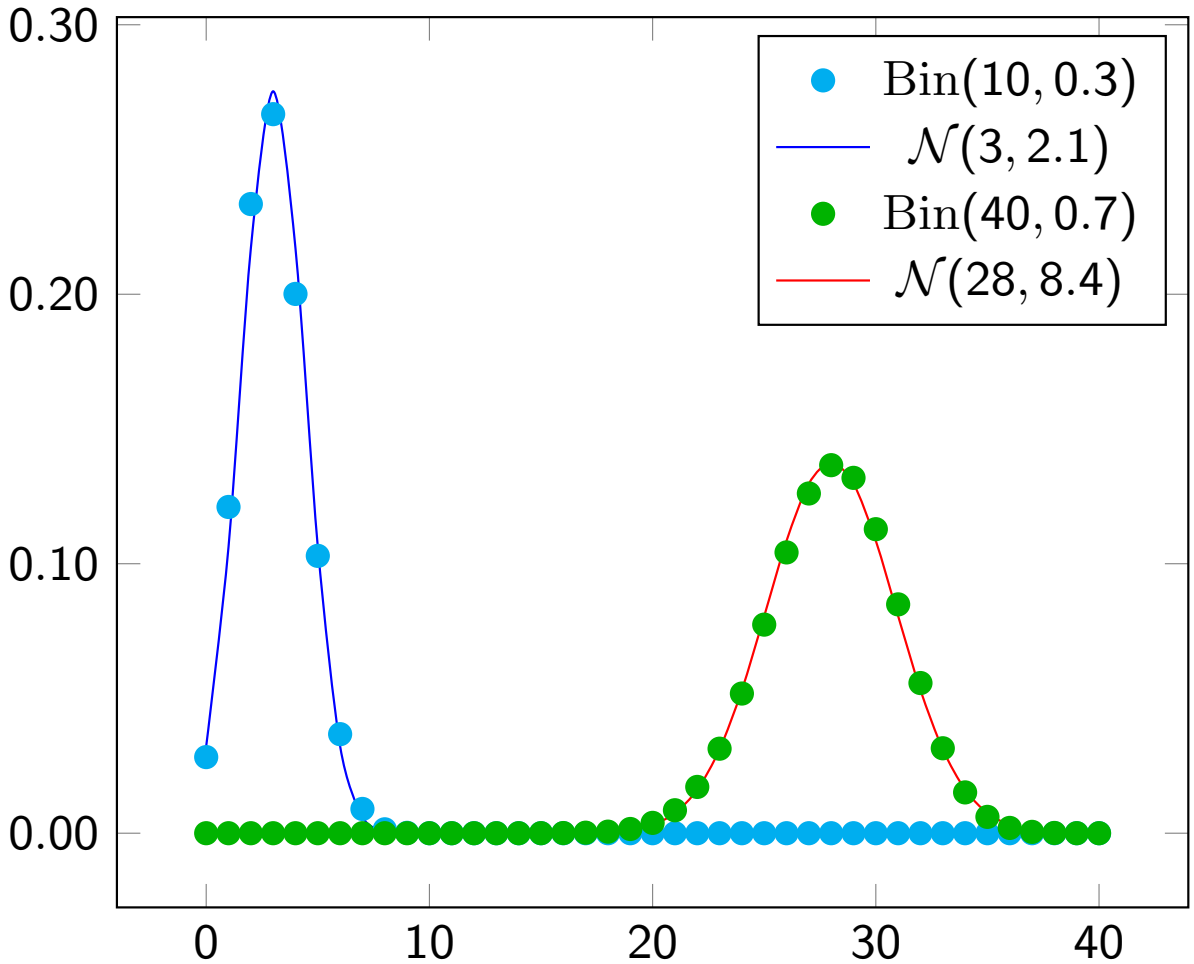
Normal distribution as limit of binomial distributions

From de Moivre's limit theorem, we can conclude:

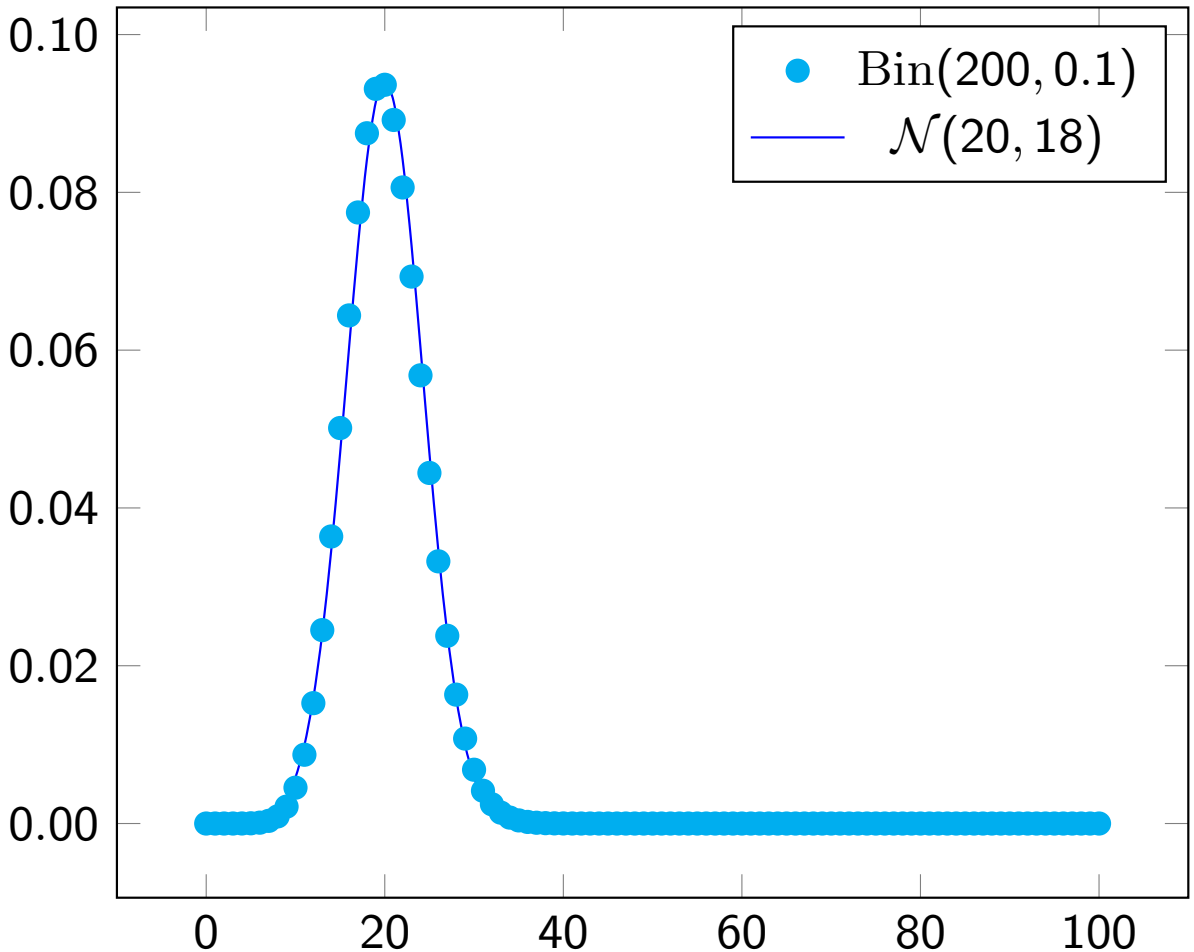
Corollary

Let $H_n \sim \mathrm{Bin}(n, p)$ be binomially distributed. Then, the distribution of H_n/n converges to $\mathcal{N}(p, p(1-p)n)$ for $n \to \infty$.

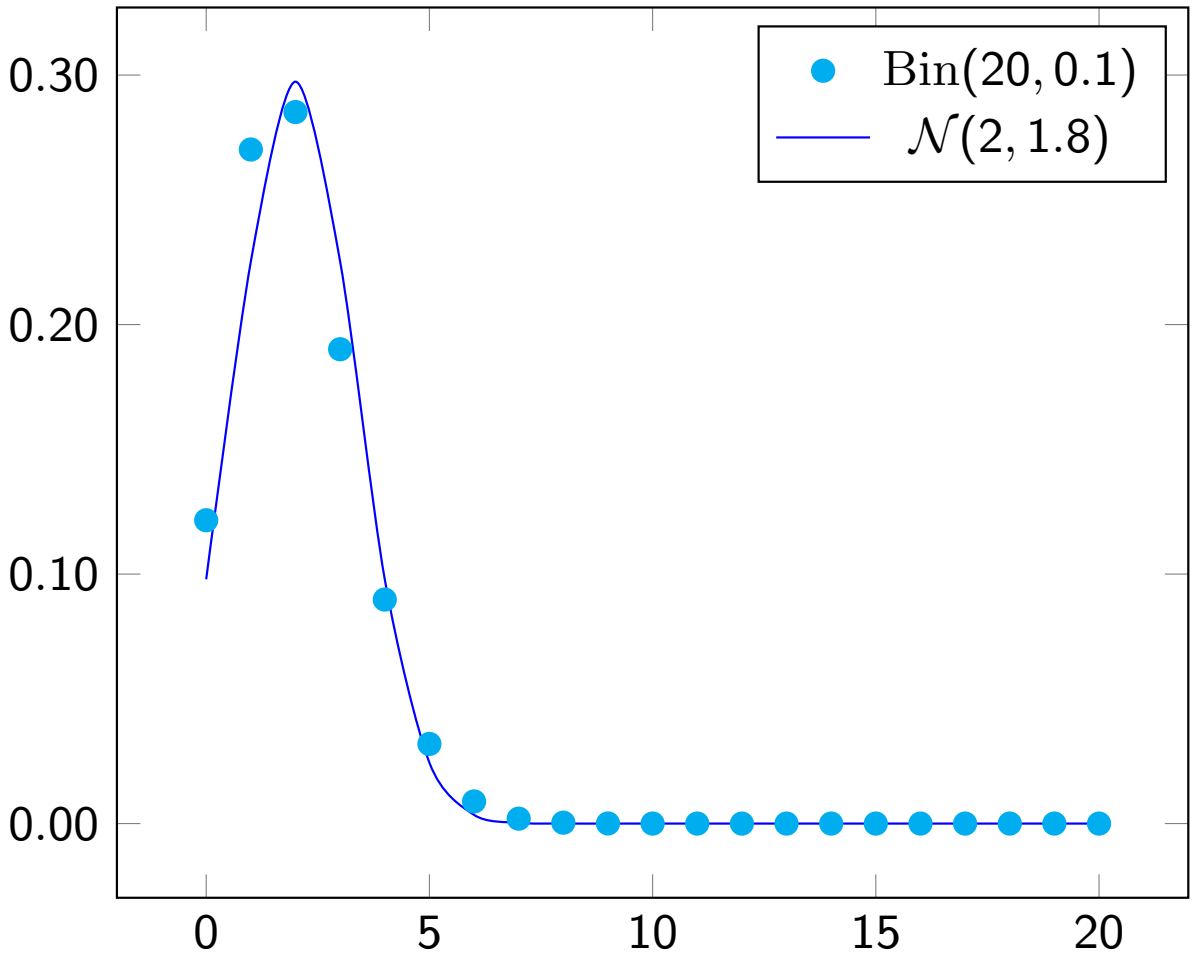
Normal distribution as limit of binomial distributions



Normal distribution as limit of binomial distributions



Normal distribution as limit of binomial distributions



Normal distribution as limit of binomial distributions

Let $H_n \sim \mathrm{Bin}(n, p)$ be a binomial random variable with cumulative distribution function F_n.

For $n \to \infty$, we have shown that

$$F_n(t) = \mathrm{Pr}(H_n \leq t) \to \Phi\left(\frac{t - np}{\sqrt{p(1-p)n}}\right).$$

For large n, it is hence reasonable to use the cumulative distribution function Φ in this way as an approximation for the cumulative distribution function of H_n.

Normal distribution as limit of binomial distributions

Example

Let us say, we want to compute the probability that an even number is rolled at least $500,500$ times when we roll a die $1,000,000$ times.

For the exact probability, we would have to compute

$$T := \sum_{i=500,500}^{1,000,000} \binom{10^6}{i} \left(\frac{1}{2}\right)^{10^6}.$$

This is hardly feasible.

The approximation with the standard normal distribution, on the other hand, requires to evaluate

$$T \approx 1 - \Phi\left(\frac{5.005 \cdot 10^5 - 5 \cdot 10^5}{\sqrt{2.5 \cdot 10^5}}\right) = 1 - \Phi\left(\frac{500}{500}\right) = 1 - \Phi(1) \approx 0.1573.$$

The numerical integration of the probability density function φ is relatively easy. Furthermore, other functions that approximate Φ very well and that are easy to evaluate are known and used in many software libraries.

Normal distribution as limit of binomial distributions

For small values of n, it can be reasonable to add a correction term to the approximation to account for continuity of Φ:

For $X \sim \mathrm{Bin}(n, p)$, one can replace

$$\Pr(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

by

$$\Pr(X \leq x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right).$$

Different approximations of the binomial distribution

- **Normal distribution:** The cumulative distribution function F_n of $\mathrm{Bin}(n, p)$ can be approximated as

$$F_n(x) \approx \Phi\left((x - np)/\sqrt{np(1-p)}\right).$$

  Rule of thumb: for $np \geq 5$ and $n(1-p) \geq 5$, this is a good approximation.
- **Poisson distribution:** $\mathrm{Bin}(n, p)$ is approximated by $\mathrm{Po}(np)$. This works well for rare events, i.e., for small p.
  Rule of thumb: for $p \leq 0.05$ and $n \geq 30$, this is a good approximation.
- **Chernoff bounds:** For $X \sim \mathrm{Bin}(n, p)$, we have for every $\delta > 0$,

$$\Pr(X \geq (1+\delta)np) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np}.$$

  This is useful to estimate the "tails" of the binomial distribution. Furthermore, the bound is a provably correct bound that can be used to prove statements about the probability of events.

Outline

Inferential statistics

In this section, we will cover the basics of

inferential statistics (a.k.a. inductive statistics).

Inferential statistics aims to analyze data samples to infer properties of the underlying probability distribution.

In contrast, descriptive statistics is concerned with properties of the observed data and its understandable presentation, e.g., by computing average values.

Example

We consider the number X of messages sent by a system until the first message is lost.

We assume that X is geometrically distributed as we assume that all messages are lost with the same probability p and that messages losses occur independently.

So, we assume for $i \geq 1$,

$$\Pr(X = i) = (1 - p)^{i-1} p.$$

We do, however, not know the parameter p. We observe n independent copies of the system until their first message loss to obtain n samples of the random variable X.

Our goal is now to estimate the value of p from the sampled data. (Instead of p, we can also estimate the expected value $\mathbb{E}(X)$ as $\mathbb{E}(X) = \frac{1}{p}$.)

Inferential statistics

We have already seen that after repeating an experiment many times, the average result converges to the expected result. The more repetitions of the experiment we consider the more precise and reliable claims can be made about the underlying probability distribution.

Recall in particular the weak law of large numbers and the central limit theorem.

Outline

Estimators

Definition

Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $F(x)$. Then, $X_1, X_2, \ldots, X_n$ is a *random sample* of size n from that probability distribution $F(x)$.

In this context, the underlying probability distribution is sometimes also called *population*.

Estimators

Example

We want to estimate the expected value $\mathbb{E}(X)$ of some random variable X from a random sample $X_1, \ldots, X_n$ from the distribution of X.
In order to empirically determine an estimate for $\mathbb{E}(X)$, we consider the *sample mean*

$$\bar{X} := \frac{\sum_{i=1}^{n} X_i}{n}.$$

We have

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=0}^{n} \mathbb{E}(X_i) = \frac{1}{n} \sum_{i=0}^{n} \mathbb{E}(X) = \mathbb{E}(X). \qquad (*)$$

So, $\bar{X}$ yields, in expectation, the desired value $\mathbb{E}(X)$.

As we use $\bar{X}$ to estimate $\mathbb{E}(X)$, we call $\bar{X}$ an *estimator* for $\mathbb{E}(X)$.

Because of $(*)$, we call $\bar{X}$ an *unbiased* estimator.

Estimators

Definition

Let X be a random variable with probability density $f(x;\theta)$. An *estimator* for the parameter θ is a random variable U depending on a random sample $X_1, \ldots, X_n$ from the distribution $f(x;\theta)$.

An estimator U for θ is called *unbiased* if

$$\mathbb{E}(U) = \theta.$$

The value $\mathbb{E}(U - \theta)$ is called *bias* of the estimator U. For unbiased estimators, the bias is zero.

Definition

Let U be an estimator for a parameter θ. The *mean squared error (MSE)* of U is defined as

$$\mathrm{MSE}(U) := \mathbb{E}((U-\theta)^2).$$

If U is unbiased,

$$\mathrm{MSE}(U) = \mathbb{E}((U-\mathbb{E}(U))^2) = \mathrm{Var}(U).$$

Example

For the estimator $\bar{X}$ for $\mathbb{E}(X)$, we obtain, using the independence of $X_1, \ldots, X_n$,

$$\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}\mathrm{Var}(X).$$

Definition

Let A and B be estimators for a parameter θ. The estimator A is called *more efficient* than the estimator B if $\mathrm{MSE}(A) < \mathrm{MSE}(B)$.

An estimator U for θ is called *consistent in mean square* if $\mathrm{MSE}(U) \to 0$ for $n \to \infty$ where n is the size of the random sample from which U is formed.

Example

Let X be a random variable with finite variance.

The estimator $\bar{X}$ for $\mathbb{E}(X)$ is consistent in mean square because

$$\mathrm{MSE}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X)$$

tends to 0 for $n \to \infty$.

Estimators

An estimator U for a parameter θ is called weakly consistent if, for all $\varepsilon > 0$,

$$\Pr\left(|U - \theta| \geq \varepsilon\right) \to 0$$

for $n \to \infty$.

For the estimator $\bar{X}$ for $\mathbb{E}(X)$, we obtain

$$\Pr r\left(\left|\bar{X} - \mathbb{E}(X)\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(\bar{X})}{\varepsilon^2}$$

using Chebyshev's inequality and the fact that $\mathbb{E}(\bar{X}) = \mathbb{E}(X)$.

If we assume that $\mathrm{Var}(X) < \infty$, we have seen that $\bar{X}$ is consistent in mean square. So, $\mathrm{Var}(\bar{X}) \to 0$ for $n \to \infty$ and hence $\bar{X}$ is also weakly consistent.

Estimators

Consider the following estimator S:

$$S := \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(X_i - \bar{X})^2}.$$

We want to compute the expected value of S^2 to show that S^2 is an unbiased estimator for $\mathrm{Var}(X)$.

Let $\mu = \mathbb{E}(X) = \mathbb{E}(\bar{X}) = \mathbb{E}(X_i)$. Then, we first compute

$$\begin{aligned}
(X_i - \bar{X})^2 &= (X_i - \mu + \mu - \bar{X})^2 \\
&= (X_i - \mu)^2 + (\mu - \bar{X})^2 + 2(X_i - \mu)(\mu - \bar{X}) \\
&= (X_i - \mu)^2 + (\mu - \bar{X})^2 - 2(X_i - \mu)\frac{\sum_{j=1}^{n} X_j - \mu}{n} \\
&= \frac{n-2}{n}(X_i - \mu)^2 + (\mu - \bar{X})^2 - \frac{2}{n}\sum_{j\neq i}(X_i - \mu)(X_j - \mu).
\end{aligned}$$

Estimators

We have shown

$$(X_i - \bar{X})^2 = \frac{n-2}{n}(X_i - \mu)^2 + (\mu - \bar{X})^2 - \frac{2}{n}\sum_{j \neq i}(X_i - \mu)(X_j - \mu).$$

As X_i and X_j are independent for $j \neq i$, we get

$$\mathbb{E}((X_i - \mu)(X_j - \mu)) = \mathbb{E}(X_i - \mu) \cdot \mathbb{E}(X_j - \mu) = 0 \cdot 0 = 0.$$

So,

$$\begin{aligned}\mathbb{E}\left((X_i - \bar{X})^2\right) &= \frac{n-2}{n}\mathbb{E}\left((X_i - \mu)^2\right) = \mathbb{E}\left((\mu - \bar{X})^2\right) \\ &= \frac{n-2}{n} \cdot \operatorname{Var}(X_i) + \operatorname{Var}(\bar{X}).\end{aligned}$$

Estimators

We derived

$$\mathbb{E}\left((X_i - \bar{X})^2\right) = \frac{n-2}{n} \cdot \mathrm{Var}(X_i) + \mathrm{Var}(\bar{X}).$$

Because $\mathrm{Var}(X_i) = \mathrm{Var}(X)$ and $\mathrm{Var}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X)$, we obtain

$$\mathbb{E}\left((X_i - \bar{X})^2\right) = \frac{n-1}{n} \cdot \mathrm{Var}(X).$$

So, for S^2, we indeed obtain

$$\mathbb{E}(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}\left((X_i - \bar{X})^2\right) = \frac{1}{n-1} \cdot n \cdot \frac{n-1}{n} \cdot \mathrm{Var}(X) = \mathrm{Var}(X).$$

So, S^2 is indeed an unbiased estimator for $\mathrm{Var}(X)$.

Estimators

Note that the computations show that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is not an unbiased estimator for the variance.

Recall that

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Estimators

Definition

The random variables

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are called sample mean and sample variance of the random sample $X_1, \ldots, X_n$.

As we have seen, $\bar{X}$ is an unbiased estimator for the expected value and S^2 is an unbiased estimator for the variance.

Maximum likelihood estimators

Maximum likelihood estimation is a method to construct estimators for a parameter of a distribution.

Let $\vec{X} = (X_1, \ldots, X_n)$ be a random sample from a distribution with probability mass function (discrete case) or probability density function (continuous case) $f(x; \theta)$. We want to estimate the parameter θ from the values $\vec{x} = (x_1, \ldots, x_n)$ that the random variables $\vec{X}$ take.

Maximum likelihood estimators

In the discrete case, the idea is now to choose θ such that the probability that the random variables $\vec{X} = (X_1, \ldots, X_n)$ take the values $\vec{x} = (x_1, \ldots, x_n)$ is maximized.

This probability is

$$L(\vec{x}; \theta) := \prod_{i=1}^{n} f(X_i; \theta) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$$

where the last equality follows from the independence of the variables $\vec{X}$. The function $L(\vec{x}, \theta)$ is called the *likelihood function*.

To emphasize that the probability depends on θ, the notation $\Pr_\theta(X_1 = x_1, \ldots, X_n = x_n)$ is sometimes used.

Analogously to the discrete setting, also in the continuous setting the likelihood function is defined as

$$L(\vec{x}, \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Maximum likelihood estimators

In both cases, the goal is now to maximize the likelihood function by estimating the parameter θ.

Definition

An estimator $\hat{\theta}$ for a parameter θ of a distribution $f(x;\theta)$ is called *maximum likelihood estimator* if for any $\vec{x}$,

$$L(\vec{x};\theta) \leq L(\vec{x},\hat{\theta}) \qquad \text{for all } \theta.$$

Maximum likelihood estimators

Example

Consider a Bernoulli distribution $f(x;p)$ with parameter p. So, $f(0;p) = 1-p$ and $f(1;p) = p$. The random sample $\vec{X} = (X_1, \ldots, X_n)$ consists of n independent random variables with probability mass function $f(x;p)$.

Now, assume $\vec{X}$ takes the values $\vec{x} = (x_1, \ldots, x_n)$. Which parameter p makes this outcome most likely?

Observe that we can write $\Pr_p(X_i = x_i) = p^{x_i} \cdot (1-p)^{1-x_i}$ as $x_i \in \{0,1\}$. So,

$$L(\vec{x};p) = \prod_{i=1}^{n} p^{x_i} \cdot (1-p)^{1-x_i}.$$

Example (continued)

Instead of L, we will maximize $\log \circ L$ and obtain

$$\begin{aligned}\log(L(\vec{x}; p)) &= \sum_{i=1}^{n} (x_i \log(p) + (1 - x_i) \log(1 - p)) \\ &= n\bar{x} \log(p) + (n - n\bar{x}) \log(1 - p)\end{aligned}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ denotes the arithmetic mean.

We obtain the maximum by finding the zero of the derivative, i.e., by solving:

$$\frac{\partial \log(L(\vec{x}; p))}{\partial p} = \frac{n\bar{x}}{p} - \frac{n - n\bar{x}}{1 - p} = 0.$$

The solution is $p = \bar{x}$.

Maximum likelihood estimators

Example

Let $X_1, \ldots, X_n$ be random samples from a uniform distribution on $[0, \theta]$.

Suppose the samples are 1.3, 3.2, 4.1, and 8.2. What would be a good estimate for the parameter θ?

The maximum likelihood estimator just estimates the maximum number seen, i.e. 8.2, for the parameter θ.

$$X_{\max} = \max(X_1, \ldots, X_n).$$

Example (continued)

For which factor k is $k \cdot X_{\max}$ an unbiased estimator for θ?

Let us compute $\mathbb{E}(X_{\max})$. We first compute the cumulative distribution function $F_{\max}(x;\theta)$ of $X_{\max}$. For $x \in [0,\theta]$, we have

$$F_{\max}(x;\theta) = \Pr(X_{\max} \leq x) = \Pr(X_1 \leq x, ..., X_n \leq x) = \prod_{i=1}^{n} \Pr(X_i \leq x) = \left(\frac{x}{\theta}\right)^n.$$

For $x < 0$, we have $F_{\max}(x;\theta) = 0$ and, for $x \geq \theta$, we have $F_{\max}(x;\theta) = 1$.

By taking the derivative, we obtain the probability density function $f_{\max}(x;\theta)$ of $X_{\max}$:

$$f_{\max}(x;\theta) = \frac{\mathrm{d}F_{\max}(x;\theta)}{\mathrm{d}x} = \begin{cases} 0 & \text{for } x < 0, \\ n \cdot x^{n-1} \cdot \frac{1}{\theta^n} & \text{for } x \in (0,\theta), \\ 0 & \text{for } x > \theta. \end{cases}$$

Example (continued)

Now, we can compute $\mathbb{E}(X_{\max})$:

$$\begin{aligned}\mathbb{E}(X_{\max}) &= \int_0^\theta x \cdot f_{\max}(x;\theta)\mathrm{d}x = \int_0^\theta x \cdot n \cdot x^{n-1} \cdot \frac{1}{\theta^n}\mathrm{d}x \\ &= \int_0^\theta n \cdot \frac{x^n}{\theta^n}\mathrm{d}x = \left[\frac{n}{n+1}\frac{x^{n+1}}{\theta^n}\right]_0^\theta \\ &= \frac{n}{n+1}\theta.\end{aligned}$$

So, $\frac{n+1}{n}X_{\max}$ is an unbiased estimator for θ.

Example

Now, consider a Poisson distribution $f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$ with parameter λ and a random sample $\vec{X} = (X_1, \ldots, X_n)$ from this distribution.

Now, $\vec{X}$ takes the values $\vec{x} = (x_1, \ldots, x_n)$. The likelihood function is

$$L(\vec{x};\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \frac{e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

Again, we compute $\log \circ L$:

$$\log(L(\vec{x};\lambda)) = -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

Example (continued)

Setting the derivative equal to zero, we obtain

$$\frac{\partial \log(L(\vec{x}; \lambda))}{\partial \lambda} = -n + \sum_{i=1}^{n} \frac{x_i}{\lambda} = 0.$$

We obtain

$$\lambda = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}.$$

As λ is the expected value of the Poisson distribution, the sample mean does certainly seem like a reasonable estimator for λ.

Maximum likelihood estimators

Example

Now, we will apply the maximum likelihood principle to estimate two parameter simultaneously.

Let $X \sim \mathcal{N}(\mu, \sigma)$. We want to estimate the parameters μ and σ from a random sample from the distribution of X. For n values $\vec{x}$, we have

$$L(\vec{x}; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \cdot \prod_{i=1}^{n} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

Taking the logarithm, we get

$$\log L(\vec{x}; \mu, \sigma^2) = -n(\log \sqrt{2\pi} + \log \sigma) + \sum_{i=1}^{n} \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right).$$

Example (continued)

Setting the derivative to 0, we obtain the following two equations:

$$\frac{\partial \log L(\vec{x}; \mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0.$$

$$\frac{\partial \log L(\vec{x}; \mu, \sigma^2)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} = 0.$$

Solving these equations, we obtain

$$\mu = \bar{x} \qquad \text{and} \qquad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

So, the maximum likelihood estimator for μ is the sample mean.

Example (continued)

The maximum likelihood estimator for σ^2, however, is not the sample variance, which would be

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2.$$

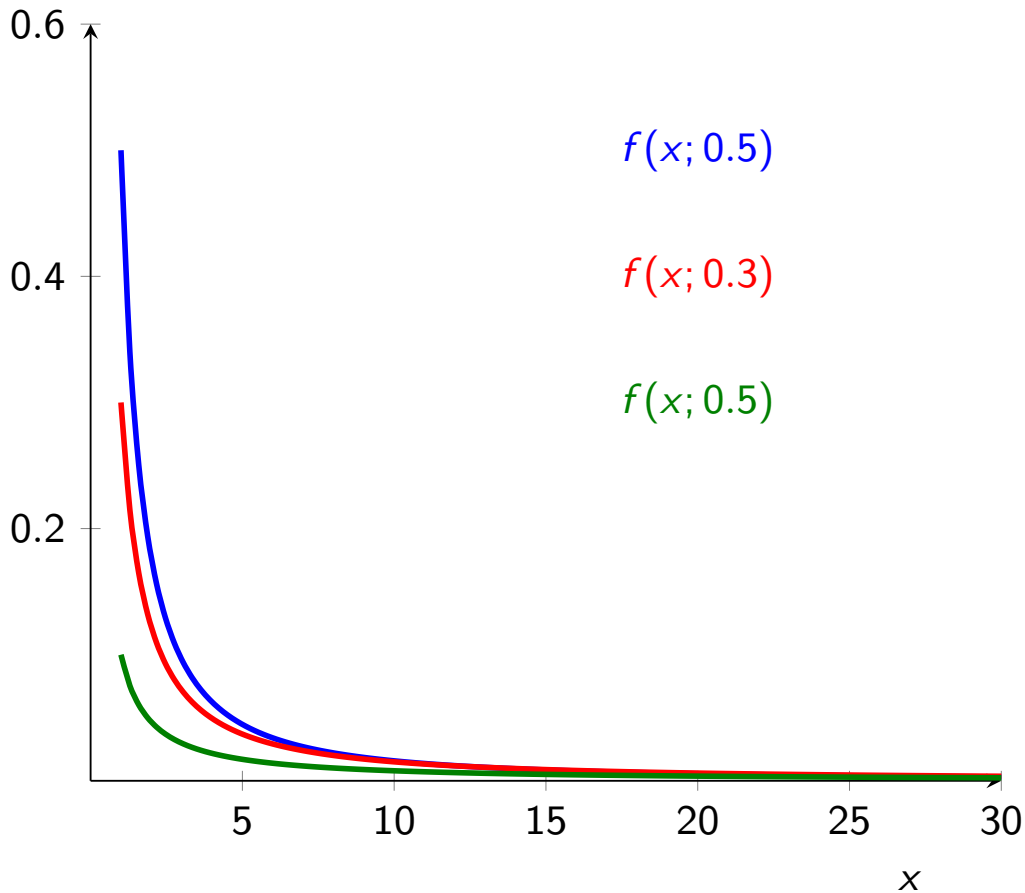
In particular, the maximum likelihood estimator for σ^2 is biased.

Maximum likelihood estimators

Example

Suppose, we know that a random variable has a probability density function of the form

$$f(x;\theta) = \begin{cases} \frac{\theta}{x^{\theta+1}} & \text{for } x > 1, \\ 0 & \text{otherwise.} \end{cases}$$



Example (continued)

Suppose we sample from this distribution and obtain values 2.1, 3.2, 5, 13.5, and 20.

What is the maximum likelihood estimation for θ?

The likelihood function for $\vec{x} = (x_1, \ldots, x_n) \in [1, \infty)^n$ is given by

$$L(\vec{x}; \theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^{\theta+1}} = \frac{\theta^n}{(\prod_{i=1}^{n} x_i)^{\theta+1}}.$$

Example (continued)

Applying the logarithm, we obain

$$\log L(\vec{x}; \theta) = n \log(\theta) - (\theta + 1) \sum_{i=1}^{n} \log(x_i).$$

Taking the derivative, we get

$$\frac{\partial \log L(\vec{x}; \theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} \log(x_i).$$

Setting this derivative to 0, we get

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log(x_i)}.$$

The second derivative is $-n/\theta^2$ and so $\log \circ L$, and hence L, obtains its maximum at $\hat{\theta}$.

Maximum likelihood estimators

Example (continued)

For our sample, we get

$$\hat{\theta} = \frac{5}{\log(2.1) + \log(3.2) + \log(5) + \log(13.5) + \log(20)} \approx 0.55.$$

Outline

Confidence intervals

Whether an estimator is *unbiased* or *consistent in mean square* are qualitative statements about the estimator.

We know that the larger the sample size, the better the estimation will be. However, this is unsatisfactory if we want to know how reliable a concrete estimation is.

To overcome this shortcoming, a solution is to provide two estimators U_1 and U_2 for a parameter θ such that, for a fixed α,

$$Pr(U_1 \leq \theta \leq U_2) \geq 1 - \alpha.$$

The probability $1 - \alpha$ is called *confidence level* and can be adjusted.

The interval $[U_1, U_2]$ is called the *confidence interval*.

Often, a confidence interval is constructed from one estimator U by choosing the interval $[U - \delta, U + \delta]$ for a suitable δ.

Confidence intervals

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and let $X_1, \ldots, X_n$ be a random sample from the distribution of X.

By the additivity of the normal distribution, the sample mean $\bar{X}$ is also normally distributed with

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

We want to derive a symmetric confidence interval with confidence level $1 - \alpha$ for the mean μ from $\bar{X}$.

We know that

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \cdot \frac{\bar{X} - \mu}{\sigma}$$

is standard normally distributed.

Confidence intervals

We will choose a value c such that

$$\Pr(-c \leq Z \leq c) = 1 - \alpha.$$

We can rewrite

$$\Pr(-c \leq Z \leq c) = \Pr\left(-c \leq \sqrt{n} \cdot \frac{\bar{X} - \mu}{\sigma} \leq c\right) = \Pr\left(\bar{X} - \frac{\sigma c}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{\sigma c}{\sqrt{n}}\right).$$

After we have chosen the right value for c, the confidence interval will be

$$C = \left[\bar{X} - \frac{\sigma c}{\sqrt{n}}, \bar{X} + \frac{\sigma c}{\sqrt{n}}\right].$$

Confidence intervals

It remains to choose c such that

$$\Pr(-c \leq Z \leq c) = \Phi(c) - \Phi(-c) = 1 - \alpha$$

where Φ is the cumulative distribution function of the standard normal distribution.

Due to the symmetry of Φ, we have $\Phi(-c) = 1 - \Phi(c)$ and we obtain

$$1 - \alpha = 2 \cdot \Phi(c) - 1.$$

This is equivalent to

$$\Phi(c) = 1 - \frac{\alpha}{2}$$

and so we choose

$$c := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Confidence intervals

Definition

Let X be a continuous random variable with cumulative distribution function F_X. The least value x_γ with

$$F_X(x_\gamma) = \gamma$$

is called γ-*quantile* of X or of F_X, respectively.

Definition

For the standard normal distribution, the γ-quantile is denoted by z_γ.

Confidence intervals

Table 5.1 Area $\Phi(x)$ Under the Standard Normal Curve to the Left of X.

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Picture: Sheldon Ross: A first course in Probability

Confidence intervals

Applying this to the confidence interval from the previous example, we can now write

$$C = \left[\bar{X} - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right].$$

So, with probability $1 - \alpha$, the true mean μ of the sampled normal distribution lies within the interval C.

Estimating the mean

In general, for any distribution with finite expected value μ and finite variance σ^2, the following argument can be used:

For large sizes n of random samples $X_1, \ldots, X_n$, the sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is approximately normally distributed with expected value μ and variance σ^2/n.

Let

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

For a desired level of confidence $1 - \alpha$, we obtain that

$$\Pr(z_{\alpha/2} < Z < z_{1-\alpha/2}) \approx 1 - \alpha.$$

Estimating the mean

Reformulating, we obtain

$$\Pr\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1 - \alpha.$$

So, if a random sample returns $\vec{X} = \vec{x}$ and we know the variance σ^2 of the distribution, a confidence interval for μ with confidence level $1 - \alpha$ is given by

$$C = \left[\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Due to the symmetry of the standard normal distribution, we can write

$$C = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Estimating the mean

Example

The average zinc concentration of 36 measurements taken from different locations of a river is $2.60 \frac{\mathrm{mg}}{\mathrm{l}}$.

We assume a standard deviation of $0.3 \frac{\mathrm{mg}}{\mathrm{l}}$.

Then, a 95%-confidence interval ($\alpha = 0.05$) for the average zinc concentration is given by

$$C = \left[2.60 - z_{1-\alpha/2} \cdot \frac{0.3}{\sqrt{36}}, 2.60 + z_{1-\alpha/2} \cdot \frac{0.3}{\sqrt{36}}\right].$$

From a table for Φ, we obtain $z_{1-0.05/2} \approx 1.96$. So,

$$C = [2.50, 2.70].$$

For $\alpha = 0.01$, we get $z_{1-0.01/2} \approx 2.575$ and hence

$$C = [2.47, 2.73].$$

Estimating the mean

We might also want to be $(1-\alpha)$-confident that the error of our estimation for μ is not larger than some bound e.

From the confidence interval

$$C = \left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

we can compute the necessary sample size.

We want

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq e.$$

Solving for n, we obtain

$$n \geq \left(\frac{z_{1-\alpha/2}\sigma}{e}\right)^2.$$

If we want to make the error bound half as big, we have to increase the sample size by a factor of 4.

Estimating the mean

Example

In the previous example about the zinc concentration in a river, we now want to make sure that our estimation is off by less than $0.05 \frac{\text{mg}}{\text{l}}$ with probability 95%. How many samples do we need?

Plugging in, we get

$$n \geq \left(\frac{z_{1-\alpha/2}\sigma}{e} \right)^2 = \left(\frac{1.96 \cdot 0.3}{0.05} \right)^2 \approx 138.3.$$

Unknown variance

Often, we do not know the variance of the distribution from which we sample.

We know, however, that the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

is an unbiased estimator for the variance.

We might want to try to replace σ by $S = \sqrt{S^2}$ in the previous approach to obtain confidence intervals.

Unknown variance

Assume, $X_1, \ldots, X_n$ is a random sample from a normal distribution with mean μ and (unknown) variance σ^2.

Define

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Definition

The distribution of T is called *t-distribution* with $n-1$ degrees of freedom. This distribution is also known as *Student's t-distribution*.

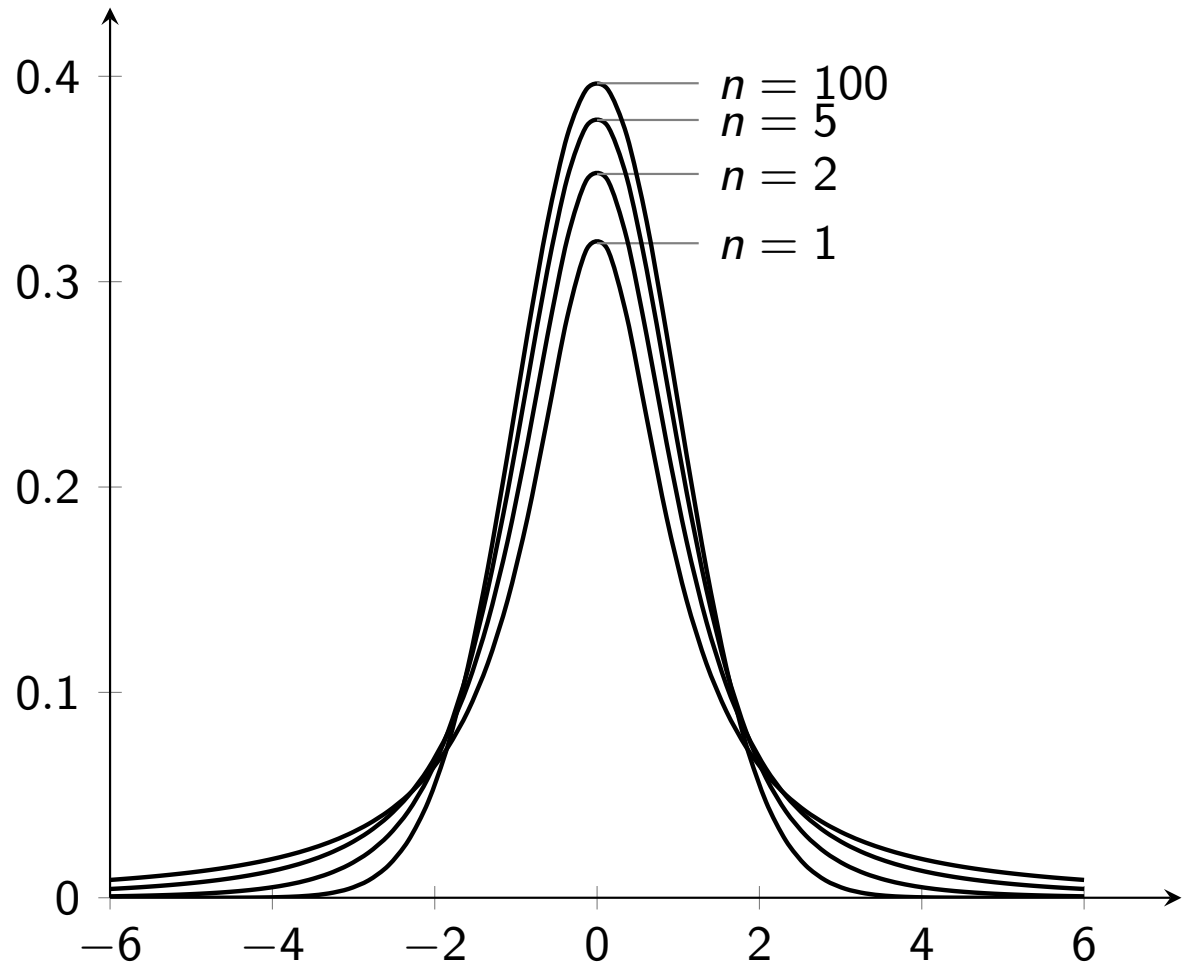
The quantiles of the t-distribution are denoted by t_α. I.e., t_α is the unique value such that

$$\Pr(T \leq t_\alpha) = \alpha.$$

The t-distribution is symmetric around 0. So, $t_\alpha = -t_{1-\alpha}$ for all α.

Unknown variance

The probability density function of the t-distribution:



For $n \to \infty$, the t-distribution converges to the standard normal disitrbution.

Unknown variance

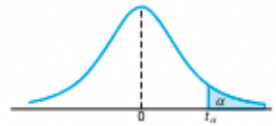


Table A.4 Critical Values of the t-Distribution

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Picture: Walpole, Myers, Myers, Ye: Probability and Statistics for Engineers and Scientists (Ninth edition)

Unknown variance

When sampling from a normal distribution, a confidence interval for μ with level of confidence $1 - \alpha$ is given by

$$C = \left[\bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2} \frac{S}{\sqrt{n}} \right].$$

Note that $t_{1-\alpha/2}$ depends on the size n of the random sample.

Large-sample confidence interval

Even when the random sample is taken from a non-normal distribution and when σ is unknown, we can reasonably replace σ by the square root $S = \sqrt{S^2}$ of the sample variance for large n.

In fact, already for $n \geq 30$, this is usually fine.

So, if a sample returns $\vec{x}$, we can provide the following confidence interval for μ with level of confidence (approximately) $1 - \alpha$:

$$C = \left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

where

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

This is known as a *large-sample confidence interval*. Intuitively, this works well because, for large n, the sample standard deviation s is close to the real standard deviation σ (with high probability).

Prediction interval

Confidence intervals provide good information about an unknown parameter.

Sometimes, we might be interested in the future value of a single observation.

Example

Suppose, we measure the sizes of certain parts manufactured in a factory. After many (or at least several) samples, we are able to determine a confidence interval C with level of confidence $1-\alpha$ for the average size μ.

Now, a customer wants to buy a single part. For the customer, the interval C is not useful.

Instead, the customer would like to be $(1-\alpha)$-sure that the size of the bought part will lie within an interval P.

If we knew the exact distribution of sizes, we could compute such an interval.

But, we can only obtain some information about the real distribution from the random sample we observed.

Prediction interval, normal distribution with known variance

First, suppose the underlying probability distribution is a normal distribution with *unknown* mean μ and *known* variance σ^2.

From a random sample $\vec{X}$ of size n, we obtain the sample mean $\bar{X}$. We know that the variance of the sample mean is σ^2/n.

Let X_0 be a new observation. Then, $X_0 - \bar{X}$ is also normally distributed.

$$\mathbb{E}(X_0) = \mathbb{E}(\bar{X}) = \mu.$$

So, $\mathbb{E}(X_0 - \bar{X}) = 0$. But what is the variance?

As X_0 and $\bar{X}$ are independent, the variance is

$$\mathrm{Var}(X_0 - \bar{X}) = \mathrm{Var}(X_0) + \mathrm{Var}(\bar{X}) = \sigma^2 + \sigma^2/n.$$

Prediction interval, normal distribution with known variance

We conclude that

$$Z := \frac{X_0 - \bar{X}}{\sqrt{\sigma^2 + \sigma^2/n}} = \frac{X_0 - \bar{X}}{\sigma\sqrt{1 + 1/n}}$$

is standard normally distributed.

Hence,

$$1 - \alpha = \mathsf{Pr}\left(1 - z_{1-\alpha/2} < Z < z_{1-\alpha/2}\right).$$

We conclude

$$1 - \alpha = \mathsf{Pr}\left(\bar{X} - z_{1-\alpha/2}\sigma\sqrt{1 + 1/n} < X_0 < \bar{X} + z_{1-\alpha/2}\sigma\sqrt{1 + 1/n}\right).$$

So, the prediction interval, into which a new observation will fall with probability $1 - \alpha$ given the observed sample $\vec{x}$ of size n, is

$$P = \left[\bar{x} - z_{1-\alpha/2}\sigma\sqrt{1 + 1/n}, \bar{x} + z_{1-\alpha/2}\sigma\sqrt{1 + 1/n}\right].$$

Prediction interval, normal distribution with known variance

Example

A bank receives mortgage applications. A recent sample of 50 mortgage loans resulted in an average loan amount of $\$257,300$. A research institute found out that the loan amount is normally distributed with standard deviation $\$25,000$.

What is a 95%-prediction interval for the future loan of a new customer?

From a table, we get $z_{0.975} = 1.96$. Plugging in the remaining values, we obtain

$$\begin{aligned} P &= \left[257300 - 1.96 \cdot 25000 \cdot \sqrt{1 + 1/50}, 257300 + 1.96 \cdot 25000 \cdot \sqrt{1 + 1/50}\right] \\ &= [207812, 306788]. \end{aligned}$$

Note that the 95%-confidence interval for the average loan amount is

$$\begin{aligned} C &= \left[\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right] \\ &= [250370, 264230]. \end{aligned}$$

Prediction interval, normal distribution with unknown variance

Still assume that we sample from a normal distribution. Of course, the variance of the underlying normal distribution is often also unknown.

Without going into details here, we can replace the standard deviation σ by the sample standard deviation $S = \sqrt{S^2}$ if we use quantiles of Student's t-distribution instead of quantiles of the standard normal distribution.

So, the prediction interval, into which a new observation will fall with probability $1 - \alpha$ given the observed sample $\vec{x}$ of size n, is

$$P = \left[\bar{x} - t_{1-\alpha/2} s \sqrt{1 + 1/n}, \bar{x} + t_{1-\alpha/2} s \sqrt{1 + 1/n}\right]$$

where

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Prediction interval, normal distribution with unknown variance

Example

To illustrate the difference between the prediction intervals in the case of known and unknown variance, consider the following situation:

We sample from a normal distribution. We obtain a random sample of size 10 with sample mean 20.

1. Suppose, we know the underlying distribution has a standard deviation of 2. Then, a 95%-prediction interval for a future observation is

$$P = \left[\bar{x} - z_{1-\alpha/2}\sigma\sqrt{1+1/n}, \bar{x} + z_{1-\alpha/2}\sigma\sqrt{1+1/n}\right] = [15.89, 24.11].$$

2. Now, suppose instead that σ is unknown, but that the sample standard deviation is 2. Then, a 95%-prediction interval for a future observation is

$$P = \left[\bar{x} - t_{1-\alpha/2}s\sqrt{1+1/n}, \bar{x} + t_{1-\alpha/2}s\sqrt{1+1/n}\right] = [15.26, 24.74].$$

   Note, however, that here we assumed a situation in which the sample standard deviation is equal to the real standard deviation.

Outline

Testing hypoteses

So far, we investigated how to estimate certain parameters of a distribution.

Often, one is not interested in the exact parameters, but rather wants to check whether statements related to these parameters are correct.

We will begin by illustrating the idea how to test whether a *hypothesis* about a probability distribution is correct.

Testing hypoteses

Let X be a Bernoulli random variable with success probability p. So, $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$.

We want to test whether $p < 1/3$ or $p \geq 1/3$.

To this end, we consider a random sample $X_1, \ldots, X_n$ from the distribution of X. For a resulting sample vector $\vec{x}$, we have to decide whether we accept or reject the hypothesis "$p \geq 1/3$".

The *critical region* of our test is the set

$$K := \{\vec{x} \in \mathbb{R}^n \mid \vec{x} \text{ leads to rejection of the hypothesis}\}.$$

Testing hypoteses

To construct K, we usually construct a new random variable T from $X_1, \ldots, X_n$ called the *test statistic*.

Then, the value range of T is divided into regions that lead to the rejection of the hypothesis or not.

Let $\tilde{K} \subseteq \mathbb{R}$ be the region of the value range of T that leads to rejection of the hypothesis. Then, often also $\tilde{K}$ (instead of K itself) is called critical region.

We have

$$K = T^{-1}(\tilde{K}) \subseteq \mathbb{R}^n.$$

Testing hypoteses

The hypothesis we want to test is usually denoted by H_0. Hence it is called the *null hypothesis*.

In our earlier example, we could choose

$$H_0 : \ p \geq 1/3.$$

If we reject H_0, we accept the *alternative hypothesis* H_1, which in this case is

$$H_1 : \ p < 1/3.$$

Often, H_1 is not explicitly stated and implicitly states that H_0 does not hold. In this case, H_1 is called the *trivial* alternative hypothesis.

A non-trivial alternative hypothesis would be

$$H_1' : \ p \leq 1/6.$$

Example

We investigate a hard drive, for which we know that it is of one of two possible types. The average access time for type 1 is $9\mathrm{ms}$, while for type 2 it is $12\mathrm{ms}$.

We want to determine the type by measuring the access time of n independent accesses. In this case, we could formulate

$$H_0 : \ \mu \leq 9$$

and

$$H_1 : \ \mu \geq 12$$

where μ is the expected access time in ms.

Possible errors in hypothesis testing

It could happen that we reject or do not reject the null hypothesis wrongfully based on the random sample, we obtain.

Definition

Rejection of the null hypothesis when it is true is called a *type I error*.

Non-rejection of the null hypothesis when it is false is called a *type II error*.

	H_0 is true	H_0 is false
do not reject H_0	correct decision	type II error
reject H_0	type I error	correct decision

Possible errors in hypothesis testing

Of course, we want to keep the probability of errors of both types small.

However, keeping the probability of type I and of type II errors small are contrary objectives.

If we choose $K = \emptyset$, i.e., if we never reject H_0, the probability for type I errors is 0. Clearly, this is a useless test.

Possible errors in hypothesis testing

The probability of type I errors (rejecting H_0 although it is true) is typically denoted by α. Sometimes, such an error is hence also called α-error.

The probability α is called the *level of significance*.

In practice, it is common to define a level of significance α. The critical region K is then constructed such that the probability of a type I error is α.

Construction of a test

We construct a test for the parameter p of a Bernoulli random variable.

The hypotheses are

$$H_0 \colon\ p \geq p_0 \qquad \text{and} \qquad H_1 \colon\ p < p_0.$$

The test statistic we use is

$$T := X_1 + \ldots + X_n.$$

In this case, we want to reject the null hypothesis if T returns a small value. The critical region for T will hence be of the form $[0, k]$ for a $k \in \mathbb{R}$ to be determined.

This is a *one-sided* test. If the null hypothesis were $p = p_0$, a *two-sided* test where we reject the null hypothesis for small and large values of T would be appropriate.

Construction of a test

The test statistic $T \sim \mathrm{Bin}(n, p)$ is binomially distributed with parameters n and p.

As we assume the size n of the random sample to be large, we can approximate the distribution of T by a normal distribution.

Let

$$\tilde{T} := \frac{T - np}{\sqrt{np(1-p)}}.$$

Then, $\tilde{T}$ is approximately standard normally distributed.

Construction of a test

For the critical region $K = [0, k]$ for T, we now compute the (worst-case) probability of a type I error, i.e., the level of significance.

We get

$$\begin{aligned}
\Pr(\text{type I error}) &= \max_{p \geq p_0} \Pr_p(T \in K) = \max_{p \geq p_0} \Pr_p(T \leq k) = \Pr_{p=p_0}(T \leq k) \\
&= \Pr_{p=p_0}\left(\tilde{T} \leq \frac{k - np}{\sqrt{np(1-p)}}\right) \\
&= \Pr\left(\tilde{T} \leq \frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right) \\
&\approx \Phi\left(\frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right)
\end{aligned}$$

Construction of a test

Using the quantiles z_q of the standard normal distribution, we obtain the following requirement:

The bound k should be chosen such that

$$\frac{k - np_0}{\sqrt{np_0(1-p_0)}} = z_\alpha.$$

This means that k (depending on n) should be chosen as

$$k = z_\alpha \sqrt{np_0(1-p_0)} + np_0.$$

For smaller values of k, the level of significance would be higher. However, this would increase the probability of type II errors. Hence, we choose k such that the level of significance is (approximately) equal to the specified α.

Construction of a test

What can we say about the probability of type II errors?

Problem:

$$H_0 : \ p \geq p_0 \qquad \text{and} \qquad H_1 : \ p < p_0 .$$

So, if H_1 is the case, the probability p can still be arbitrarily close to p_0.

For the worst-case probability of a type II error, we obtain

$$\Pr(\text{type II error}) = \sup_{p<p_0} \Pr_p(T > k) = \Pr_{p=p_0}(T > k) \approx 1 - \alpha.$$

If the true situation only slightly deviates from the null hypothesis, with very high probability, we will not reject the null hypothesis although it is false.

For test with a non-trivial alternative hypothesis, however, we can choose the critical region such that both error probabilities are small.

Construction of a test

Example

Suppose in a message processing system, two different components may be used. One leads to a probability of message loss of $1/3$; the other to a probability of $1/6$. In this case, we could choose

$$H_0 \colon\ p \geq 1/3 \qquad \text{and} \qquad H_1 \colon\ p \leq 1/6.$$

For the type II error probability, we now obtain

$$\begin{aligned}
\Pr(\text{type II error}) &= \sup_{p \leq 1/6} \Pr_p(T > k) = \Pr_{p=1/6}(T > k) \\
&= \Pr\left(\tilde{T} > \frac{k - n/6}{\sqrt{(1/6) \cdot (5/6) \cdot n}}\right) \\
&\approx 1 - \Phi\left(\frac{k - n/6}{\sqrt{(1/6) \cdot (5/6) \cdot n}}\right).
\end{aligned}$$

The probability of a type II error is often denoted by β.

Construction of a test

Put together, for the hypothesis

$$H_0 \colon\ p \geq 1/3 \qquad \text{and} \qquad H_1 \colon\ p \leq 1/6,$$

we have seen the following:

To achieve a level of significance of α, we have to choose

$$k = z_\alpha \sqrt{n \cdot (1/3) \cdot (1-1/3)} + n/3.$$

For $\alpha = 0.05$, we obtain

$$k \approx 0.33 \cdot n - 0.77 \cdot \sqrt{n}.$$

The resulting probability of a type II error is

$$\beta = \Pr(\text{type II error}) \approx 1 - \Phi\left(\frac{k - n/6}{\sqrt{(1/6) \cdot (5/6) \cdot n}}\right) \approx 1 - \Phi(0.45\sqrt{n} - 2.07).$$

Construction of a test

$$H_0 \colon\ p \geq 1/3 \qquad \text{and} \qquad H_1 \leq 1/6.$$

For $\alpha = 0.05$, we get the following values for different sample sizes n:

n	k	$\beta = \Pr(\text{type II error})$
25	4.4	$1 - \Phi(0.18) \approx 0.43$
100	25.3	$1 - \Phi(2.43) \approx 0.0075$
900	276.9	$1 - \Phi(11.43) \approx 0$

For the trivial alternative $H_1' \colon\ p < 1/3$, the probability of a type II error is always (approximately) $1 - \alpha$.

Application of statistical tests

The test we constructed above is called *approximate binomial test*.

Assumptions: $X_1, \ldots, X_n$ are independent and identically distributed with $\Pr(X_i = 1) = p$ and $\Pr(X_i = 0) = 1 - p$ for all i and an unknown p.

Further, n is sufficiently large; we assume we can approximate the binomial distribution $\mathrm{Bin}(n, p)$ with a normal distribution.

Hypotheses:
a) $H_0 : \ p = p_0$ and $H_1 : \ p \neq p_0$ (two-sided).
b) $H_0 : \ p \geq p_0$ and $H_1 : \ p < p_0$ (one-sided).
c) $H_0 : \ p \leq p_0$ and $H_1 : \ p > p_0$ (one-sided).

Test statistic:

$$Z := \frac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}}.$$

Critical region (rejection of H_0) for level of significance α:
a) $|Z| > z_{1-\alpha/2}$.
b) $Z < z_\alpha$.
c) $Z > z_{1-\alpha}$.

Application of statistical tests

Typical values of α are 0.05 and 0.01.

For the two-sided test, H_0 is rejected if $|Z| > z_{1-\alpha/2}$.

In the one-sided tests, H_0 is rejected if $Z < z_\alpha$ or $Z > z_{1-\alpha}$, respectively.

α	0.05	0.01
$z_{1-\alpha/2}$	1.96	2.58
z_α	-1.64	-2.33
$z_{1-\alpha}$	1.64	2.33

The quantiles in the table specify by how many standard deviations the observed number of successes has to deviate from what is to be expected for $p = p_0$ in order to reject the null hypothesis.

Application of statistical tests

In general, one can follow the following procedure to test hypotheses while ensuring a fixed probability of type I errors:

1. State the null hypothesis and the alternative hypothesis.
2. Choose a fixed level of significance α.
3. Choose an appropriate test statistic and establish the critical region based on α.
4. Compute the test statistic from a random sample. Reject H_0 if the computed test statistic lies in the critical region. Otherwise, do not reject.
5. Draw conclusions.

P-values

Instead of fixing a level of significance α and computing a critical region in which H_0 is rejected, one can compute a so called *p-value*.

Example

Consider an approximate binomial test with the hypotheses

$$H_0 : \ p = 1/2 \qquad \text{and} \qquad H_1 : \ p \neq 1/2.$$

Assume, we consider a random sample of size $n = 100$. Suppose we observe 59 successes in the random sample.

We compute

$$Z = \frac{59 - 100 \cdot 1/2}{\sqrt{100 \cdot (1/2) \cdot (1/2)}} = \frac{9}{5} = 1.8.$$

P-values

Example (continued)

For the fixed level of significance $\alpha = 0.05$, we cannot reject the null hypothesis:

$$1.8 = |Z| < z_{1-\alpha/2} \approx 1.96.$$

However, we see that the test statistic is close to its critical region. Hence, we might wonder: For which level of significance α would we reject the null hypothesis based on the observed random sample?

We are interested in

$$\inf\{\alpha \mid 1.8 < z_{1-\alpha/2}\} \approx 0.072.$$

Definition

The *P-value* of an observed test statistic is the lowest level of significance for which the null hypothesis would be rejected.

Types of statistical tests

There are several criteria according to which statistical tests can be classified.

Number of involved random variables

If there is only one distribution that is analyzed, tests are called *one-sample tests*.

If there are multiple distributions that are compared, tests are called *two-sample tests*.

Example

One sample tests can be used to answer questions like "Is the average access time to the server at most $10\mathrm{ms}$?".

Two sample tests can be used to answer questions like "Is the average access time of server A lower than that of server B?".

Types of statistical tests

In the case of two involved random variables, it is important to know whether the random variables are independent.

To detect (in-)dependence, there are various ways of *correlation analysis*.

In order to provide a quantitative connection between random variables, *regression analysis* can be used.

Types of statistical tests

Formulation of the null hypothesis

Often, the null hypothesis is a statement about a *location parameter* of a distribution.

These parameters could be the *expected value*, the *variance*, or the *median*, i.e., the least value x such that $F(x) = 1/2$ where F is the cumulative distribution function of the distribution.

A two-sample test could, e.g., address the question whether the median of two distributions is the same.

Besides hypotheses about location parameters, one can also formulate hypotheses like "The analyzed distribution has the distribution function F".

Types of statistical tests

Assumptions on the analyzed distributions

Usually, some assumptions about the distribution are made when constructing a test.

E.g., a certain type of distribution is assumed or the expected value or the variance is assumed to be a certain value.

Selection of some statistical tests

We have already seen the **approximate binomial test**:

Assumptions: $X_1, \ldots, X_n$ are independent and identically distributed with $\Pr(X_i = 1) = p$ and $\Pr(X_i = 0) = 1 - p$ for all i and an unknown p.

Further, n is sufficiently large; we assume we can approximate the binomial distribution $\mathrm{Bin}(n, p)$ with a normal distribution.

Hypotheses:
a) $H_0 : \ p = p_0$ and $H_1 : \ p \neq p_0$ (two-sided).
b) $H_0 : \ p \geq p_0$ and $H_1 : \ p < p_0$ (one-sided).
c) $H_0 : \ p \leq p_0$ and $H_1 : \ p > p_0$ (one-sided).

Test statistic:

$$Z := \frac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}}.$$

Critical region (rejection of H_0) for level of significance α:
a) $|Z| > z_{1-\alpha/2}$.
b) $Z < z_{\alpha}$.
c) $Z > z_{1-\alpha}$.

Selection of some statistical tests

As we used the central limit theorem, to assume that the test statistic is approximately normally distributed, the approximate binomial test is a special case of the following Z-**test**.

Assumptions: $X_1, \ldots, X_n$ are independent and identically distributed with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known.
Alternatively, $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ and the sample size n is large.

Hypotheses:
a) $H_0 : \ \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ (two-sided).
b) $H_0 : \ \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$ (one-sided).
c) $H_0 : \ \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ (one-sided).

Test statistic:

$$Z := \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}.$$

Critical region (rejection of H_0) for level of significance α:
a) $|Z| > z_{1-\alpha/2}$.
b) $Z < z_\alpha$.
c) $Z > z_{1-\alpha}$.

Selection of some statistical tests

The downside of the Z-test is that the variance σ^2 has to be known.

As we have seen before, it is an obvious attempt to replace σ^2 with the sample variance S^2 if the variance is not known.

When using S^2 instead of the true σ^2, the test statistic has a t-distribution (with $n-1$ degrees of freedom). For $n \geq 30$, this distribution is very similar to the standard normal distribution.

Selection of some statistical tests

t-test:

Assumptions: $X_1, \ldots, X_n$ are independent and identically distributed with $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
Alternatively, $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$ and the sample size n is large.

Hypotheses:
a) $H_0 : \ \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$ (two-sided).
b) $H_0 : \ \mu \geq \mu_0$ and $H_1 : \mu < \mu_0$ (one-sided).
c) $H_0 : \ \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ (one-sided).

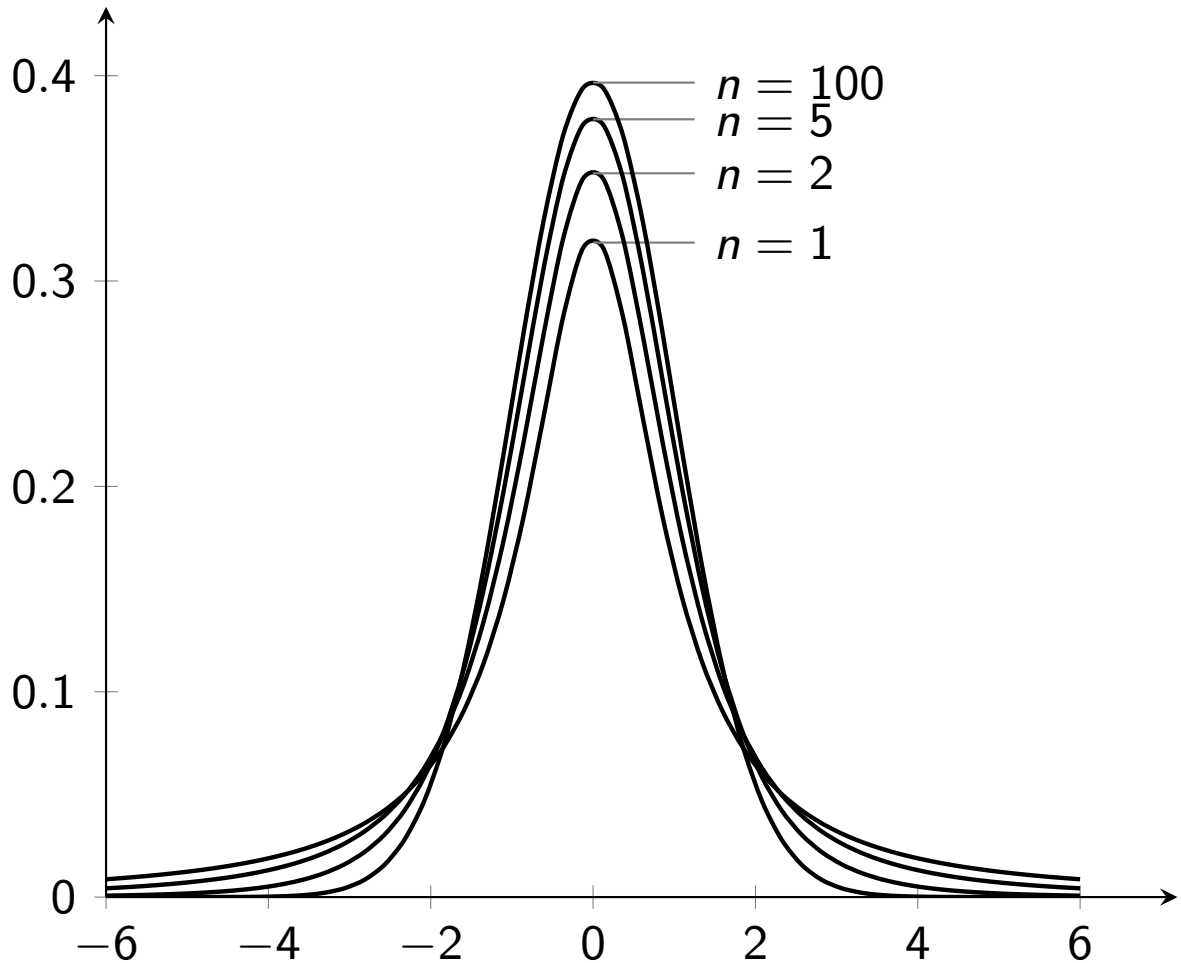
Test statistic:

$$T := \frac{\bar{X} - \mu_0}{S} \sqrt{n}.$$

Critical region (rejection of H_0) for level of significance α:
a) $|T| > t_{1-\alpha/2}$.
b) $T < t_\alpha$.
c) $T > t_{1-\alpha}$.

Selection of some statistical tests

The probability density function of the t-distribution with n degrees of freedom:



For $n \to \infty$, the t-distribution converges to the standard normal disitrbution.

Selection of some statistical tests

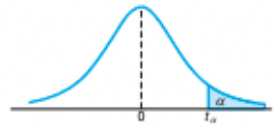


Table A.4 Critical Values of the t-Distribution

	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

Picture: Walpole, Myers, Myers, Ye: Probability and Statistics for Engineers and Scientists (Ninth edition)

Selection of some statistical tests

Now, we take a quick look at a two-sample test.

For two random variables X and Y, we want to test whether the expected values $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$ are equal.

A random sample now consists of independent copies $X_1, \ldots, X_n$ of X and $Y_1, \ldots, Y_m$ of Y.

Selection of some statistical tests

Two-sample-t-test:

Assumptions: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent and identically distributed with $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$.
Furthermore, the variances are equal, i.e., $\sigma_X^2 = \sigma_Y^2$.

Hypotheses:
a) $H_0 \colon \mu_X = \mu_Y$ and $H_1 \colon \mu_X \neq \mu_Y$ (two-sided).
b) $H_0 \colon \mu_X \geq \mu_Y$ and $H_1 \colon \mu_X < \mu_Y$ (one-sided).
c) $H_0 \colon \mu_X \leq \mu_Y$ and $H_1 \colon \mu_X > \mu_Y$ (one-sided).

Test statistic:

$$T := \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{n+m-2}{1/m + 1/n}}.$$

Critical region (rejection of H_0) for level of significance α:
a) $|T| > t_{1-\alpha/2}$.
b) $T < t_\alpha$.
c) $T > t_{1-\alpha}$.
The number of degrees of freedom in the t-distribution is $n + m - 2$ in all cases.

Selection of some statistical tests

There are several variants of the two-sample-t-test in the literature.

These variants can be applied if the two distributions do not have the same variance.

Selection of some statistical tests

Finally, we take a look at a test that is not concerned with a location parameter.

Example

We want to test a die and find out whether it is fair.

So, our hypothesis is that the outcome has a uniform distribution with probability mass function

$$f(x) = 1/6 \quad \text{for } x = 1, \ldots, 6.$$

We toss the die 120 times and observe the following:

Outcome:	1	2	3	4	5	6
Expected:	20	20	20	20	20	20
Observed:	20	22	17	18	19	24

Selection of some statistical tests

For this hypothesis, we can use a **goodness-of-fit test**.

This goodness-of-fit test makes use of the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

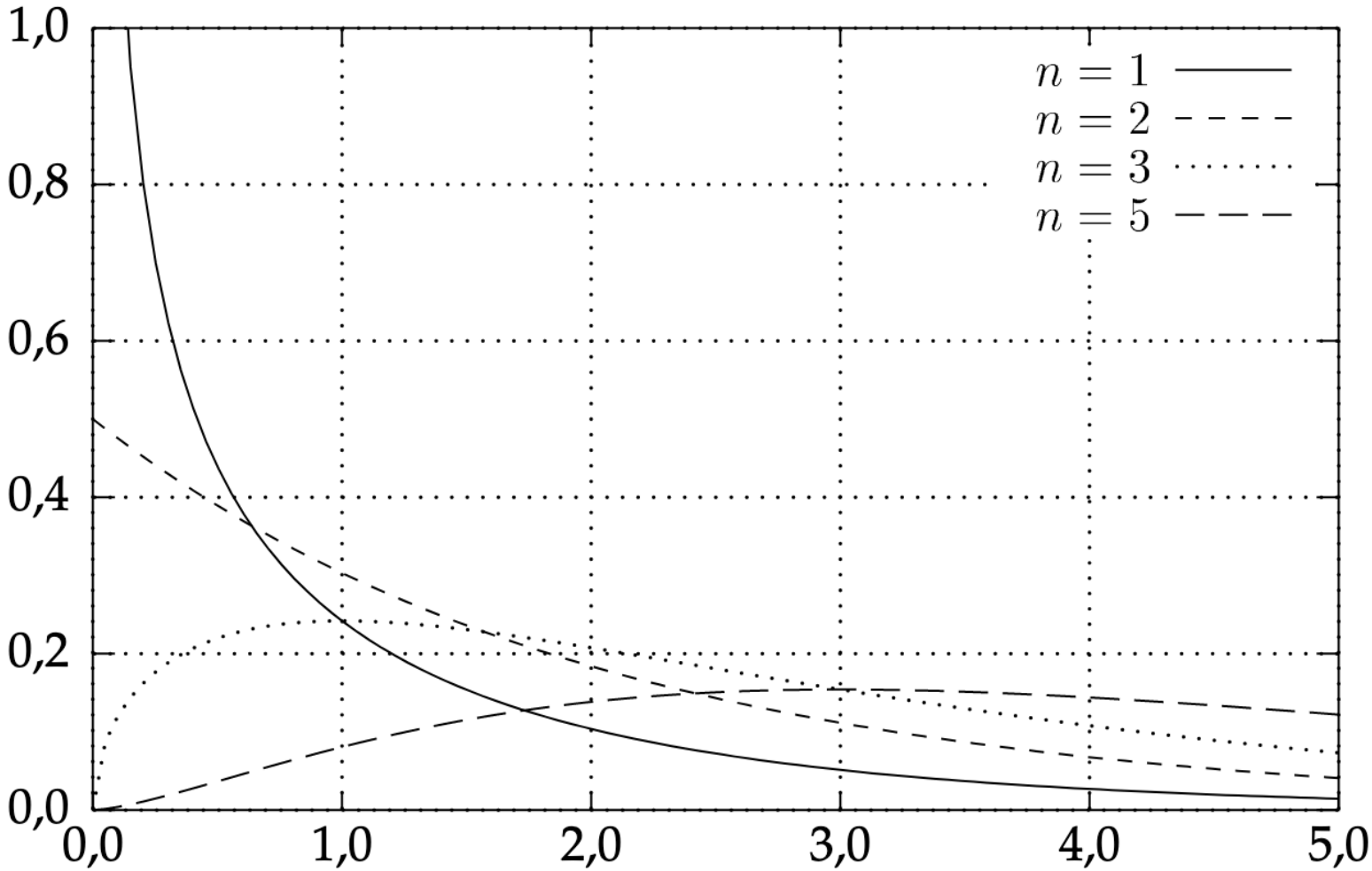
where o_i are the observed frequencies for $i = 1, \ldots, k$ and e_i are the corresponding expected frequencies.

The distribution of the test statistic χ^2 is approximated very closely by the so-called *chi-squared distribution* with $k - 1$ degrees of freedom.

If χ^2 is large, the observed frequencies are a poor fit for the expected frequencies and the null hypothesis should be rejected.

The critical region for a level of significance $1 - \alpha$ can be obtained via the quantiles χ^2_α of the chi-squared distribution. (In the following table χ^2_α is what we would denote by $\chi^2_{1-\alpha}$.)

Selection of some statistical tests



Plot of the probability density function of the chi-squared distribution with n degrees of freedom.

Selection of some statistical tests

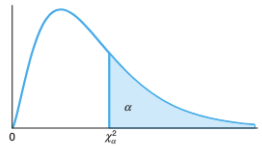


Table A.5 Critical Values of the Chi-Squared Distribution

	α									
v	**0.995**	**0.99**	**0.98**	**0.975**	**0.95**	**0.90**	**0.80**	**0.75**	**0.70**	**0.50**
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	0.102	0.148	0.455
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	0.575	0.713	1.386
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.213	1.424	2.366
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	1.923	2.195	3.357
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	2.675	3.000	4.351
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.455	3.828	5.348
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671	6.346
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527	7.344
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393	8.343
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267	9.342
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148	10.341
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034	11.340
13	3.565	4.107	4.765	5.009	5.892	7.041	8.634	9.299	9.926	12.340
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821	13.339
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721	14.339
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624	15.338
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531	16.338
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	13.675	14.440	17.338
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352	18.338
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266	19.337
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182	20.337
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	17.240	18.101	21.337
23	9.260	10.196	11.293	11.689	13.091	14.848	17.187	18.137	19.021	22.337
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943	23.337
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867	24.337
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792	25.336
27	11.808	12.878	14.125	14.573	16.151	18.114	20.703	21.749	22.719	26.336
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647	27.336
29	13.121	14.256	15.574	16.047	17.708	19.768	22.475	23.567	24.577	28.336
30	13.787	14.953	16.306	16.791	18.493	20.599	23.364	24.478	25.508	29.336
40	20.707	22.164	23.838	24.433	26.509	29.051	32.345	33.66	34.872	39.335
50	27.991	29.707	31.664	32.357	34.764	37.689	41.449	42.942	44.313	49.335
60	35.534	37.485	39.699	40.482	43.188	46.459	50.641	52.294	53.809	59.335

Picture: Walpole, Myers, Myers, Ye: Probability and Statistics for Engineers and Scientists (Ninth edition)

Selection of some statistical tests

Table A.5 (continued) Critical Values of the Chi-Squared Distribution

	α									
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.001**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.466
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.515
6	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	8.383	9.037	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.321
8	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	15.119	15.984	16.985	19.812	22.362	24.736	25.471	27.688	29.819	34.527
14	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.124
15	17.322	18.245	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.698
16	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.791
18	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.819
20	22.775	23.828	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.314
21	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.796
22	24.939	26.039	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.619
26	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.051
27	30.319	31.528	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.475
28	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.994	56.892
29	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.335	58.301
30	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.702
40	44.165	45.616	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.403
50	54.723	56.334	58.164	63.167	67.505	71.420	72.613	76.154	79.490	86.660
60	65.226	66.981	68.972	74.397	79.082	83.298	84.58	88.379	91.952	99.608

Picture: Walpole, Myers, Myers, Ye: Probability and Statistics for Engineers and Scientists (Ninth edition)

Selection of some statistical tests

In our example, we had

Outcome:	1	2	3	4	5	6
Expected:	20	20	20	20	20	20
Observed:	20	22	17	18	19	24

We compute the value

$$\chi^2 = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} \approx 1.7.$$

For a level of significance of 95%, we get from the table that $\chi^2_{0.05}$ with five degrees of freedom is 11.070. So, there is no reason to reject the null hypothesis.

Selection of some statistical tests

χ^2-goodness-of-fit test:

Assumptions: $X_1, \ldots, X_n$ are independent and identically distributed with $W_{X_i} = \{1, \ldots, k\}$.

Hypotheses:
H_0: $\Pr(X_i = i) = p_i$ for $i = 1, \ldots, k$.
H_1: trivial alternative, i.e., H_0 does not hold.

Test statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - n \cdot p_i)^2}{n \cdot p_i}$$

where o_i is the observed frequency of value i in the random sample.

Critical region (rejection of H_0) for level of significance α:
$\chi^2 > \chi^2_\alpha$ for $k - 1$ degrees of freedom.
(Or in our notation for quantiles $\chi^2 > \chi^2_{1-\alpha}$.)

The expected frequencies of all outcomes should be sufficiently large, e.g., all at least 5.

A more detailed rule of thumb asks all expected frequencies to be at least 1 and 80% of the expected frequencies to be at least 5.

Outline

Outline

Stochastic processes

A *stochastic process* describes a probabilistic system that changes over time.

Mathematically, this can be described by a sequence of random variables

$$(X_t)_{t \in T}.$$

The random variable X_t describes the state of the process at time t.

If $T = \mathbb{N}$, we call the process a stochastic process with *discrete time*.
If $T = \mathbb{R}_{\geq 0}$, we call it a stochastic process with *continuous time*.

Example

Let N_t be the number of bacteria in a population t hours after the begin of the experiment.

For each time point $t \in \mathbb{N}$, we get a different random variable.

Of course, there are strong connections between these random variables.

Certainly, N_{100} and N_{101} are not independent.

Often, the conditional distribution of N_{t+1} given the values of N_0, N_1, $\ldots$, N_t, i.e.,

$$\Pr(N_{t+1} = i \mid N_0 = j_0, \ldots, N_t = j_t).$$

can be described.

If we consider the continuous setting $T = \mathbb{R}_{\geq 0}$, expressing the connection between the random variables typically becomes more involved.

Example (Random walk)

An example of a random walk is given by the following sequence of random variables $(X_t)_{t\in\mathbb{N}}$: At time 0, $X_0 = 0$ and for all $t \in \mathbb{N}$

$$\Pr(X_{t+1} = k+1 \mid X_t = k) = \frac{1}{2},$$
$$\Pr(X_{t+1} = k-1 \mid X_t = k) = \frac{1}{2}.$$

Alternatively, this can be formulated as follows: Let $Z_0, Z_1, \ldots$ be a sequence of independent random variables such that

$$\Pr(Z_i = 1) = \Pr(Z_i = -1) = \frac{1}{2}.$$

Then,

$$X_t = \sum_{i=0}^{t-1} Z_i.$$

Example (continued)

For this example, we observe that

$$\mathbb{E}(X_t) = \sum_{i=0}^{t-1} \mathbb{E}(Z_i) = 0.$$

Furthermore,

$$\begin{aligned}
\mathbb{E}(X_t^2) &= \mathbb{E}\left(\left(\sum_{i=0}^{t-1} Z_i\right)^2\right) \\
&= \mathbb{E}\left(\sum_{i=0}^{t-1} Z_i^2 + 2 \sum_{0 \le i < j \le t-1} Z_i Z_j\right) \\
&= \sum_{i=0}^{t-1} \mathbb{E}(Z_i^2) + 2 \sum_{0 \le i < j \le t-1} \mathbb{E}(Z_i Z_j) = n.
\end{aligned}$$

Markov chains

Finite-state *Markov chains* are a special case of discrete-time stochastic processes.

Definition

A Markov chain over the state space $S = \{1, \dots, n\}$ consists of a sequence of random variables $(X_t)_{t \in \mathbb{N}}$ with value range S satisfying the following condition:

There are values $p_{ij} \geq 0$ for $1 \leq i, j \leq n$ such that for any sequence of values $s_0, \dots, s_{t-1} \in S$,

$$\Pr(X_{t+1} = j \mid X_0 = s_0, \dots, X_{t-1} = s_{t-1}, X_t = i) = \Pr(X_{t+1} = j \mid X_t = i) = p_{ij}.$$

Markov chains are named after Andrey Andreyevich Markov (1856–1922).

Markov chains

The condition for Markov chains states that the probability to move to state j at time $t+1$ depends only on the state at time t.

The "memory" of a Markov chain consists only of the current state. It does not "remember" how it got there.

The property that the probability to move from a state i to a state j does also not depend on the current time t is sometimes called time-homogeneity, and time-inhomogeneous variants of Markov chains are also studied.

Definition

Let $(X_t)_{t\in\mathbb{N}}$ be a Markov chain with state space $S = \{1, \dots, n\}$. The *initial distribution* q of the Markov chain is the (row) vector

$$(q_1, \dots, q_n) = (\Pr(X_0 = 1), \dots, \Pr(X_0 = n)) \in \mathbb{R}^n.$$

The entries $q_1, \dots, q_n$ are all non-negative and sum up to 1.

The *transition probability matrix* $P \in \mathbb{R}^{n\times n}$ of the Markov chain is given by

$$p_{ij} = \Pr(X_{t+1} = j \mid X_t = i).$$

Recall that p_{ij} is well-defined.

Note that the matrix P has the property that all entries are non-negative and that all rows sum up to 1. Such a matrix is called a *stochastic matrix*.

Markov chains

We also view P as a function from $S \times S$ to $[0, 1]$ by setting $P(i, j) = p_{ij}$.

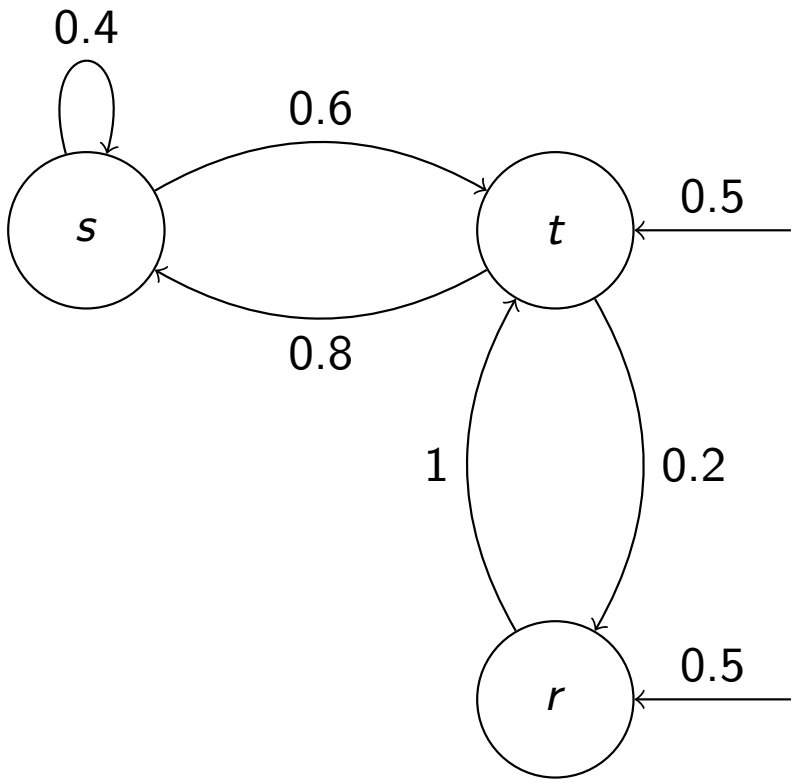
We can view Markov chains as triples $M = (S, P, q)$ of state space, transition probability matrix/function, and initial distribution.

Often, we use arbitrary finite (ordered) sets of states S instead of $\{1, \ldots, n\}$ and write $P(s, t)$ for the transition probability from state s to state t and $q(s)$ for $\Pr(X_0 = s)$.

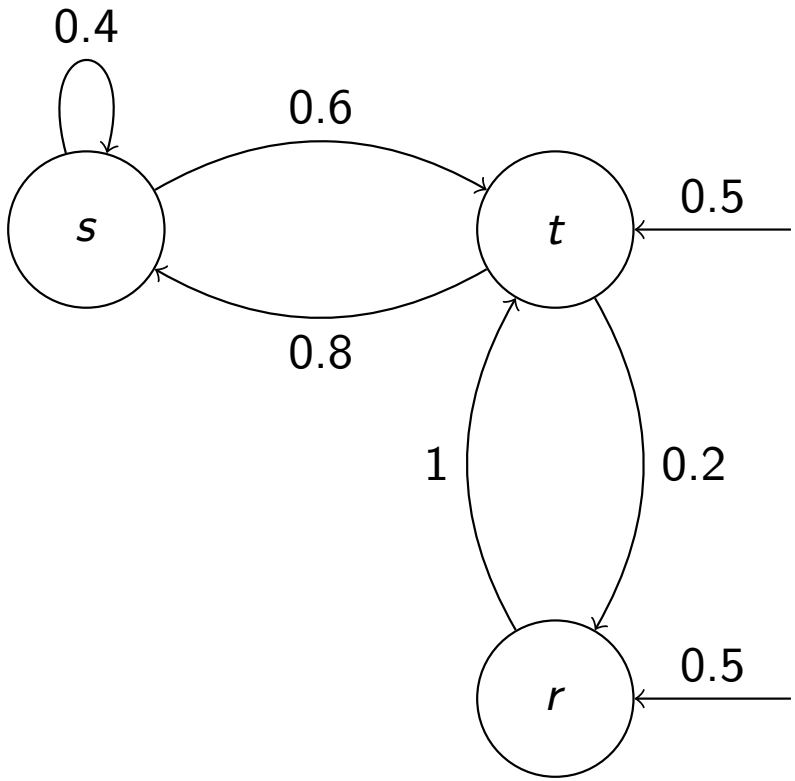
Markov chains

Given a Markov chain $M = (S, P, q)$, we can represent the Markov chain graphically.

The representation is an annotated directed graph with nodes S and edges between two nodes s and t if $P(s, t) > 0$. The edges are annotated with probabilities. The initial distribution is indicated by arrows not coming from a state.



Markov chains



This Markov chain is given by $S = \{s, t, r\}$,

$$P = \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.8 & 0 & 0.2 \\ 0 & 1 & 0 \end{pmatrix}, \qquad q = (0, 0.5, 0.5).$$

Markov chains

Definition

Given a Markov chain $M = (S, P, q)$, we call the graph $(S, \rightarrow)$, where for all states s and t we have $s \rightarrow t$ (i.e., $(s, t) \in \rightarrow$) iff $P(s, t) > 0$, the *graph of the Markov chain M*.

We say that a state t is reachable from s in M if t is reachable from s in the graph of M.

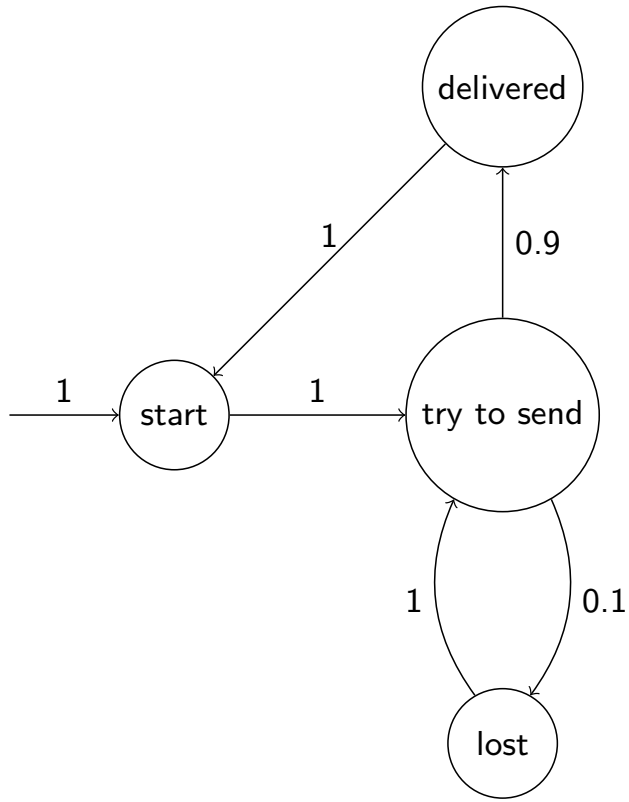
Markov chains can be seen as operational system models like transition systems.

Instead of non-deterministic transitions between states as in the case of transition systems, transitions are chosen randomly.

In this view, Markov chains can be used to model, e.g., randomized algorithms or systems interacting with a probabilistic environment (e.g., certain components of system might fail with a certain probability).

Example

Model of a simple communication protocol:



We might want to answer questions like "What is the probability that a message is eventually delivered?" or "How many time steps does it take in average to deliver a message after a message loss?".

Definition

Let $M = (S, P, q)$ be a Markov chain. If we observe the Markov chain for N time steps, possible outcomes are paths in $\Omega = S^{N+1}$.

The probability of a path $w = s_0 s_1 \dots s_N \in \Omega$ is given by

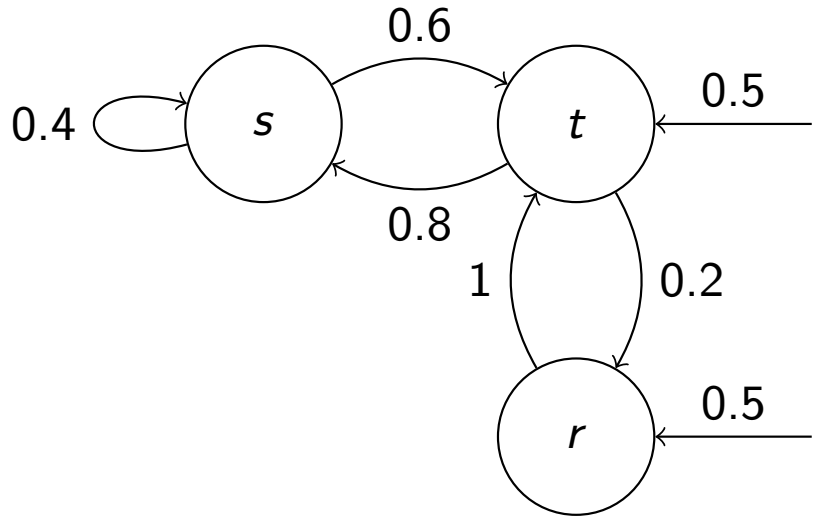
$$\Pr(w) = q(s_0) \cdot \prod_{i=0}^{N-1} P(s_i, s_{i+1}).$$

This defines a discrete probability space.

For an event $A \subseteq \Omega$, the probability is given by

$$\Pr(A) = \sum_{w \in A} \Pr(w).$$

Markov chains



Let the time horizon be $N = 3$. We want to know the probability to reach s within these three steps.

Let $A = tsS^2 \cup rtsS \cup \{trts\}$. This event collects all paths with non-zero probability that reach s within three time steps.

$$\Pr(A) = \sum_{w \in A} \Pr(w) = 0.5 \cdot 0.8 + 0.5 \cdot 1 \cdot 0.8 + 0.5 \cdot 0.2 \cdot 1 \cdot 0.8 = 0.88.$$

Markov chains

If we observe a Markov chain $M = (S, P, q)$ infinitely long, the possible outcomes are

$$\Omega = S^\omega.$$

This is an uncountable sample space. We can define a probability space by specifying the σ-algebra generated by all cylinder sets.

For a finite path $w = (s_0, \ldots, s_k)$, the cylinder set $Cyl(w)$ is defined as

$$Cyl(w) = \{\rho \in \Omega \mid w \text{ is a prefix of } \rho\}.$$

We let $\mathfrak{A}$ be the smallest σ-algebra that contains all cylinder sets.

For cylinder sets, we define

$$\Pr(Cyl(w)) = q(s_0) \cdot \prod_{i=0}^{k-1} P(s_i, s_{i+1}).$$

This can be extended to a probability measure on the σ-algebra $\mathfrak{A}$. In this way, we obtain a general probability space $(\Omega, \mathfrak{A}, \Pr)$.

If the Markov chain is not clear from context, we also write $\Pr_M$ for the probability measure.

Example

Usually, all events that we are interested in are members of the σ-algebra $\mathfrak{A}$.

Let $M = (S, P, q)$ be a Markov chain. The event A that a certain state s is reached is a union of cylinder sets:

$$A = \bigcup_{w \in (S \setminus \{s\})^* \{s\}} Cyl(w).$$

So, the probability $\Pr(A)$ is well-defined. As the cylinder sets in the representation above are disjoint, we get

$$\Pr(A) = \sum_{w \in (S \setminus \{s\})^* \{s\}} \Pr(Cyl(w)).$$

Computation of transition probabilities

Given a Markov chain $M = (S, P, q)$ with state space $S = \{1, \dots, n\}$, let X_t be the random variable that describes the state after t steps for $t \in \mathbb{N}$. So, the value range of X_t is S.

As we did for the initial distribution, we can describe the situation after t steps by a distribution vector q_t. This distribution after t steps is the row vector

$$q_t = ((q_t)_1, \dots, (q_t)_n) = (\Pr(X_t = 1), \dots, \Pr(X_t = n)) \in \mathbb{R}^n.$$

In particular, $q_0 = q$.

Computation of transition probabilities

Observe that

$$\Pr(X_{t+1} = j) = \sum_{i=1}^{n} \Pr(X_{t+1} = j \mid X_t = i) \cdot \Pr(X_t = i).$$

So,

$$(q_{t+1})_j = \sum_{i=1}^{n} p_{ij} \cdot (q_t)_i.$$

In matrix formulation, we obtain

$$q_{t+1} = q_t \cdot P.$$

Inductively, we see that

$$q_t = q_0 \cdot P^t = q \cdot P^t.$$

Computation of transition probabilities

The probability to transition from state i to state j in exactly k steps is given by $(P^k)_{ij}$.

More precisely,

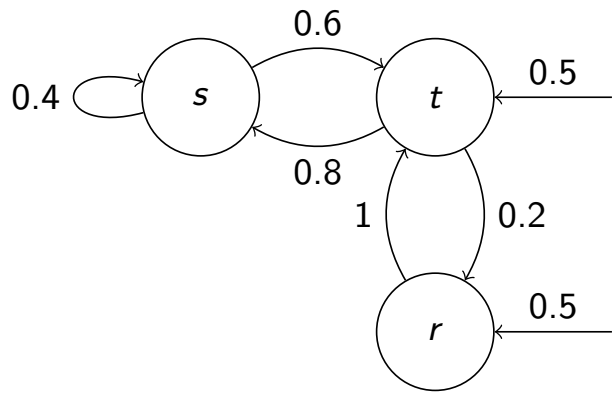
$$p_{ij}^{(k)} := \Pr(X_{t+k} = j \mid X_t = i) = (P^k)_{ij}.$$

Computation of transition probabilities

Example

Consider the Markov chain $M = (S, P, q)$ with $S = \{s, t, r\}$,

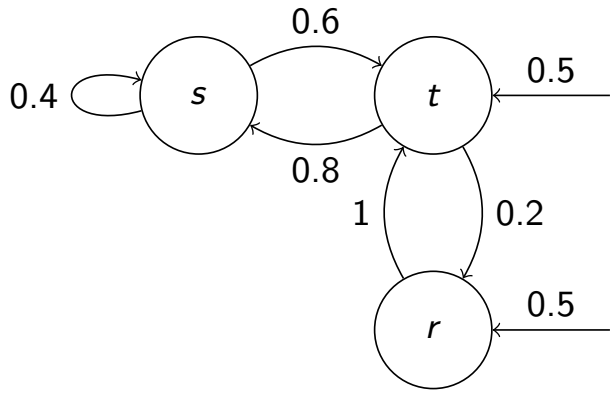
$P = \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.8 & 0 & 0.2 \\ 0 & 1 & 0 \end{pmatrix}$, and $q = (0, 0.5, 0.5)$.



The probability to be in state s four time steps after being in state r is given by

$$(P^4)_{3,1} = \begin{pmatrix} 0.5824 & 0.3168 & 0.1008 \\ 0.4224 & 0.5392 & 0.0384 \\ 0.672 & 0.192 & 0.136 \end{pmatrix}_{3,1} = 0.672$$

Computation of transition probabilities



Alternatively, we can compute the probabilities of paths from r to s of length $4+1$. We change the initial distribution to $\Pr(X_0 = r) = 1$ and compute

$$\begin{aligned}\Pr(\{rtsss, rtsts, rtrts\}) &= 1 \cdot 0.8 \cdot 0.4^2 + 1 \cdot 0.8 \cdot 0.6 \cdot 0.8 + 1 \cdot 0.2 \cdot 1 \cdot 0.8 \\ &= 0.128 + 0.384 + 0.16 = 0.672.\end{aligned}$$

Clearly, for long paths and large Markov chains, this becomes infeasible quite quickly while the exponentiation of the transition probability matrix can easily be done for large Markov chains and large numbers of steps.

Exponentiation of matrices

For diagonalizable matrices P, we can find a diagonal matrix D and an invertible matrix B such that

$$P = B \cdot D \cdot B^{-1}.$$

The entries on the diagonal of D are the eigenvalues of P while the columns of B are the eigenvectors.

This makes exponentiation of P very easy:

$$P^k = (B \cdot D \cdot B^{-1})^k = B \cdot D^k \cdot B^{-1}.$$

Further

$$D^k = \begin{pmatrix} d_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_{nn} \end{pmatrix}^k = \begin{pmatrix} d_{11}^k & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_{nn}^k \end{pmatrix}.$$

Exponentiation of matrices

Example

Consider the matrix

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix}.$$

We determine the roots of the characteristic polynomial $\det(P - \lambda \cdot I)$ where I is the unit matrix.

We obtain 0.7 and 1 as well as the corresponding eigenvectors

$$v_1 = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Setting

$$D = \begin{pmatrix} 0.7 & 0 \\ 0 & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix}, \text{ and hence } B^{-1} = \begin{pmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix},$$

we obtain, for example,

$$P^3 = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0.7 & 0 \\ 0 & 1 \end{pmatrix}^3 \begin{pmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix} \approx \begin{pmatrix} 0.56 & 0.44 \\ 0.22 & 0.78 \end{pmatrix}.$$

Exponentiation of matrices

Besides that, we just want to remark that repeated squaring can be used to compute exponents of a matrix.

Example

To compute P^{1024}, we can successively compute

$$P^2 = P^1 \cdot P^1$$
$$P^4 = P^2 \cdot P^2$$
$$P^8 = P^4 \cdot P^4$$
$$\dots$$
$$P^{1024} = P^{512} \cdot P^{512}.$$

Further, e.g.,

$$P^{1600} = P^{1024} \cdot P^{512} \cdot P^{64}.$$

Skolem Problem

We have seen that the probability to be in a certain state s after k steps is relatively easy to compute:

$$\Pr(X_k = s) = (q_k)_s = (q \cdot P^k)_s.$$

The following problem, however, has been open for many decades:

Definition (Skolem Problem)

Given a Markov chain $M = (S, P, q)$ and a state s, the *Skolem Problem* asks whether there is a natural number k such that

$$\Pr(X_k = s) = \frac{1}{2}.$$

Another formulation of the problem asks, given a transition probability matrix P of a Markov chain, whether there is a k such that $(P^k)_{1,1} = \frac{1}{2}$.

The problem is furthermore inter-reducible to the question whether, for an arbitrary matrix $A \in \mathbb{Z}^{n \times n}$, there is number k such that $(A^k)_{1,1} = 0$.

Skolem Problem

Remarkably, we can decide for a Markov chain whether there are *infinitely many* $k \in \mathbb{N}$ such that

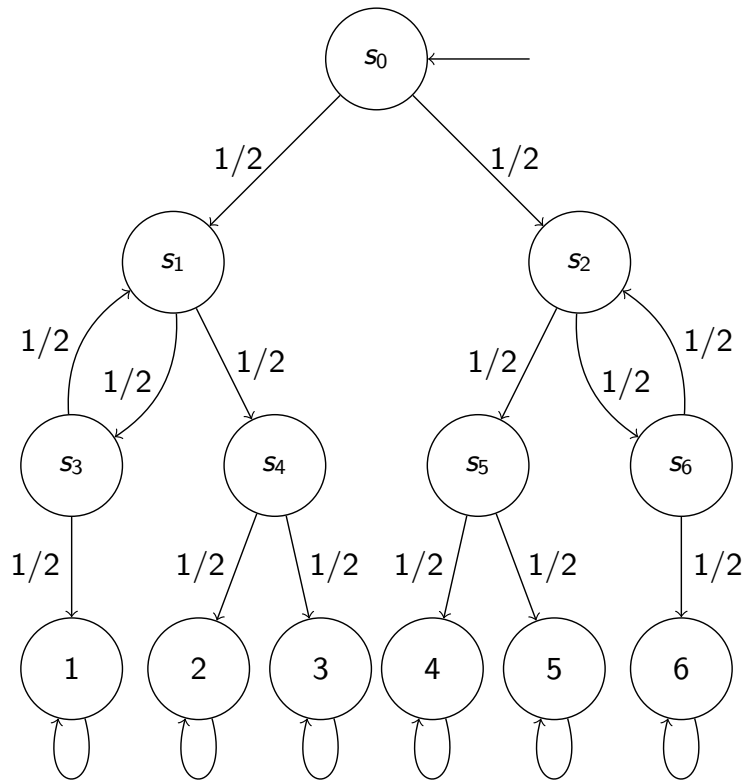
$$\Pr(X_k = s) = \frac{1}{2}.$$

One can even compute a representation of a set B that contains all time steps k with $\Pr(X_k = s) = \frac{1}{2}$ except for possibly finitely many exceptions.

Reachability probabilities

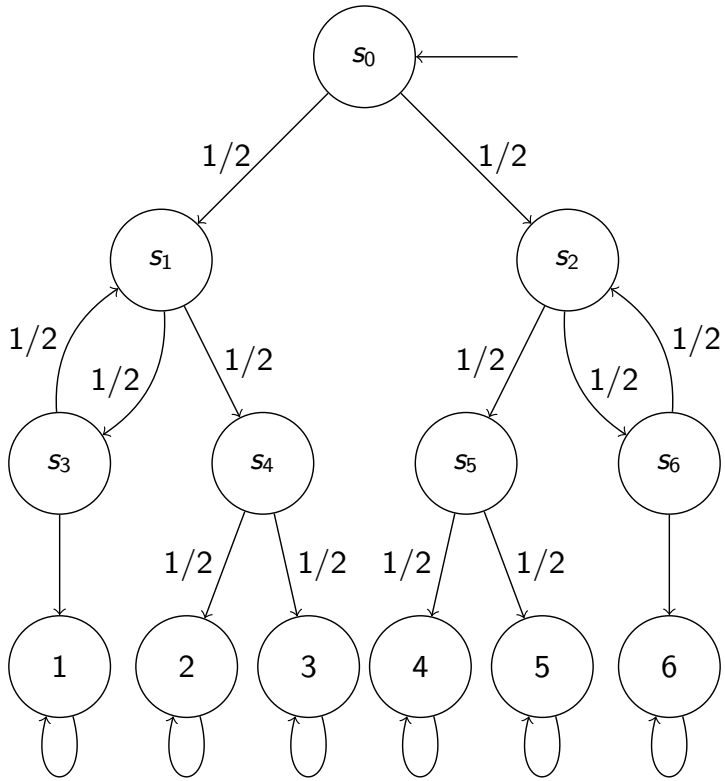
A typical question about a Markov chain one wants to answer is "What is the probability that *eventually* a state s is reached?".

Example (Knuth's dice)



What is the probability to obtain a 6 by this simulation of a die with a fair coin?

Reachability probabilities



Let $x_0, \ldots, x_6$ be the probabilities to reach 6 from states $s_0, \ldots, s_6$, respectively. We immediately see that $x_1 = x_3 = x_4 = x_5 = 0$. Furthermore,

$$x_0 = 1/2 \cdot x_1 + 1/2 \cdot x_2 = 1/2 \cdot x_2$$
$$x_2 = 1/2 \cdot x_5 + 1/2 \cdot x_6 = 1/2 \cdot x_6$$
$$x_6 = 1/2 \cdot 1 + 1/2 \cdot x_2$$

Reachability probabilities

We solve the system

$$\begin{aligned} x_0 &= 1/2 \cdot x_2 \\ x_2 &= 1/2 \cdot x_6 \\ x_6 &= 1/2 \cdot 1 + 1/2 \cdot x_2. \end{aligned}$$

We obtain

$$\begin{aligned} x_0 &= 1/6 \\ x_2 &= 1/3 \\ x_6 &= 2/3. \end{aligned}$$

So, the probability to obtain a 6 in the simulation of a die with a fair coin is indeed $\frac{1}{6}$.

Analogously, we can check that all outcomes have probability $1/6$.

Reachability probabilities

Let $M = (S, P, q)$ be a Markov chain. We (implicitly) followed the following procedure to compute the reachability probability in the example:

1. For each state s introduce a variable x_s, representing the probability to reach the target state g from s.
2. Analyze the graph of the Markov chain to determine the states r with $x_r = 0$.
3. Analyze the graph of the Markov chain to determine the states t with $x_t = 1$. In particular, $x_g = 1$.
4. For all remaining states s, formulate the equation

$$x_s = \sum_{t \in S} P(s, t) \cdot x_t.$$

5. Compute the vector x by solving the equation system.
6. The desired reachability probability is $q \cdot x$.

Reachability probabilities

In more detail, the pre-processing works as follows:

1. Change the transition probability function by setting $P(g, g) = 1$ and $P(g, s) = 0$ for all $s \neq g$. In this way, the state g is made *absorbing*.
2. Compute the set

$$S_0 = \{s \in S \mid g \text{ is not reachable from } s\}.$$

3. Compute the set

$$S_1 = \{s \in S \mid S_0 \text{ is not reachable from } s\}.$$

4. Let $S_? = S \setminus (S_0 \cup S_1)$ and solve the equation system with the following equation for each state $s \in S_?$:

$$x_s = \sum_{t \in S_?} P(s, t) \cdot x_t + \sum_{r \in S_1} P(s, r).$$

One can show that S_1 contains exactly the states from which g is reached with probability 1 and that S_0 contains exactly the states from which g is reached with probability 0.

Reachability probabilities

The equation system takes the form

$$I \cdot \vec{x} = A \cdot \vec{x} + b$$

where $\vec{x}$ consists of the variables x_s with $s \in S_?$ and A is a matrix, b a vector, and I the identity matrix of appropriate size.

The equation system is equivalent to

$$(I - A)x = b.$$

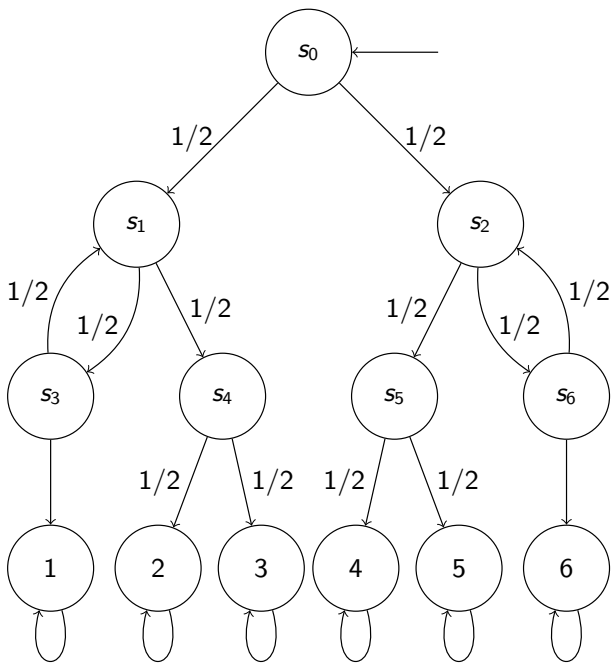
One can show that after the pre-processing described above the matrix $(I - A)$ is invertible and the system hence has the unique solution $x = (I - A)^{-1} b$.

Adding the variables set to value 0 for states in S_0 and to 1 for states in S_1, we can hence compute a (column) vector $\vec{v}$ containing the probabilities to reach g from each state $s \in S$. The probability to reach g in the investigated Markov chain is then $q \cdot v$ where q is the (row) vector containing the initial distribution.

Reachability probabilities

Example

Let us more formally apply the procedure to the Markov chain M we have seen before. This time, we compute the probability to reach 5.



First, we compute
$S_0 = \{s \in S \mid 5 \text{ not reachable from } s\}$.

$S_0 = \{s_1, s_3, s_4, 1, 2, 3, 4, 6\}$.

Next, we compute
$S_1 = \{s \in S \mid S_0 \text{ not reachable from } s\}$.

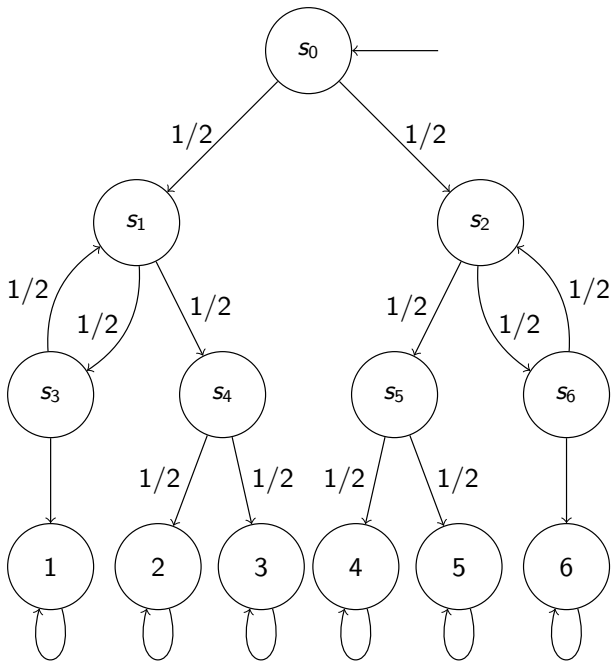
$S_1 = \{5\}$.

For $s \notin S_? = S \setminus (S_0 \cup S_1)$, we the following equation to an equation system:

$$x_s = \sum_{t \in S_?} P(s,t) \cdot x_t + \sum_{r \in S_1} P(s,r).$$

Example (continued)

We obtain the following equation system:



$$
\begin{aligned}
x_{s_0} &= 1/2 \cdot x_{s_2} \\
x_{s_2} &= 1/2 \cdot x_{s_5} + 1/2 \cdot x_{s_6} \\
x_{s_5} &= 1/2 \\
x_{s_6} &= 1/2 \cdot x_{s_2}
\end{aligned}
$$

We obtain the solution:

$$
\begin{aligned}
x_{s_0} &= 1/6 \\
x_{s_2} &= 1/3 \\
x_{s_5} &= 1/2 \\
x_{s_6} &= 1/6
\end{aligned}
$$

As the initial distribution assigns probability 1 to s_0, the probability to reach 5 is $x_{s_0} = 1/6$.

Example (continued)

Note that the initial distribution q is given by the vector

$$q = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \in \mathbb{R}^{13}.$$

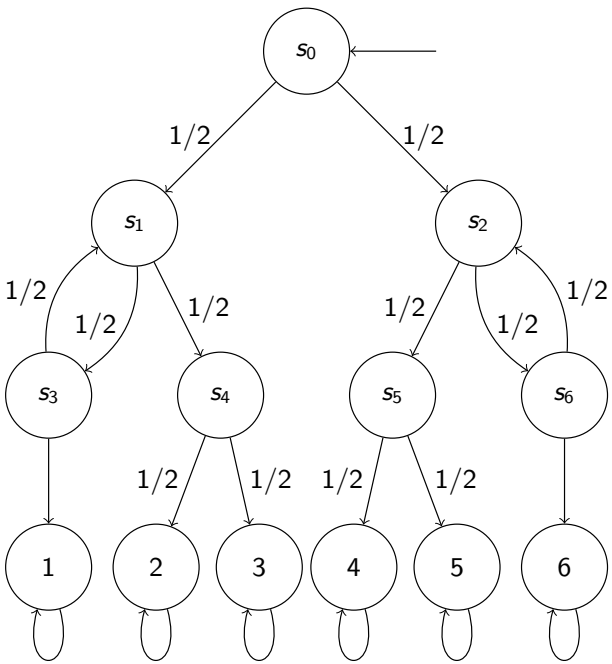
The full solution vector with entries for states in S_0 and S_1 is

$$x = \begin{pmatrix} 1/6 \\ 0 \\ 1/3 \\ 0 \\ 0 \\ 1/2 \\ 1/6 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

So, $q \cdot x = 1/6$.

Reachability probabilities

Example (continued)



If we do not compute S_0 and S_1 first, we run into problems.

Our equation system would contain equations like

$$x_1 = x_1 .$$

So, we could choose any value for x_1.

This value would be propagated and hence also x_{s_0} could take arbitrary values.

Bottom strongly connected components

A directed graph is called *strongly connected* if all nodes are reachable from any node.

Definition

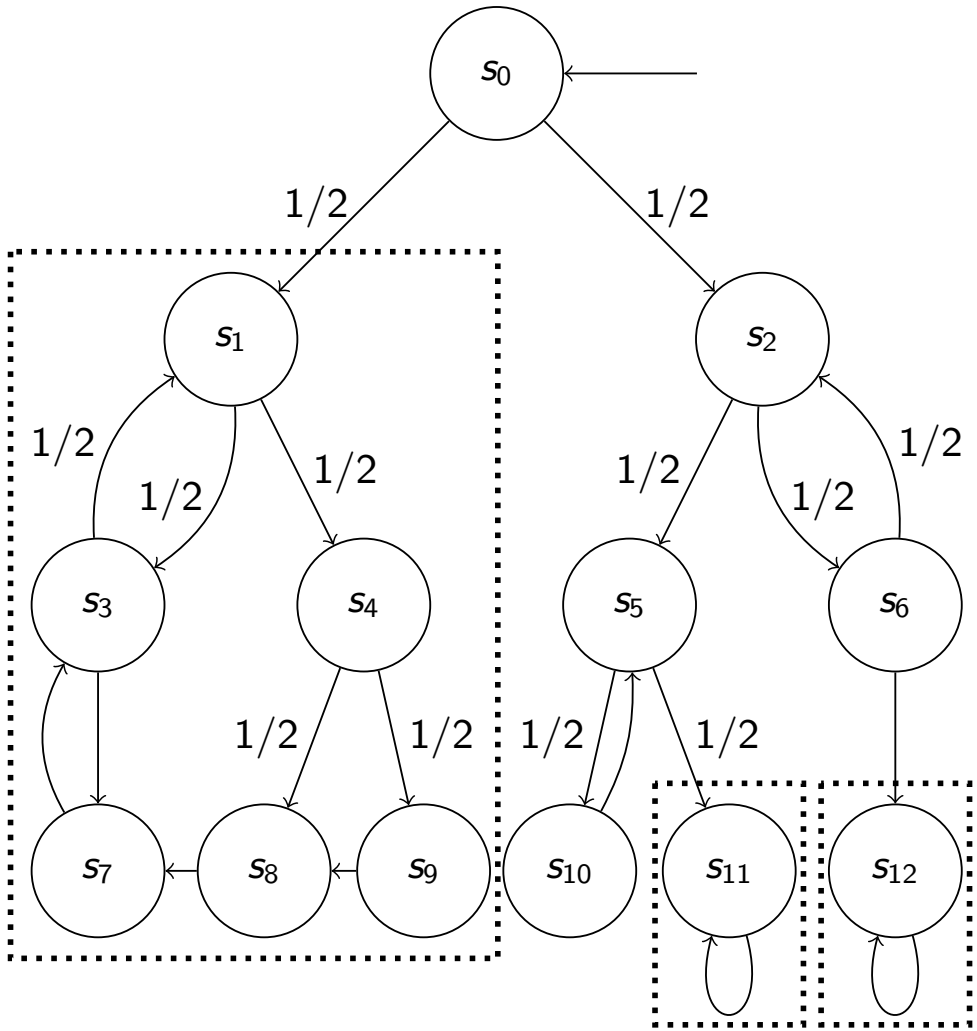
Let $M = (S, P, q)$ be a Markov chain and let $\to \subseteq S \times S$ be the induced edge relation on S. (I.e., for states $s, t \in S$, we have $s \to t$ iff $P(s, t) > 0$.)

A non-empty set $B \subseteq S$ together with the restriction of the transition probability function $P \restriction B \times B$ is called a *bottom strongly connected component* (BSCC) of M if

- B together with $\to \restriction B \times B$ is strongly connected, and
- there are no states $s \in B$ and $t \in S \setminus B$ with $s \to t$.

In other words, a BSCC is a sub-Markov chain in which all states are reachable from all states and which cannot be left.

Bottom strongly connected components



BSCCs in this Markov chain

Realization of BSCCs of (finite) Markov chains

Given an infinite path w, we denote by $\mathit{Inf}(w)$ the set of states that are visited infinitely often along w.

Theorem (Realization of BSCCs of Markov chains)

Let $M = (S, P, q)$ be a Markov chain. With probability 1, the states that are visited infinitely often on a run form a BSCC.

More formally, let

$$A = \{w \in S^\omega \mid \mathit{Inf}(w) \text{ is a BSCC of } M\}.$$

Then, the event A has probability 1.

Realization of BSCCs of (finite) Markov chains

Our goal is to prove this theorem. We start with the following result:

Theorem (Probabilistic choice as strong fairness)

Let M be a Markov chain and s and t states of M such that t is reachable from s. Then,

$$\Pr(s \textit{ is visited infinitely often}) = \Pr(s \textit{ and } t \textit{ are visited infinitely often}).$$

Proof.

We fix a simple (i.e., no state is repeated) finite path fragment $\pi = s_0 \dots s_k$ with $s_0 = s$ and $s_k = t$.

First, we show that

$$\Pr(s \text{ is visited infinitely often}) \leq \Pr(\text{eventually the path } \pi \text{ is taken}).$$

Let $p = \prod_{i=0}^{k-1} P(s_i, s_{i+1})$. Then, $0 < p \leq 1$.

Realization of BSCCs of (finite) Markov chains

Proof continued.

Now, let E_n be the event that "s is visited at least n times, but the path fragment π is never taken".

Observe that $E_1 \supseteq E_2 \supseteq \dots$.

Furthermore, $\Pr(E_n) \leq (1-p)^n$.

Let $E = \bigcap_{n>1} E_n$. So, E is the event that s is visited infinitely often but π is never taken.

As $E_1 \supseteq E_2 \supseteq \cdots \supseteq E$, we get that

$$0 \leq \Pr(E) \leq \lim_{n\to\infty} \Pr(E_n) = \lim_{n\to\infty} (1-p)^n = 0.$$

So,

$$\Pr(s \text{ is visited infinitely often but } \pi \text{ is never taken}) = 0$$

and hence

$$\Pr(s \text{ is visited infinitely often}) \leq \Pr(\text{eventually the path } \pi \text{ is taken}).$$

Proof continued.

Now, let F_n be the event that s is visited infinitely often, but after more than n steps π is never taken.

$$\Pr(F_n) = \sum_{r \in S} \Pr(X_n = r) \cdot \Pr_{M_r}(E)$$

where M_r is the Markov chain M in which the initial distribution has been changed to choose r with probability 1.

Our proof that $\Pr(E) = 0$ works for any initial distribution. So, $\Pr_{M_r}(E) = 0$ for all $r \in S$ and hence $\Pr(F_n) = 0$.

Let $F = \bigcup_{n \geq 1} F_n$. So, F states that s is visited infinitely often and from some moment on π is not taken anymore.

As $F_1 \subseteq F_2 \subseteq \ldots$, we conclude

$$\Pr(F) = \lim_{n \to \infty} \Pr(F_n) = 0.$$

Realization of BSCCs of (finite) Markov chains

Proof continued.

We conclude

$$\begin{aligned}&\Pr(s \text{ is visited infinitely often and } \pi \text{ is taken infinitely often})\\ =\,&\Pr(s \text{ is visited infinitely often}) - \Pr(F)\\ =\,&\Pr(s \text{ is visited infinitely often}).\end{aligned}$$

So, indeed,

$$\Pr(s \text{ is visited infinitely often}) \leq \Pr(s \text{ and } t \text{ are visited infinitely often}).$$

The other inequality holds obviously and so the two probabilities are equal. □

In fact, we have shown a stronger statement. For any path fragment π starting in s, the probability that π is taken infinitely often is equal to the probability that s is visited infinitely often almost surely!

This can be seen as a strong form of *fairness*: Every finite sequence of transitions that could be taken infinitely often is taken infinitely often.

Realization of BSCCs of (finite) Markov chains

Now, we turn to the proof of the Realization of BSCCs of (finite) Markov chains.

Theorem (Realization of BSCCs of Markov chains)

Let $M = (S, P, q)$ be a Markov chain. With probability 1, the states that are visited infinitely often by a run form a BSCC.

Proof.

By the previous theorem, we conclude the following: If
$\Pr(s \text{ is visited infinitely often}) > 0$, then

$$\Pr(t \text{ is visited infinitely often} \mid s \text{ is visited infinitely often}) = 1$$

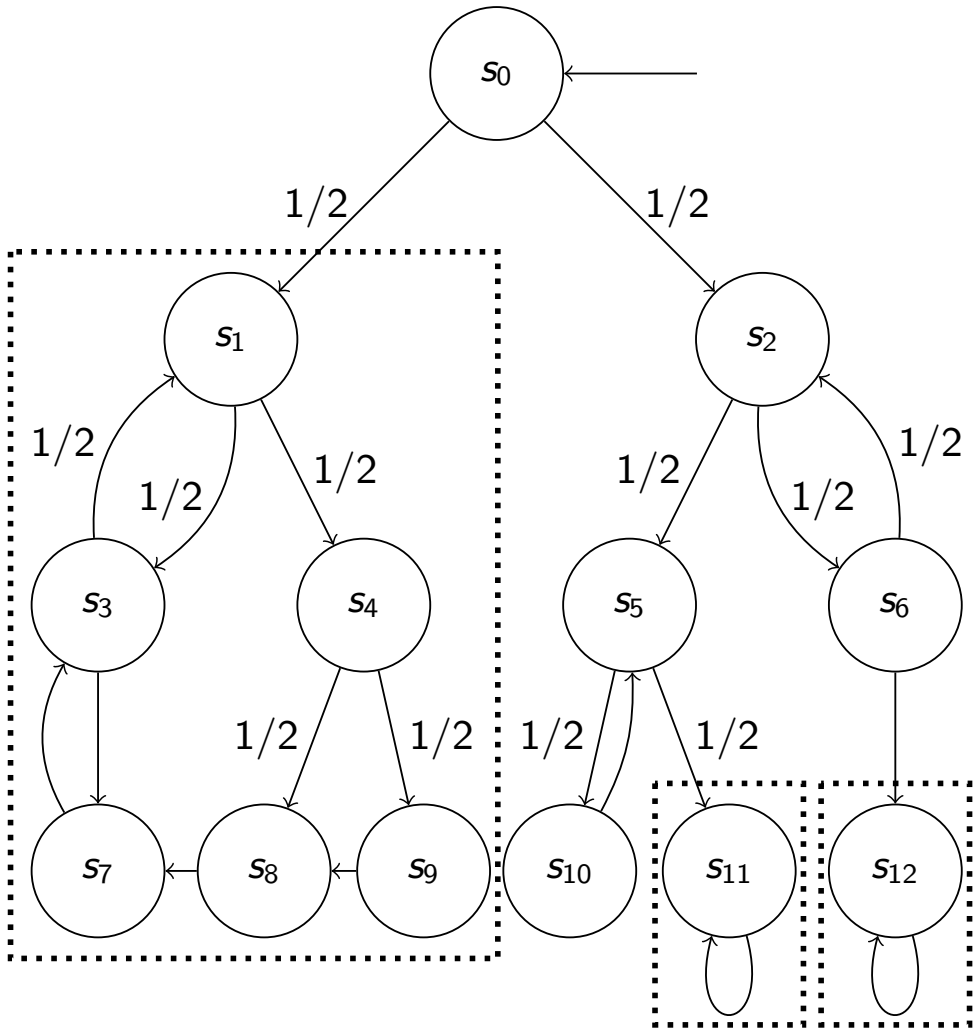
for all states t that are reachable from s.

So, with probability 1, the set of states that are visited infinitely often on a run is closed under successors.

Furthermore, the set $\mathit{Inf}(w)$ is strongly connected for any path w.

A strongly connected subset of states in a Markov chain that is closed under successors is a BSCC. □

Realization of BSCCs of (finite) Markov chains



Realization of BSCCs of (finite) Markov chains

When computing reachability probabilities, we made the target state g absorbing.

In this way, the only BSCCs are the state g with a self-loop and BSCCs from which g is not reachable.

This shows that indeed either g or a state from which g is not reachable is reached with probability 1.

So, the sets

$$S_0 = \{s \in S \mid g \text{ is not reachable from } s\}$$

and

$$S_1 = \{s \in S \mid S_0 \text{ is not reachable from } s\}$$

are exactly the sets of states from which g is reached with probability 0 and 1, respectively.

Stationary distribution

Dynamical systems often run for a very long time. Hence, it can be interesting to analyze the behavior for when time tends towards ∞.

Example

Recall that we determined that

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0.7 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}.$$

So,

$$P^k = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0.7^k & 0 \\ 0 & 1^k \end{pmatrix} \begin{pmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}.$$

This means

$$\lim_{k\to\infty} P^k = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix} = \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix}.$$

Stationary distribution

Example (continued)

For an arbitrary initial distribution $q = (a, 1-a)$, we observe

$$\lim_{k\to\infty} q_k = \lim_{k\to\infty} q \cdot P^k = (a, 1-a) \cdot \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix} = (1/3, 2/3).$$

Independently of the initial distribution, the system converges to a *stationary distribution*.

The vector $\pi := (1/3.2/3)$ has an interesting property:

$$\pi \cdot P = (1/3, 2/3) \cdot \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix} = (1/3, 2/3) = \pi.$$

So, π is an eigenvector of P with eigenvalue 1 with respect to multiplication from the left. If the Markov chain reaches the distribution over states π, this distribution will be preserved by all future transitions.

Stationary distribution

Definition

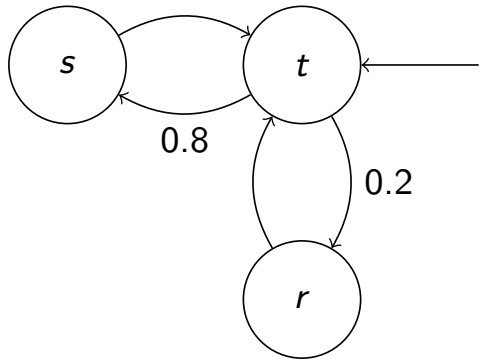
Let P be the transition probability matrix of a Markov chain. A distribution vector π with $\pi \cdot P = \pi$ is called *stationary distribution* of the Markov chain.

Not all Markov chains converge to a unique stationary distribution independently of the initial distribution.

In fact, there are two obstacles: periodicity and reducibility.

Periodicity

Consider the following Markov chain:



We can compute the first members of the sequence q_k of distributions over states after k steps:

$$
\begin{aligned}
q_0 &= (0, 1, 0) \\
q_1 &= (0.8, 0, 0.2) \\
q_2 &= (0, 1, 0) \\
q_3 &= (0.8, 0, 0.2) \\
q_4 &= (0, 1, 0) \\
q_5 &= (0.8, 0, 0.2)
\end{aligned}
$$

Periodicity

We observe $q_{2\ell} = (0, 1, 0)$ for all $\ell \in \mathbb{N}$ while $q_{2\ell+1} = (0.8, 0, 0.2)$. So, $(q_k)_{k\in\mathbb{N}}$ does not converge.

Definition

The *period* of a state j in a Markov chain $M = (S, P, q)$ with state space $S = \{1, \dots, n\}$ is the largest number ξ such that

$$\{k \in \mathbb{N} \mid p_{jj}^{(k)} > 0\} \subseteq \{i \cdot \xi \mid i \in \mathbb{N}\}.$$

A state with period 1 is called *aperiodic*. A Markov chain is called aperiodic if all its states are aperiodic.

Note that $\{k \in \mathbb{N} \mid p_{jj}^{(k)} > 0\} \subseteq \{i \cdot \xi \mid i \in \mathbb{N}\}$. means that all numbers in $\{k \in \mathbb{N} \mid p_{jj}^{(k)} > 0\}$ are divisble by ξ.

Periodicity

For a state j and a natural number k, we have $p_{jj}^{(k)} > 0$ iff there is a path from j to j of length k in the graph of the Markov chain.

Under the following conditions, a state j is certainly aperiodic:

- If there is a self-loop from j to j, i.e., if $p_{jj} > 0$, or
- if there are two paths w_1 and w_2 from j to j such that the length ℓ_1 and ℓ_2 of these paths are co-prime.

For the second condition, observe that in this case

$$\ell_1, \ell_2 \in \{k \in \mathbb{N} \mid p_{jj}^{(k)} > 0\}.$$

There is no $\xi > 1$ such that ξ divides ℓ_1 and ℓ_2 as they are co-prime. Hence, state j is aperiodic.

Periodicity

Lemma

Let $M = (S, P, q)$ be a Markov chain with state space $S = \{1, \dots, n\}$. A state i is aperiodic if and only if there is a natural number k_0 such that

$$p_{ii}^{(k)} > 0$$

for all $k \geq k_0$.

Proof.

As two successive natural numbers are always co-prime, the existence of a k_0 as in the lemma implies that i is aperiodic.

Now, assume that i is aperiodic.

We use the following number-theoretic fact that can be shown using Euclid's algorithm: For natural numbers a and b with greatest common divisor d, there is a natural number ℓ_0 such that, for all natural numbers $\ell > \ell_0$, there are $x, y \in \mathbb{N}$ with $\ell d = xa + yb$.

Periodicity

Proof continued.

Observe that

$$p_{ii}^{(xa+yb)} \geq (p_{ii}^{(a)})^x \cdot (p_{ii}^{(b)})^y.$$

We conclude that if $p_{ii}^{(a)} > 0$ and $p_{ii}^{(b)} > 0$ for $a, b \in \mathbb{N}$ with greatest common divisor d, then there is ℓ_0 such that

$$p_{ii}^{(\ell d)} > 0 \qquad (*)$$

for all $\ell > \ell_0$.

If $d > 1$, then there is a $a' \in \mathbb{N}$ such that $a' \notin d\mathbb{N}$, but $p_{ii}^{(a')} > 0$ because i is aperiodic.

Now, let c be a prime larger than a' and larger than ℓ_0. Then, $p_{ii}^{(c \cdot d)} > 0$ and the greatest common divisor of $c \cdot d$ and a' is a d' less than d.

We can apply this argument inductively until we find numbers a^* and b^* that are co-prime and that satisfy $p_{ii}^{(a^*)} > 0$ and $p_{ii}^{(b^*)} > 0$.

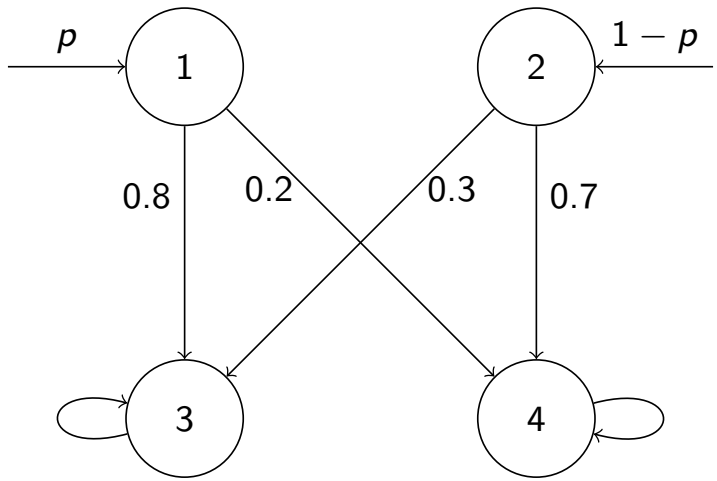
Applying $(*)$ to a^* and b^* finishes the proof. □

Reducibility

Definition

A strongly connected Markov chain is also called *irreducible*. A Markov chain is called *reducible* if it is not strongly connected.

Consider the following Markov chain:



The sequence of distributions $q_0, q_1, \ldots$ over states of this Markov chain is

$$(p, 1-p, 0, 0), (0, 0, 0.3 + 0.5p, 0.7 - 0.5p), (0, 0, 0.3 + 0.5p, 0.7 - 0.5p), \ldots$$

So, this sequence does not converge to a unique stationary distribution independently of the initial distribution.

Cesàro limit

Instead of considering the limit of the sequence of distributions, we can also look at the average probability to be in state s after many steps.

Let $M = (S, P, q)$ be a Markov chain with sequence of distributions $q_0, q_1, \ldots$. We define

$$\theta = \lim_{t \to \infty} \frac{1}{t+1} \sum_{k=0}^{t} q_k.$$

Note that componentwise, this means

$$\theta_j = \lim_{t \to \infty} \frac{1}{t+1} \sum_{k=0}^{t} (q_k)_j.$$

The distribution θ is the so-called *Cesàro limit* of $(q_t)_{t \in \mathbb{N}}$. In finite Markov chains, this limit always exists. It depends on q if M is reducible.

Cesàro limit

Lemma

For the Cesàro limit θ of $(q_t)_{t\in\mathbb{N}}$, we have

$$\theta \cdot P = \theta.$$

Proof.

As

$$\theta = \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} q_k,$$

we observe

$$\begin{aligned}
\theta \cdot P &= \left(\lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} q_k \right) \cdot P \\
&= \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} q_k \cdot P \\
&= \lim_{t\to\infty} \frac{1}{t+1} \sum_{k=0}^{t} q_{k+1} = \theta.
\end{aligned}$$

□

Ergodic Markov chains

The results for the Cesàro limit show that a stationary distribution always exists. Now, we show that aperiodicity and irreducibility ensure that the distributions over states after k steps indeed converge to a stationary distribution and that the stationary distribution is unique and independent of the initial distribution in this case.

Definition

A Markov chain is called *ergodic* if it is irreducible and aperiodic.

Note that a state in a strongly connected Markov chain is aperiodic iff all states are aperiodic.

Theorem (Fundamental theorem for ergodic Markov chains)

Let $M = (S, P, q)$ be an ergodic Markov chain with sequence of distributions $q_0, q_1, \ldots$. Then, there is a unique stationary distribution π with $\pi \cdot P = \pi$ such that

$$\lim_{t \to \infty} q_t = \pi.$$

The distribution π does not depend on q.

Proof.
We know that a stationary distribution π exists.

We fix a stationary distribution π and we will show that for arbitrary states i and j, we have

$$p_{ij}^{(n)} \to \pi_j \text{ for } n \to \infty.$$

This implies that π is unique and does not depend on q.

Ergodic Markov chains

Proof continued.

Let $(X_t)_{t\in\mathbb{N}}$ be the random variables denoting the state in M after t steps and let $(Y_t)_{t\in\mathbb{N}}$ be an independent copy of $(X_t)_{t\in\mathbb{N}}$.

The sequence of pairs $(Z_t)_{t\in\mathbb{N}} := ((X_t, Y_t))_{t\in\mathbb{N}}$ can be seen as a parallel execution of two independent copies of the Markov chain.

Due to the independence, we have

$$\begin{aligned}&\Pr((X_{t+1}, Y_{t+1}) = (j_x, j_y) \mid (X_t, Y_t) = (i_x, i_y))\\ =&\Pr(X_{t+1} = j_x \mid X_t = i_x) \cdot \Pr(Y_{t+1} = j_y \mid Y_t = i_y) = p_{i_x j_x} p_{i_y j_y}.\end{aligned}$$

So, $(Z_t)_{t\in\mathbb{N}}$ is also a Markov chain.

Analogously, we obtain that the probability to reach (j_x, j_y) from (i_x, i_y) in n steps is $p^{(n)}_{i_x j_x} p^{(n)}_{i_y j_y}$.

As M is ergodic, $p^{(n)}_{i_x j_x}$ and $p^{(n)}_{i_y j_y}$ are positive for all sufficiently large n. So, the same holds for $p^{(n)}_{i_x j_x} p^{(n)}_{i_y j_y}$ and hence, the Markov chain $(Z_t)_{t\in\mathbb{N}}$ is ergodic, too.

Proof continued.

The states of the Markov chain N given by $(Z_t)_{t\in\mathbb{N}}$ are given by $S \times S$. We define the set of states

$$E = \{(x, y) \in S \times S \mid x = y\}.$$

As N is ergodic, the probability that E will be reached is 1 from any state. We define the time at which E is reached as a random variable

$$H = \min\{t \in \mathbb{N} \mid Z_t \in E\}.$$

So, $\Pr(H < \infty) = 1$ for any initial distribution.

Proof continued.

We now define the following initial distribution for N: In $Z_0 = (X_0, Y_0)$, we set $X_0 = i \in S$ and let Y_0 be chosen according to the stationary distribution π on S that we fixed. So, for all $t \in \mathbb{N}$ and $s \in S$, $\Pr(Y_t = s) = \pi_s$.

Recall that H is defined such that $X_H = Y_H$. So,

$$\Pr(X_t = s, H \leq t) = \Pr(Y_t = s, H \leq t) \qquad (*)$$

for all $s \in S$ and $t \in \mathbb{N}$.

Now, we obtain as $\Pr(X_0 = i) = 1$,

$$\begin{aligned} |p_{ij}^{(n)} - \pi_j| &= |\Pr(X_n = j) - \Pr(Y_n = j)| \\ &= |\Pr(X_n = j, H \leq n) + \Pr(X_n = j, H > n) \\ &\quad - \Pr(Y_n = j, H \leq n) - \Pr(Y_n = j, H > n)|. \end{aligned}$$

Using $(*)$, we see

$$|p_{ij}^{(n)} - \pi_j| = |\Pr(X_n = j, H > n) - \Pr(Y_n = j, H > n)|.$$

Proof continued.

We observe that $|\Pr(X_n = j, H > n) - \Pr(Y_n = j, H > n)|$ is bounded by $\Pr(H > n)$.

As $\Pr(H < \infty) = 1$, we have $\lim_{n\to\infty} \Pr(H > n) = 0$.

So,

$$\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j. \qquad \square$$

For the transition probability matrix P of an ergodic Markov chain, this means that

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi \\ \vdots \\ \pi \end{pmatrix}$$

where π is the unique stationary distribution (as a row vector).

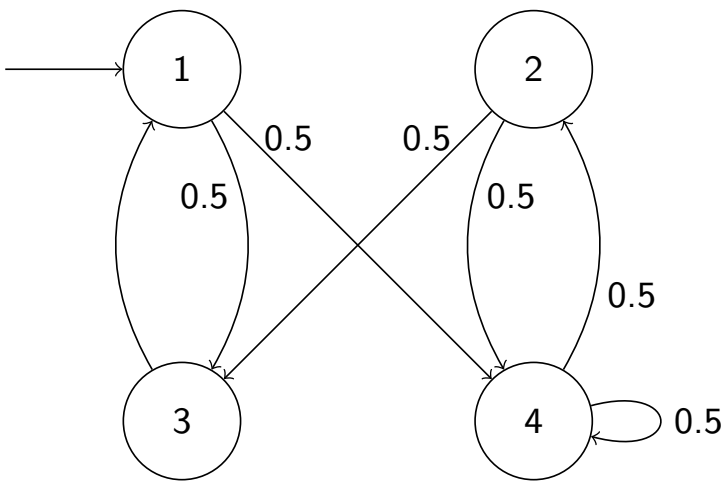
Ergodic Markov chains

One consequence of the fundamental theorem is the following:

In order to compute the unique stationary distribution π of an ergodic Markov chain $M = (S, P, q)$, we can solve the following system of equations

$$\pi = \pi P,$$
$$\sum_{j \in S} \pi_j = 1.$$

Example



$$\pi_1 = \pi_3$$
$$\pi_2 = 0.5\pi_4$$
$$\pi_3 = 0.5(\pi_1 + \pi_2)$$
$$\pi_4 = 0.5(\pi_1 + \pi_2 + \pi_4)$$
$$1 = \pi_1 + \pi_2 + \pi_3 + \pi_4.$$

The solution is $\pi = (0.2, 0.2, 0.2, 0.4)$.

Cesàro limit and steady-state probabilities

If the sequence of distributions $q_0, q_1, \ldots$ of a Markov chain converges to π, this distribution π is also called *steady-state probabilities*.

In aperiodic (but possibly reducible) Markov chains M, the limit $\pi = \lim_{n\to\infty} q_n$ exists.

In this case, π is equal to the Cesàro limit θ of the sequence.

The Cesàro limit θ, which also exists if $q_0, q_1, \ldots$ does not converge can be computed by first solving the following system of equations for each BSCC B.

$$\theta_B = \theta_B P_B,$$

$$\sum_{j\in B} (\theta_B)_j = 1.$$

where θ_B is a vector with one variable per state in B and P_B is the transition probability matrix of the BSCC B.

The Cesàro limit θ can now be compute for states s in a BSCC B by

$$\theta_s = \Pr(B \text{ is reached}) \cdot (\theta_B)_s.$$

All states t not in a BSCC satisfy $\theta_t = 0$.

Expected hitting time

Besides the probability to reach certain states or the long-run behavior of a Markov chain, also expected times until reaching a target set are an important quantity.

Let $M = (S, P, q)$ be a Markov chain and let $G \subseteq S$ be a set of target states.

We are interested in how many steps it takes in expectation to reach the set G. We let $T = \min\{t \in \mathbb{N} \mid X_t \in G\}$ be the random variable that returns the first time point at which G is reached.

If the probability to reach G is not 1, this expected value of T is infinite. It is however, easy to check whether this is the case: G is reached with probability 1 if each reachable BSCC contains a state from G.

So, we will treat the case that G is reached with probabilty 1 in the sequel.

Expected hitting time

Let $M = (S, P, q)$ and $G \subseteq S$ be such that G is reached almost surely in M.

Introduce a variable x_s for each state $s \in S$ and consider the equation system

$$\begin{aligned} x_s &= 0 && \text{for } s \in G, \\ x_s &= 1 + \sum_{t \in S} P(s,t) \cdot x_t && \text{for } s \in S \setminus G. \end{aligned}$$

Then, the unique solution x to the equation system contains the expected values of T starting from each state s, i.e.,

$$x_s = \mathbb{E}_{M_s}(T)$$

where $\mathbb{E}_{M_s}$ denotes the expected value under the probability measure of the Markov chain M where the initial distribution is changed to always choose s.

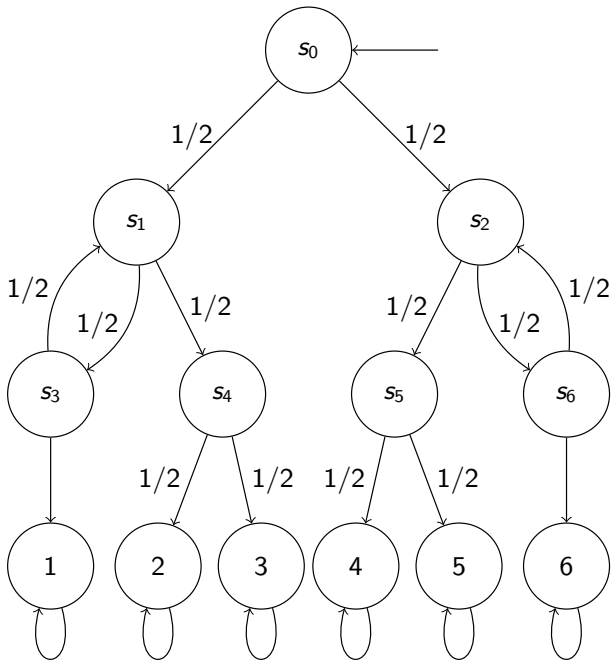
The expected value in M can be computed by multiplying the solution x with the initial distribution q, i.e., $\mathbb{E}_M(T) = q \cdot x$.

Note the similarities with the equation system for reachability probabilities.

Expected hitting time

Example

How many coin tosses does Knuth's die require in average?



For $s \in G = \{1, \dots 6\}$, we set $x_s = 0$.

Plugging in these 0s, we obtain

$$
\begin{aligned}
x_{s_0} &= 1 + 1/2 \cdot x_{s_1} + 1/2 \cdot x_{s_2} \\
x_{s_1} &= 1 + 1/2 \cdot x_{s_3} + 1/2 \cdot x_{s_4} \\
x_{s_2} &= 1 + 1/2 \cdot x_{s_5} + 1/2 \cdot x_{s_6} \\
x_{s_3} &= 1 + 1/2 \cdot x_{s_1} \\
x_{s_4} &= 1 \\
x_{s_5} &= 1 \\
x_{s_6} &= 1 + 1/2 \cdot x_{s_2}
\end{aligned}
$$

As a solution, we get

$$
\begin{aligned}
x_{s_4} = x_{s_5} &= 1, \\
x_{s_3} = x_{s_6} &= 7/3 \\
x_{s_1} = x_{s_2} &= 8/3 \\
x_{s_0} &= 11/3.
\end{aligned}
$$

Expected hitting time

Let us check that the values $\mathbb{E}_{M_s}(T)$ for $s \in S$ indeed form a solution to the equation system.

If $s \in G$, then clearly $\mathbb{E}_{M_s}(T) = 0$.

If $s \in S \setminus G$, then

$$\begin{aligned}\mathbb{E}_{M_s}(T) &= \sum_{t \in S} \Pr_{M_s}(X_1 = t) \cdot \mathbb{E}_{M_s}(T \mid X_1 = t) \\ &= \sum_{t \in S} P(s,t) \cdot (1 + \mathbb{E}_{M_t}(T)) = 1 + \sum_{t \in S} P(s,t) \cdot \mathbb{E}_{M_t}(T).\end{aligned}$$

The uniqueness of the solution can be shown similarly to showing the uniqueness of solutions in the equation system for reachability probabilities.

Expected costs

Instead of considering the time until reaching a target, we can consider more general costs that can be used to model resource and energy consumption for example.

Let $M = (S, P, q)$ be a Markov chain and let $G \subseteq S$ be a set of states that is reached with probability 1. Let

$$c \colon S \setminus G \to \mathbb{R}$$

be a cost function.

As before, we let T be the random variable denoting the time at which G is reached. We define a random variable C for the accumulated cost until reaching the target as

$$C = \sum_{k=0}^{T-1} c(X_k).$$

Expected costs

The expected costs $\mathbb{E}_{M_s}(C)$ for $s \in S$ are the unique solution to the following system of equations

$$\begin{aligned} x_s &= 0 && \text{for } s \in G, \\ x_s &= c(s) + \sum_{t \in S} P(s,t) \cdot x_t && \text{for } s \in S \setminus G. \end{aligned}$$

The uniqueness of the solution can be shown as for expected times as only the constants changed.

Mean payoff

If a system does not reach a terminal state, but runs forever, energy or resource consumption can be measured as a rate. This rate quantifies the average consumption per time step.

Let $M = (S, P, q)$ be a Markov chain with a cost function $c \colon S \to \mathbb{R}$. The mean payoff of a path $w = s_0 s_1 \dots$ is the random vairable

$$mp(w) := \lim_{n \to \infty} \frac{1}{n+1} \cdot \sum_{k=0}^{n} c(s_n).$$

In a Markov chain, this limit exists for almost all paths.

Mean payoff

Definition

The *expected mean payoff* in M with cost function c is the expected value $\mathbb{E}_M(mp)$ of the mean payoff.

The expected mean-payoff can be computed via the Cesàro limit θ of the sequence $q_0, q_1, \ldots$ of distributions of a Markov chain.

We have

$$\mathbb{E}(mp) = \sum_{s \in S} \theta_s \cdot c(s).$$

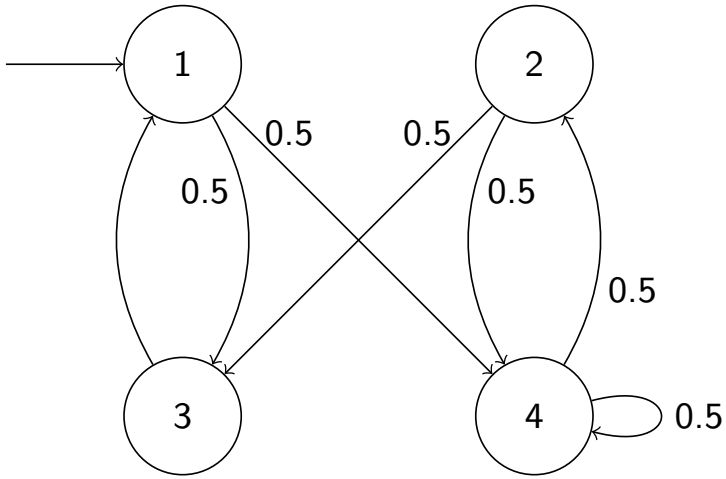
Recall that in strongly connected Markov chains, θ is the unique solution to

$$\theta = \theta \cdot P \quad \text{and} \quad \sum_{s \in S} \theta_s = 1.$$

In reducible Markov chains, we can compute a θ_B for each BSCC B. To obtain θ, the results have to weighted with the probabilities to reach the BSCC B.

Mean payoff

Example



Let the cost function be

$$c(1) = 2, \quad c(2) = 5, \quad c(3) = -1, \quad c(4) = 1.$$

We know the steady-state probabilities

$$\pi = (0.2, 0.2, 0.2, 0.4).$$

So, the expected mean payoff can easily be computed:

$$\mathbb{E}(mp) = \sum_{s \in S} \pi_s \cdot c(s) = 8/5.$$

Outline

Markov decision processes

Presentation on the board.

Not relevant for the exam.