Week 1: Probability Fundamentals

Bayesian Statistics for Machine Learning

Dr Daniel Worrall

AMLab, University of Amsterdam

September 1, 2019

This course

- Schedule
  - **Hoorcollege**: 15-17 Weds & 11-13 Fri in H0.08, **+ 13-15 27/09 in C0.05**
  - **Werk/laptopcollege**: check own timetable **Deadline: Sunday 23:59**
  - 6 EC = 168 hours in total = 28 hours / week
- Assessment
  - 30% **homework & laptop** +35% **midterm** + 35% **final**
  - All homeworks count - 25% penalty per day for late hand-ins, with max cut-off of 2 days. May be waived in sickness (need proof from studieadviseur).
  - Werkcollege not mandatory, but you get 2 points per class
  - Minimum of 5.5 in your homework and a minimum of 5.5 in combined exams to pass the course
- People
  - **Hoorcollege**: Dr Daniel Worrall
  - **Werkcollege**: Dr Gaelle Fontaine `gaelle.marie.fontaine@gmail.com`

All questions should be directed to Gaelle

Homework

- You should upload a photocopy of your homework to Canvas by 11.59pm on Sunday of the week it is set
- Plagarism is not allowed. Once = 0 points, twice = students reported to commission
- Marks comes out the following Friday
- Sick policy: you can miss an extra homework due to sickness (need to have confirmation from study-advisor) and we shall take the average of the remaining homeworks
- No working out, no points

Syllabus/Timetable

1. Probability spaces
   - What is machine learning
   - Classical probability theory
2. Random Variables
   - Random variables
   - Estimators and Moments
3. More Random Variables
   - Moment generating functions
   - Change of variables
5. Maximum likelihood
   - Maximum likelihood estimation
6. Bayesian probability and statistics
   - Bayesian inference
   - MAP estimation
7. Model comparison
   - Bayesian model comparison
   - Review

Books and resources

I have compiled this course from many sources.

- *Probability, Random Processes, and Statistical Analysis*, Hisashi Kobayashi, Brian L. Mark, and William Turin
- *Information Theory, Inference, and Learning Algorithms*, David J. C. Mackay, Ch 2 & 3, (free at `http://www.inference.org.uk/itila/book.html`)
- *Pattern Recognition and Machine Learning*, Christopher Bishop, Ch 1 - 3

I: What is Machine Learning?

I: Supervised learning

We will focus on two areas of machine learning called *supervised learning* and *unsupervised learning*.

Supervised learning problems have 3 parts

- **Data**: inputs **x** and outputs **y**
- **Model space**: a collection of *models* $\mathcal{M}$ which convert inputs into outputs
- **Algorithm**: a method to choose the best model $m \in \mathcal{M}$ from the model space[1], which best *fits* the data

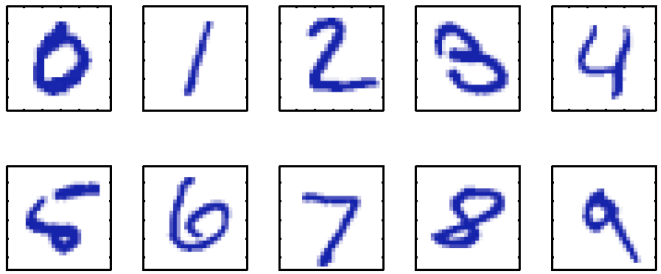
Unsupervised learning is supervised learning, without outputs **y** (more about this later)

[1] The notation '$\in$' is pronounced *in*, so $m \in \mathcal{M}$ is spoken 'm in $\mathcal{M}$'

I: Example 1 - Handwritten digit recognition

Handwritten digit recognition

- **inputs**: 28×28 pixel images, with pixel values in $\{0, ..., 255\}$
- **outputs**: *labels* in $\{0, 1, 2, ..., 9\}$
- **model**: ?



  - **Suggestion**: Collect examples of handwritten digits $\{\mathbf{x}_1, \mathbf{x}_2, ...\}$ with labels $\{\mathbf{y}_1, \mathbf{y}_2, ...\}$ and build a *lookup table*. If new input $\mathbf{x}_* = \mathbf{x}_j$, where $\mathbf{x}_j \in \{\mathbf{x}_1, \mathbf{x}_2, ...\}$, then its *predicted label* is $\mathbf{y}_* = \mathbf{y}_j$.
  - **Problem 1**: What happens if we have never seen $\mathbf{x}_*$ before?
  - **Problem 2**: What happens if we feed in something which is not a number?
  - **Problem 3**: Are there ways in which we can quantify how good our model is?

I: Example 2 - Machine Translation

'Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheuren Ungeziefer verwandelt.'
Franz Kafka, (1915)

'As Gregor Samsa awoke one morning from uneasy dreams he found himself transformed in his bed into a gigantic insect.'
Edwin and Willa Muir, (1933)

'When Gregor Samsa woke up one morning from unsettling dreams, he found himself changed in his bed into a monstrous vermin.'
Stanley Corngold, (1972)

'One morning, upon awakening from agitated dreams, Gregor Samsa found himself, in his bed, transformed into a monstrous vermin.'
Joachim Neugroschel, (1993)

I: Example 2 - Machine Translation

Machine translation is an input–output task

- **inputs**: variable length German strings
- **outputs**: variable length English strings

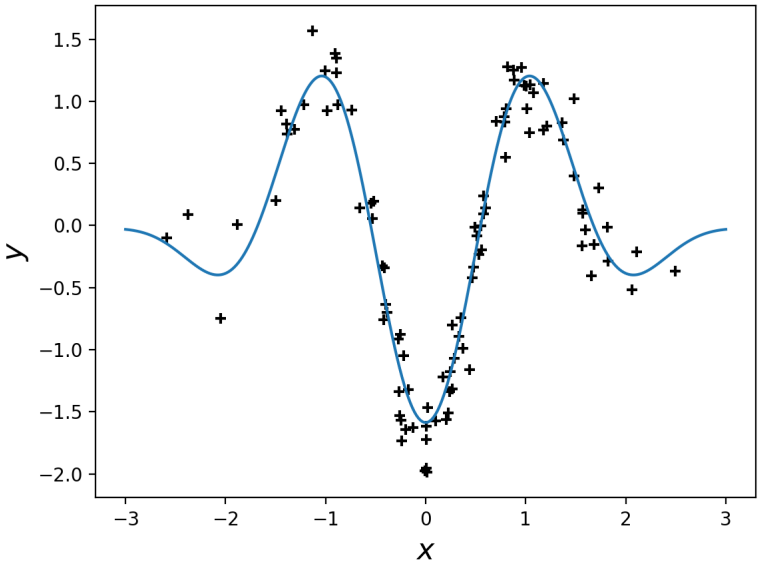
- **Problem 1**: Each word has multiple translations
- **Problem 2**: We cannot possibly collect all input–output pairs
- **Problem 3**: What even is a good translation?

Many of the problems we have seen can be addressed (to some extent) by thinking *probabilistically*. To understand what this means though, we need to learn, what probability is.

I: Example 3 - Regression

Machine translation is an input–output task

- **inputs**: real numbers $x_i \in \mathbb{R}$
- **outputs**: real numbers $y_i \in \mathbb{R}$



- **Problem 1**: How to handle residual error?
- **Problem 2**: Is a linear model the best we can do?
- **Problem 3**: What about higher dimensions?

We will see later in the course that the previous two examples are just variants to of regression (in a very liberal sense).

I - The Nature of Machine Learning

Machine learning is a primarily *conceptual* discipline.

The language of machine learning is mathematics!

- Probability theory
- Calculus and optimization in higher dimensions
- Linear algebra

We implement our mathematical models on computers, which is where programming comes in.

But do not be deceived by popularized preconceptions. Without a firm grasp of the mathematical basics it will be difficult to proceed with just coding!

II: Probability Theory Foundations

Probability versus statistics

- probability: **a mathematical formalism describing uncertain events**
- statistics: **the science of collecting and analysing data**

Bayesian statistics is a branch of statistics loved by machine learners for its computational nature.

Main questions to address:

- Why is this useful?
- What can we say about uncertain events?
- What can be measured?

Intuitive probability

Take a coin. Label heads with 1 and tails with 0. Now flip the coin N times and take the average. Now do this again multiple times.

	Trial 1	Trial 2	Trial3	Trial4	Trial 5
$N = 10$	0.5000	0.8000	0.6000	0.6000	0.2000
$N = 100$	0.4800	0.4800	0.4800	0.5400	0.5400
$N = 1000$	0.4950	0.5130	0.5080	0.5080	0.4850
$N = 10000$	0.4967	0.5031	0.4980	0.4988	0.4934

Despite the fact that in each trial we get a different result, there is a trend!

As $N \to \infty$, what do you think will happen?

Frequentist probability

Probabilities can represent *frequencies*

e.g. Flip a coin N times, define $N(H)$ to be the number of times it lands heads. The *relative frequency* $f_H(N)$ of landing heads is

$$f_H(N) = \frac{N(H)}{N}$$

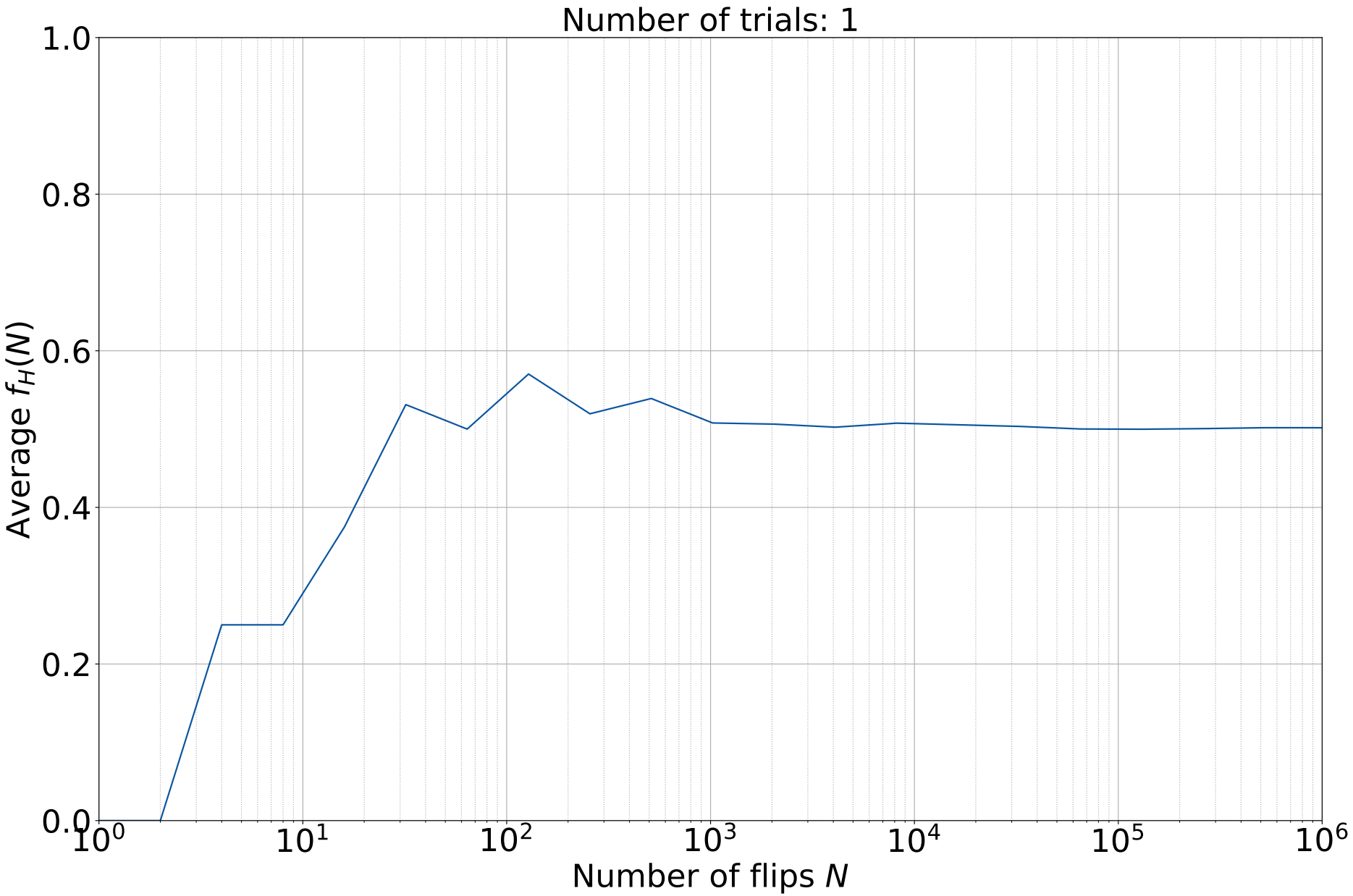
The probability of the coin landing heads, written $P(H)$ is

$$P(H) := \lim_{N \to \infty} f_H(N).$$

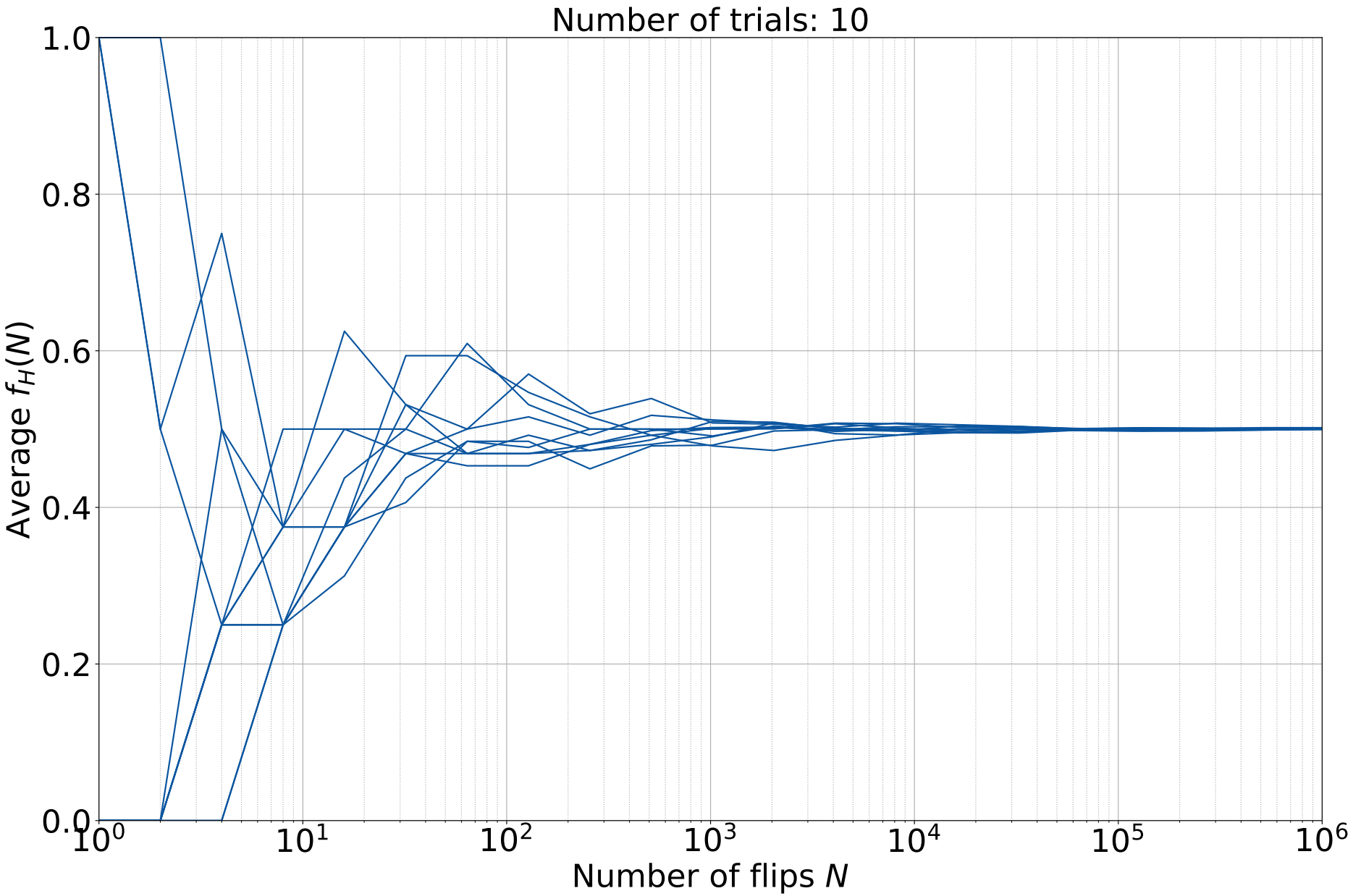
The symbol $\lim_{N \to \infty}$ is called a *limit*. It is the formal way to say "when N gets really really big".

This is the Classical (or frequentist) interpretation of probability: the probability of an *event* is defined as its long run frequency in a repeatable experiment.

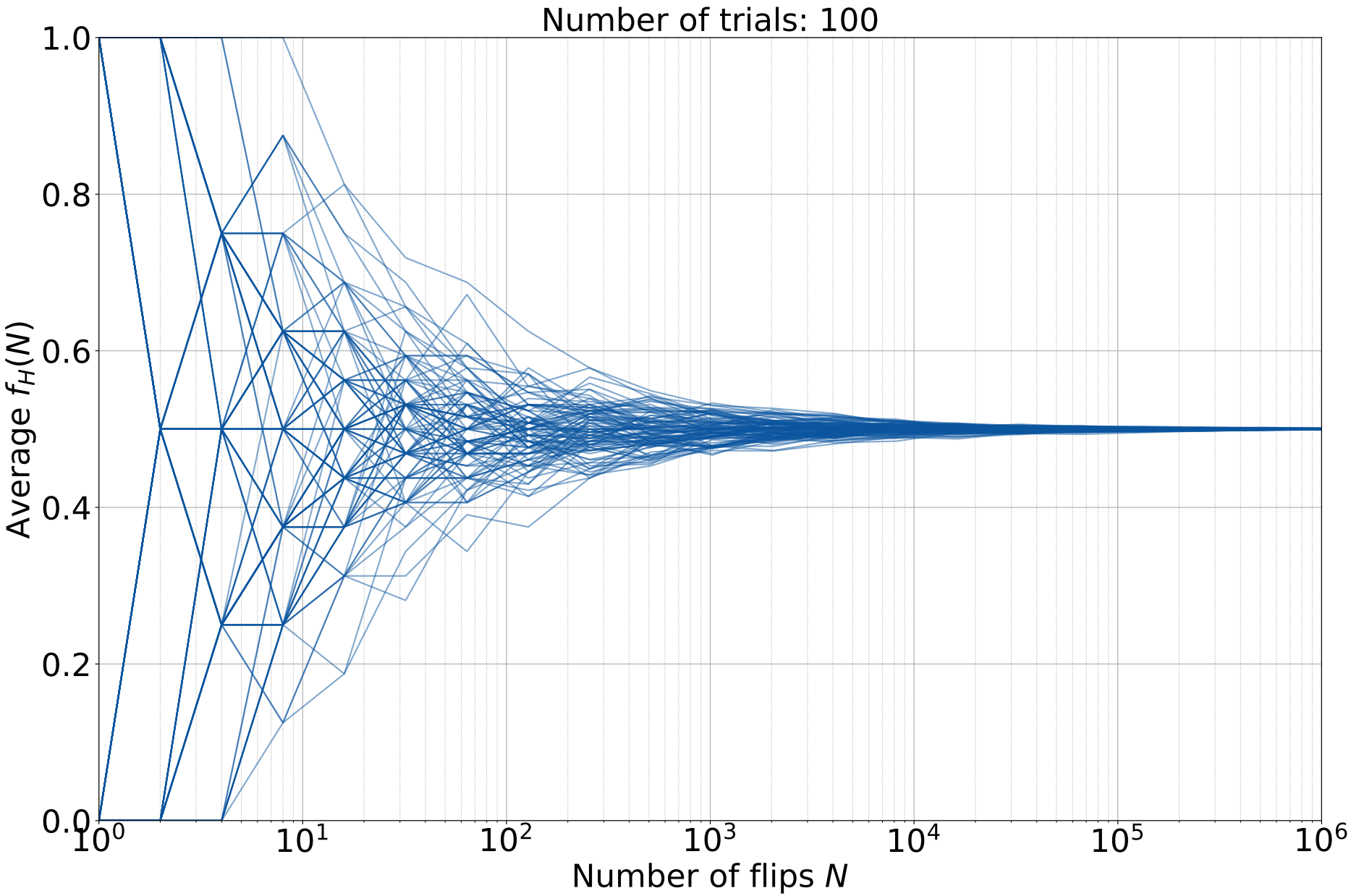
Frequentist probability



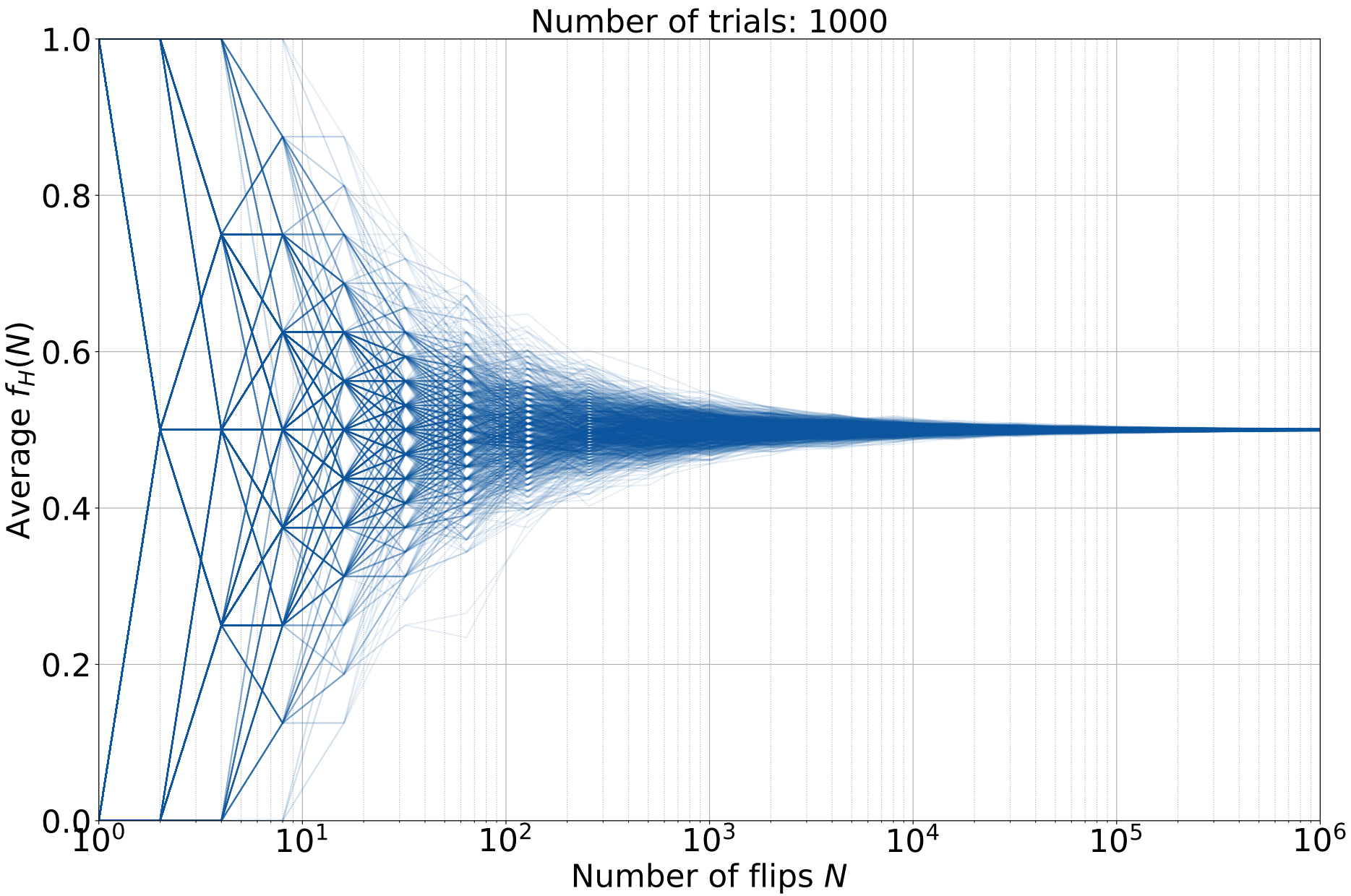
Frequentist probability



Frequentist probability



Frequentist probability



Bayesian probability

Probabilities can represent *beliefs*

e.g. Given the results of a blood test, the probability that Rutger has a nasty disease is $p\%$

e.g. The probability that the UK will leave the EU on Hallowe'en is $q\%$

Such claims cannot be verified through repeated experimentation. This subjective interpretation or *Bayesian* interpretation expresses *degree of belief* .

Both frequentist and Bayesian interpretations of probability are treated with the same theory

The first half of this course is primarily classical/frequentist, the second half is chiefly Bayesian.

Frequentist or Bayesian?

For each of the following scenarios, is the probability frequentist or Bayesian?

Q: You go Leidseplein and drop your wallet, the probability it is still there tomorrow morning is 1%.

Q: In the USA 4 people die a year from vending machine-related accidents: that's a 0.00001 % chance.

Q: You take a statistics course delivered by a (young and intelligent) machine learning researcher, the probability you pass his course is 70%.

Q: The odds of being born with 11 fingers or toes is 0.2%

Sample space

Sample Space[2]

What objects can take probabilities? A *sample space* Ω is a mathematical abstraction, describing the *set* of possible outcomes of an experiment. Each outcome is called a *sample point* or just *sample*[3] $\omega \in \Omega$. We might also use lower case letters $a, b, \ldots$ or $a_1, a_2, \ldots$ to denote samples.

e.g. We flip two coins. Denote heads as h and tails as t. There are four possible outcomes hh, ht, th, and tt. Each outcome is a sample point and

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\} = \{hh, ht, th, tt\}$$

e.g. I wait for a bus at the bus stop. In theory, I could wait forever for the bus, so the sample space is the positive half-line:

$$\Omega = \{\omega : 0 \leq \omega < \infty\}$$

[2] Due to Richard von Mises, who used the German word 'Merkmalraum'

[3] **Notation**: $\omega \in \Omega$ (pronounced ω 'in' Ω) means that ω is an element in the set Ω. e.g. $hh \in \{hh, ht, th, tt\}$, but $-4 \notin \{\omega : 0 \leq \omega < \infty\}$.

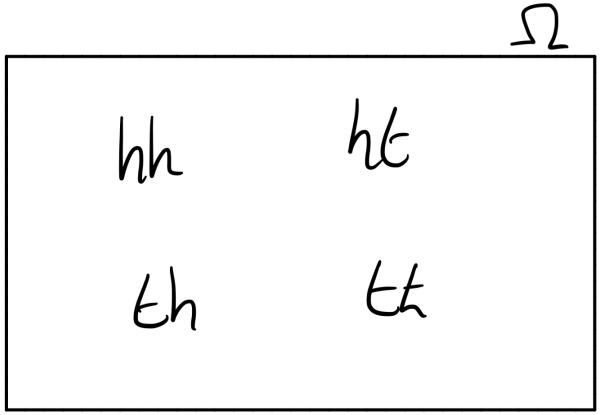
Events

Events

An event $E \subseteq \Omega$ is a set of sample points. We denote it with E or capital letters $A, B, \ldots$ or $A_1, A_2, \ldots$.

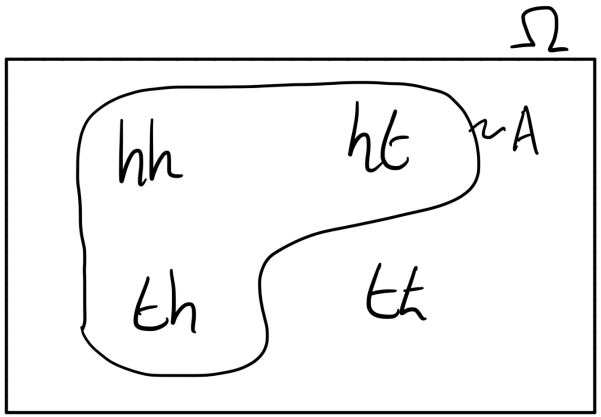
e.g. In the two coin example the event that at least one head is thrown is

$$A = \{hh, ht, th\}$$



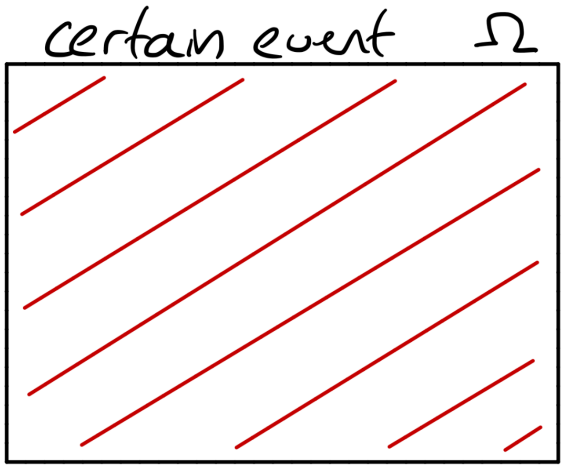
e.g. The event that I wait less than t minutes for my bus is

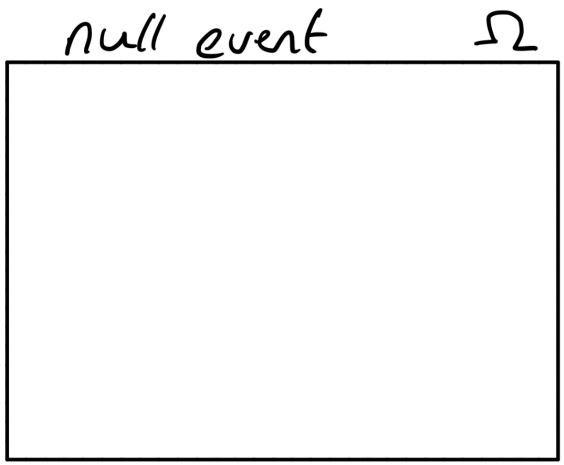
$$E_t = \{\omega : 0 \leq \omega < t \text{ minutes}\}$$

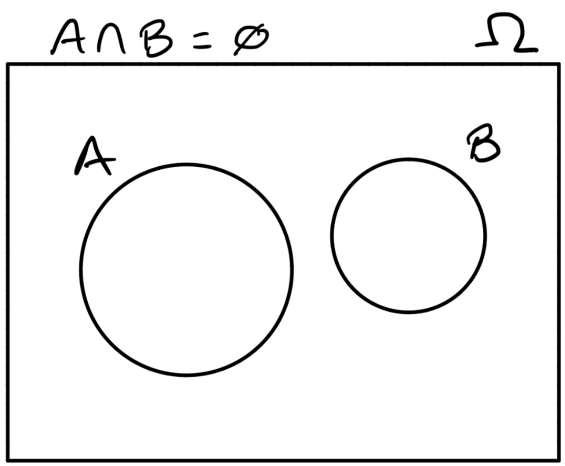


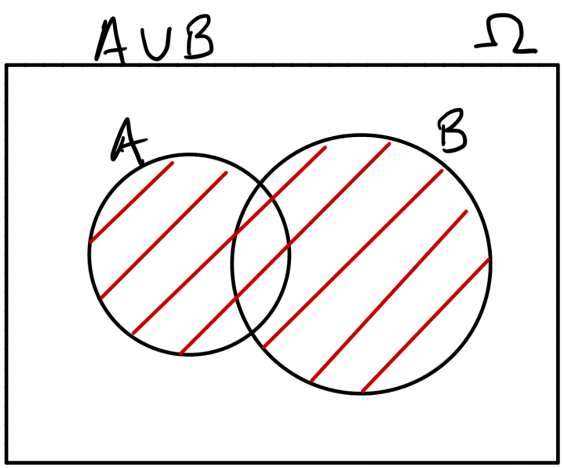
A event consisting of a single sample, e.g. $B = \{hh\}$, is called a *simple event*.

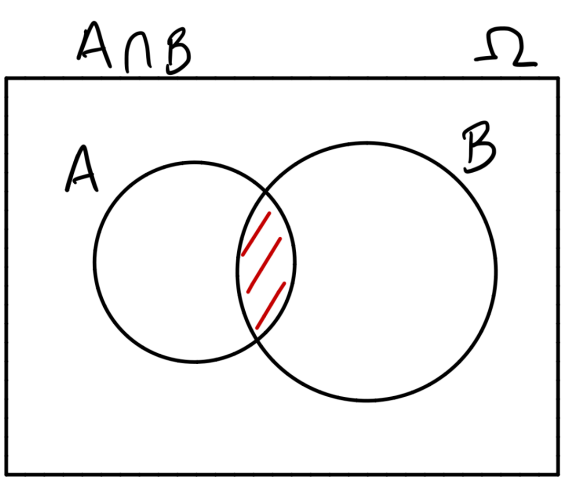
Venn Diagrams

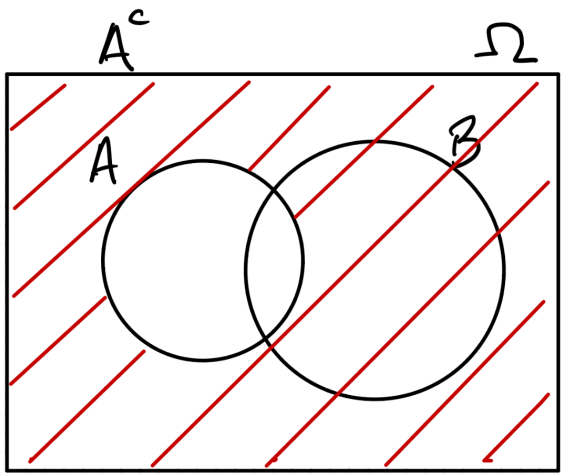












Useful definitions

Useful definitions

1. The *complement* of event A, denoted A^c is all points in Ω except those in A: $A^c := \{\omega : \omega \notin A\}$.
2. The *union* of A and B, denoted $A \cup B$ is the all samples in at least one of A or B: $A \cup B := \{\omega \in A \text{ or } B\}$.
3. The *intersection* of A and B, denoted $A \cap B$ is all samples in both A and B: $A \cap B := \{\omega \in A \text{ and B}\}$.
4. The event containing no samples is the *null event* $\varnothing$, e.g. $A \cap A^c = \varnothing$.
5. The *certain event* is the event containing all samples. It is the sample space, so $A \cup A^c = \Omega$. Clearly $\Omega^c = \varnothing$ and $\varnothing^c = \Omega$.
6. A and B are called *disjoint* or *mutually exclusive* if they have no sample points in common $A \cap B = \varnothing$.

Event spaces

The *power set* $\mathcal{P}(\Omega)$ is the set of all subsets $E \subseteq \Omega$ of set Ω.

e.g. From the set $\Omega = \{a, b\}$ derive the power set.

$$\mathcal{P}(\Omega) = \{\varnothing, \{a\}, \{b\}, \Omega\}$$

A simple way to think of it is as *the set of all events*.

In probability theory, this space is the space, on top of which we shall define probabilities. It is usually given the name of *event space* or *σ-algebra*[4], and the symbol $\mathcal{F}$.

[4] In truth, this is only true in the discrete setting. When the sample space Ω is continuous, the σ-algebra is a subset of Ω, with special conditions beyond the scope of this course.

Probability mass functions

A *probability mass function* (PMF) $P : \mathcal{F} \to [0, 1]$ assigns a number in[a] $[0, 1]$ to every event in the event space $\mathcal{F}$.

- $P(A) = 1$ means that $A \in \mathcal{F}$ is certain
- $P(A) = 0$ means that $A \in \mathcal{F}$ will never happen
- If $P(A) > P(B)$, then A is more likely than B

e.g. In the two coin example we could have

$$P(hh) = P(ht) = P(th) = P(tt) = 1/4.$$

e.g. Distribution of English letters $\Omega = \{a, b, ..., z, -\}$. PMF shown on right with corresponding Hinton diagramme.

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

[a] This means a number between 0 and 1, including both 0 and 1

Probability mass functions

Intuitively, you can imagine that $P(A)$ measures the size[5] of event A in sample space. The bigger the size of A, the bigger $P(A)$. By this logic,

$$P(\Omega) = 1, \quad \text{and} \quad P(\varnothing) = 0.$$

For disjoint events A and B, we have that

$$P(A \dot{\cup} B) = P(A) + P(B)$$

since the size of A plus the size of B is the size of the combined object. In general, if we take $P(A \cup B)$, then we have to discount the size of the intersection

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

You will prove this in the huiswerk.

[5] Indeed, the modern treatment of probability theory is as a branch of the much more general measure theory. Measure theory is the theory of how to assign sizes to mathematical objects.

The Kolmogorov Axioms*

The Kolmogorov Axioms

- The probability of an event is a non-negative real number

$$P(E) \geq 0 \quad \text{for all } E \subseteq \Omega$$

- Certain events have unit probability

$$P(\Omega) = 1$$

e.g. $P(hh) + P(ht) + P(th) + P(tt) = 1$

- Countable additivity: for disjoint events $E_1, E_2, ..., E_N$

$$P(E_1 \dot{\cup} E_2 \dot{\cup} ... \dot{\cup} E_N) = P(E_1) + P(E_2) + ... + P(E_N)$$

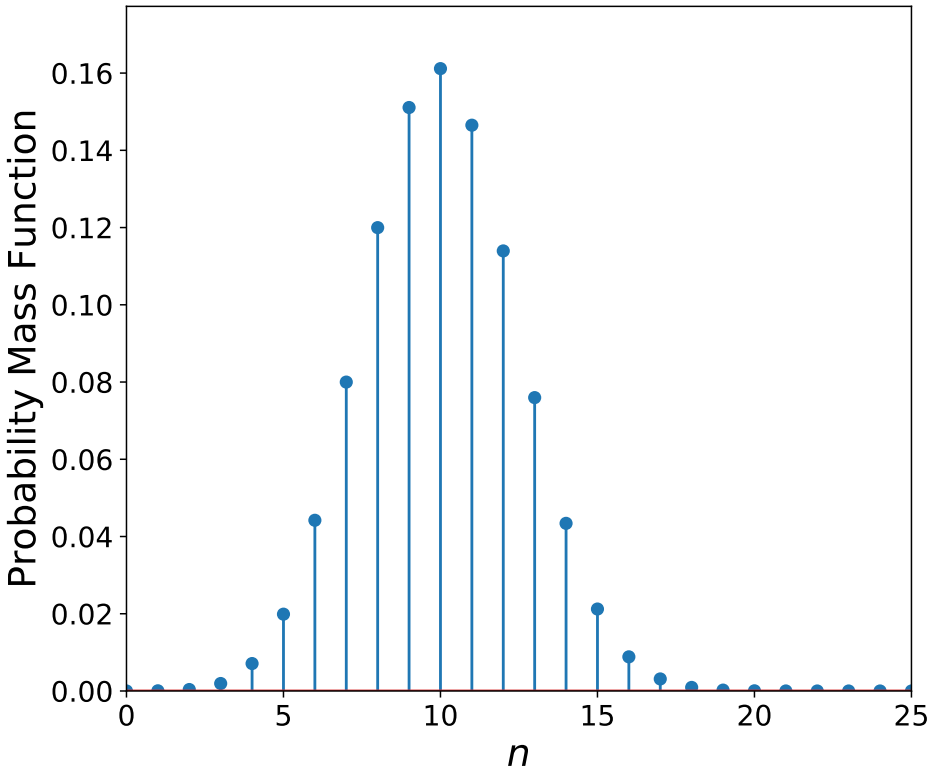
e.g. $P(hh) + P(ht) = 0.5$

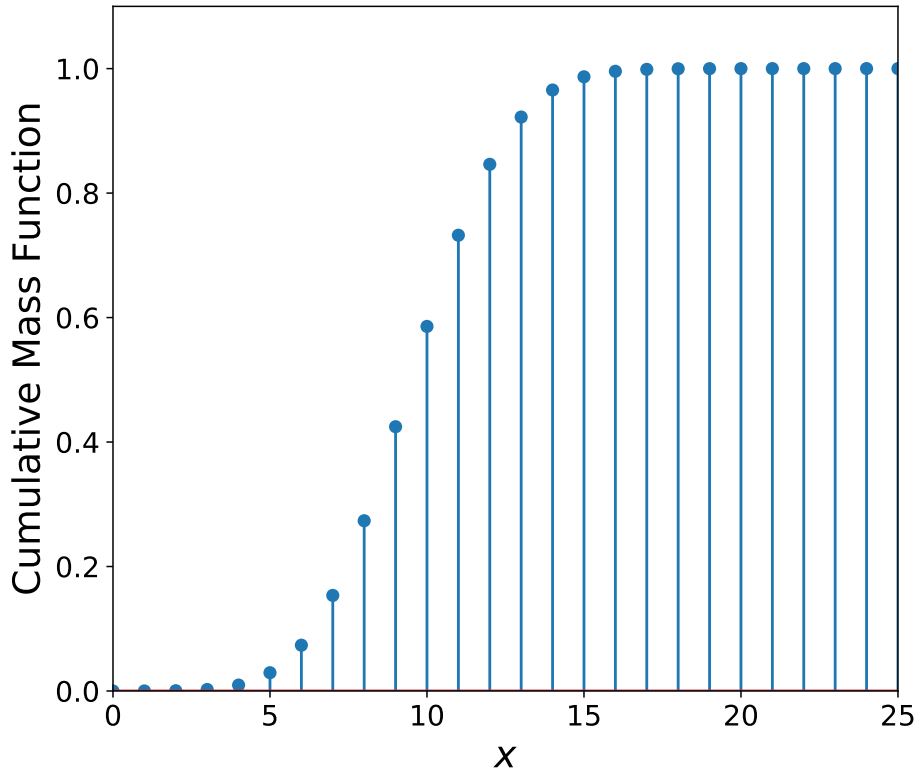
Incredibly all results in probability theory can be derived from various combinations of these 3 axioms[6].

[6] Bayesians actually have their own set of axioms called *Cox's axioms*.

Cumulative mass functions

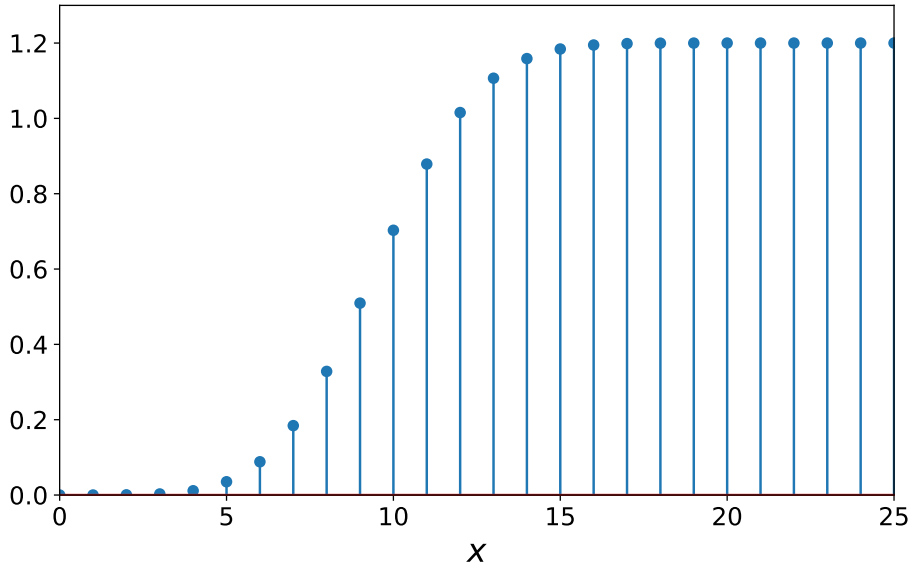
If samples have a natural ordering (e.g. $n \in \{0, 1, 2, 3, ...\}$) the *cumulative mass function* (CMF) F assigns mass to a events of the form $\{n : n \leq x\}$, thus

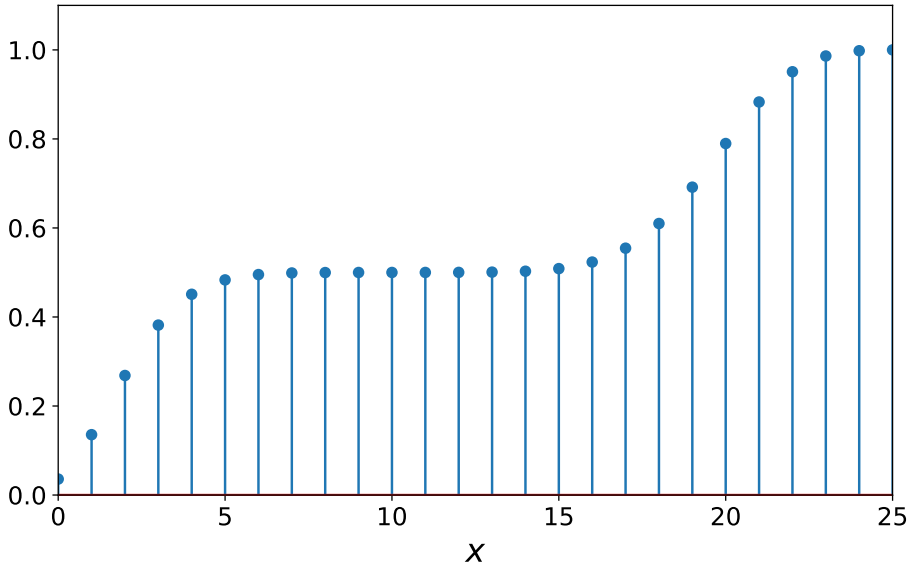


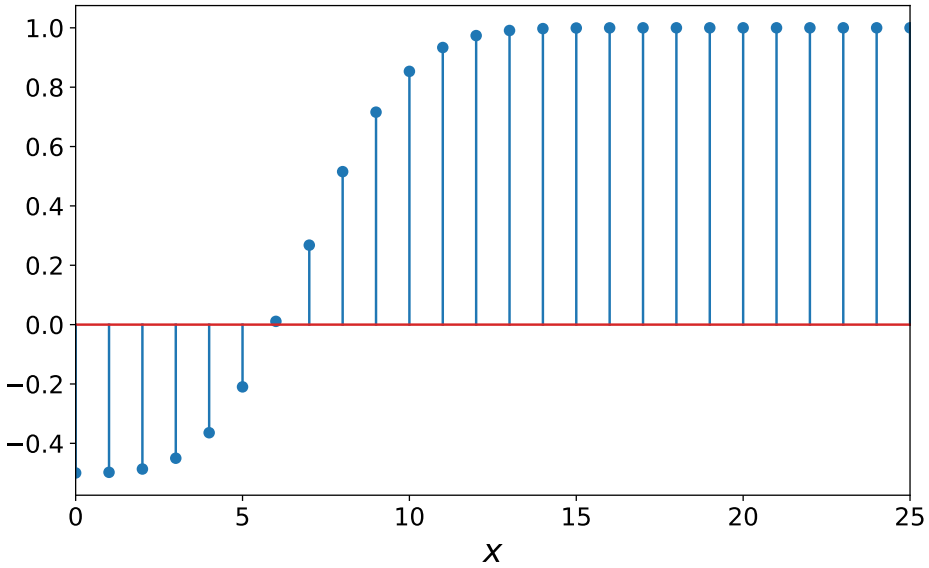


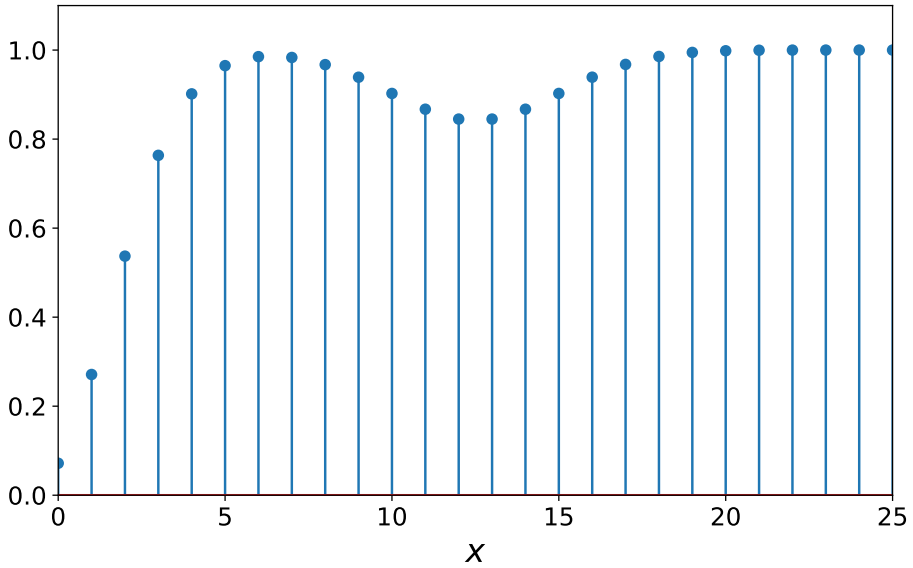
Cumulative mass functions

Which of the following, if any, are valid CMFs?







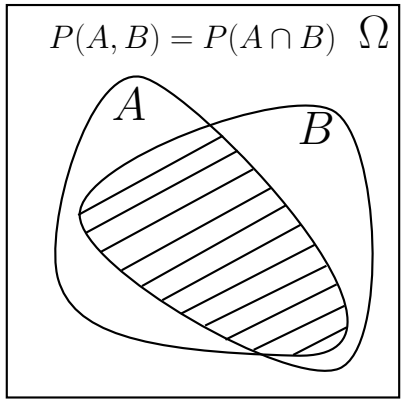


Joint and conditional probability

Joint probability

Given events A and B, the *joint probability* is written

$$P(A, B)$$



It is the probability of A and B happening together. In set-builder notation $P(A, B) = P(A \cap B)$.

e.g. I go to Albert Heijn to buy groceries. Let A be the event that they run out of hummus. Let B be the event I leave my wallet at home. Then $P(A, B)$ is the probability that AH runs out of hummus AND I leave my wallet at home.

The ordering of A and B is unimportant, so

$$P(A, B) = P(B, A).$$

Note that in general $P(A, B) \neq P(A) + P(B)$. Why?

Conditional probability

We write the conditional probability as

$$P(A|B)$$

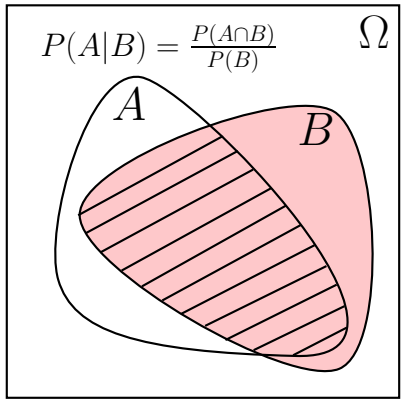
This is the probability of A occurring given that B already has.

e.g. We talk of the probability of event A (me winning the lottery), *given* event B (I purchased a lottery ticket). Clearly in this case $P(A|B^c) = 0$, since you cannot win the lottery without a ticket.

In terms of our size analogy $P(A|B)$ is the *relative size* of $A \cap B$ to B. This is just a ratio

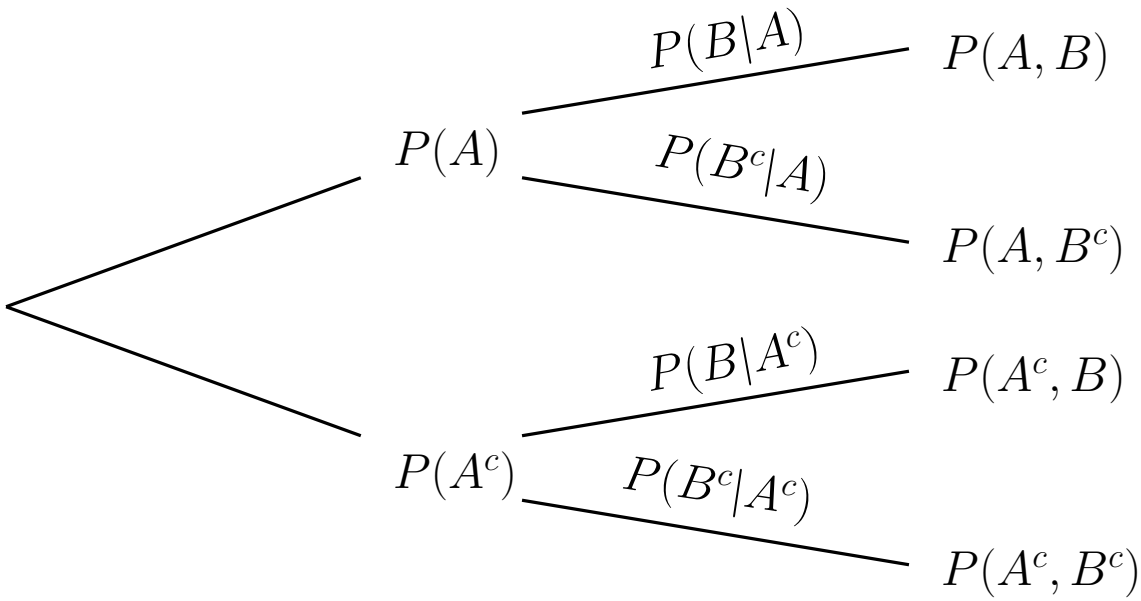
$$P(A|B) = \frac{P(A,B)}{P(B)}.$$

Obviously this only makes sense if $P(B) \neq 0$.



Conditional probability

Joints and conditionals model different parts of a sequence of events



The above *tree diagram* shows possible sequences. The conditional probabilities model the *transition probabilities* passsing from one state to another.

$$P(A,B) = P(B|A)P(A)$$

Conditional probability

e.g. Let $P(A) = 0.3$ be the probability a shop places product X at eye-level on the shelves. Let $P(B|A) = 0.05$ be the probability that a customer puts product X in their basket given that it was at eye-level.

What is the probability X in my basket and it was at eye-level?

$$P(B, A) = P(B|A)P(A) = 0.3 \cdot 0.01 = 0.003$$

e.g. What is[7]

$$\sum_B P(B|A) =$$

e.g. What is

$$\sum_A P(B|A) =$$

[7] The notation $\sum_B$ is a shorthand for sum over every possible value of B

Sum and product rule

To master probability, you only need two rules.

The **product rule**:

$$P(A, B) = P(A|B)P(B).$$

This follows directly from the definition of the conditional probability.

The **sum rule**:

$$P(A) = \sum_B P(A, B).$$

A fancy name for this is *marginalization*, and the term $P(A)$ is refered to as *marginal probability*. We see that it is true from the fact

$$\sum_B P(A, B) = \sum_B \underbrace{P(B|A)P(A)}_{\text{product rule}} = P(A) \underbrace{\sum_B P(B|A)}_{=1} = P(A)$$

Sum and product rule

e.g. Data is stored on a hard drive in binary format. At a given location, the probability that a 0 is stored is 0.47. Random corruption occurs such that 0s sometimes get read as 1s with probability p but a 1 is never read as a 0. What is the marginal probability distribution that a 1 is read?

Denoting an input of x as x_{in}, an output of y as $\mathrm{y}_{\mathsf{out}}$, we have

$$P(1_{\mathsf{out}}|1_{\mathsf{in}}) = 1 \quad P(0_{\mathsf{out}}|1_{\mathsf{in}}) = 0 \quad P(1_{\mathsf{out}}|0_{\mathsf{in}}) = p \quad P(0_{\mathsf{out}}|0_{\mathsf{in}}) = 1 - p \quad P(0_{\mathsf{in}}) = 0.47$$

Using the sum rule

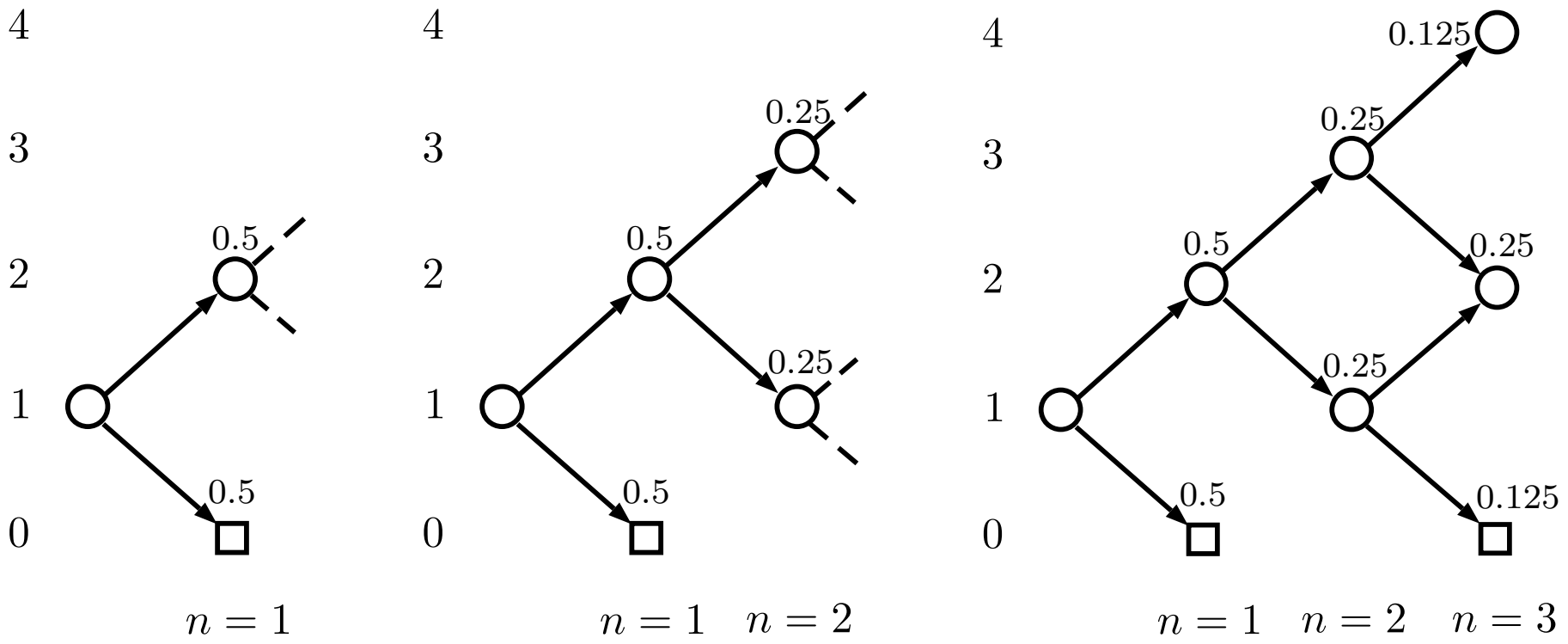
$$\begin{aligned}
P(1_{\mathsf{out}}) &= \sum_{\omega \in \{0_{\mathsf{in}}, 1_{\mathsf{in}}\}} P(1_{\mathsf{out}}, \omega) && \text{(sum rule)} \\
&= \sum_{\omega \in \{0_{\mathsf{in}}, 1_{\mathsf{in}}\}} P(1_{\mathsf{out}}|\omega) P(\omega) && \text{(product rule)} \\
&= P(1_{\mathsf{out}}|0_{\mathsf{in}}) P(0_{\mathsf{in}}) + P(1_{\mathsf{out}}|1_{\mathsf{in}}) P(1_{\mathsf{in}}) \\
&= p \cdot 0.47 + 1 \cdot (1 - 0.47) \\
&= 0.53 + 0.47p
\end{aligned}$$

If $p = 0$, then we read as many 1s are were originally written. If $p = 1$, we only ever read 1s. Otherwise, the number of ones we read is somewhere in between.

Sum and product rule: Random walks

e.g. I flip a coin. If it lands heads, I win €1. If it lands tails, I lose €1. If I have no money, I stop playing. The probability of heads is 0.5. I start with €1.

After $n = 1, 2, 3$ flips, what is the PMF of possible outcomes?



Such diagrams are called *trellises*, they are trees with merges.

Sum and product rule contd.

In general, we can have N variables in a joint, so

$$P(A_1, A_2, A_3, ..., A_N).$$

For instance, I may have N coins, which I flip.

Summing over a variable essentially deletes it from the joint, so

$$P(A_2, A_3, ..., A_N) = \sum_{A_1} P(\cancel{A_1}, A_2, A_3, ..., A_N)$$

and of course you can sum over more than one variable at a time.

What is this equal to?

$$\sum_{A_1} \sum_{A_2} \cdots \sum_{A_N} P(A_1, A_2, A_3, ..., A_N)$$

Sum and product rule contd.

e.g. I flip 3 coins X, Y, Z with probability of heads $0.4, 0.5, 0.55$, respectively. If heads is 1, tails is 0. What is the probability that the sum S of the values is 0, 1, 2, or 3?

If we always write the results of the the coin flips in order $P(X, Y, Z)$ then

$$P(S = 0) = P(0, 0, 0) = 0.6 \cdot 0.5 \cdot 0.45 = \mathbf{0.135}$$

$$\begin{aligned} P(S = 1) &= P(1, 0, 0) + P(0, 1, 0) + P(0, 0, 1) \\ &= 0.4 \cdot 0.5 \cdot 0.45 + 0.6 \cdot 0.5 \cdot 0.45 + 0.6 \cdot 0.5 \cdot 0.55 \\ &= 0.090 + 0.135 + 0.165 = \mathbf{0.39} \end{aligned}$$

$$\begin{aligned} P(S = 2) &= P(1, 1, 0) + P(1, 0, 1) + P(0, 1, 1) \\ &= 0.4 \cdot 0.5 \cdot 0.45 + 0.4 \cdot 0.5 \cdot 0.55 + 0.6 \cdot 0.5 \cdot 0.55 \\ &= 0.090 + 0.110 + 0.165 = \mathbf{0.365} \end{aligned}$$

$$P(S = 3) = P(1, 1, 1) = 0.4 \cdot 0.5 \cdot 0.55 = \mathbf{0.110}$$

Sum and product rule contd.

Likewise, the product rule can be applied to big joints. There are actually many ways to expand the joint because of the symmetry of the arguments e.g.

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\begin{aligned} P(A, B, C) &= P(A|B, C)P(B|C)P(C) \\ &= P(A|B, C)P(C|B)P(B) \\ &= P(B|A, C)P(A|C)P(C) \\ &= P(B|A, C)P(C|A)P(A) \\ &= P(C|A, B)P(A|B)P(B) \\ &= P(C|A, B)P(B|A)P(A) \end{aligned}$$

Bayes' Theorem

Using the product rule, another way to write conditional probabilities is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

This is an **extremely important** rule to "reverse" conditional probabilities.

In many cases we are not given marginal probability $P(B)$, but we can compute it from the sum rule as $P(B) = \sum_A P(B|A)P(A)$, so

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

Bayes' Theorem contd.

e.g. You go to the doctor with trouble breathing, the doctor asks you for a blood sample. A test is run and it returns positive (positive is indicative of cancer). You ask the doctor how likely it is that you really have lung cancer and she says "The blood test returns positive for 95% of all people with lung cancer and 2% for all those who are otherwise healthy. The probability that somebody your age has lung cancer of is 0.01". Unsure what this means, you apply Bayes' rule.

Let $\texttt{cancer}$ be the event you have cancer and $\texttt{pos}$ be the event that the test is positive. We want to know $P(\texttt{cancer}|\texttt{pos})$.

We know

$$P(\texttt{cancer}) = 0.01 \qquad P(\texttt{pos}|\texttt{cancer}) = 0.95 \qquad P(\texttt{pos}|\texttt{cancer}^c) = 0.02$$

So

$$\begin{aligned} P(\texttt{cancer}|\texttt{pos}) &= \frac{P(\texttt{pos}|\texttt{cancer})P(\texttt{cancer})}{P(\texttt{pos})} && \text{(Bayes' theorem)} \\ &= \frac{P(\texttt{pos}|\texttt{cancer})P(\texttt{cancer})}{P(\texttt{pos}|\texttt{cancer})P(\texttt{cancer}) + P(\texttt{pos}|\texttt{cancer}^c)P(\texttt{cancer}^c)} && \text{(sum rule)} \\ &= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.02 \cdot (1 - 0.01)} \simeq 0.324 \end{aligned}$$

III: Probability Densities

Probability on an interval

PMFs are defined for discrete sample spaces, we now turn our attention to *probability density functions* (PDFs), defined on continuous sample spaces.

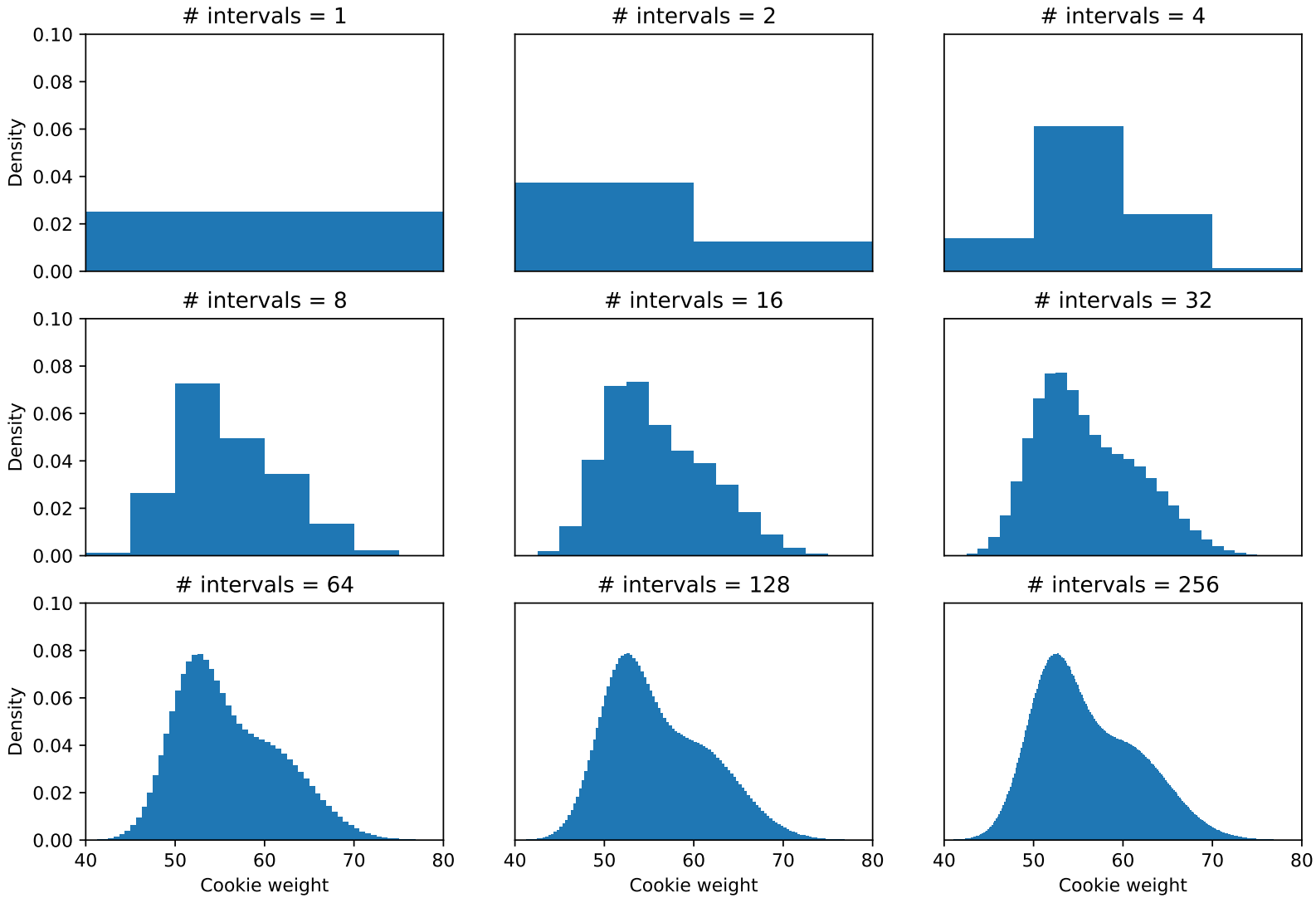
Let's first consider a toy scenario.

Probability of intervals
You buy a box of cookies from the cookie shop. Before you share them with your favourite TAs and esteemed lecturer, you weigh them all. You do this every day for a year and then compute a rough estimate of the distribution of cookie weights.

How do you measure the probability of a continuous value such as weight? With discrete valued outcomes, we simply enumerated every possible outcome and measured the relative frequency $f(N)$ in the limit of large N. We cannot do that if we have an infinite number of outcomes!

Probability on an interval

Answer: discretise the sample space into disjoint (non-overlapping) intervals.



Probability on an interval

Take an interval $(x-\delta, x+\delta)$ and denote the probability mass in this interval by $P(x-\delta < X < x+\delta)$. It is the *area* of a bar!

As $\delta \to 0$, we expect the area to go to zero

$$\lim_{\delta \to 0} P(x-\delta < X < x+\delta) = 0.$$

But this is just $P(x)$ so

$$\boxed{P(x) = 0}$$

We have to rethink probability in continuous spaces.

Cumulative density functions

Cumulative density function

Let's begin by introducing the *cumulative density function* (CDF)

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

It is the probability that an outcome (the weight of our cookie) X is less than some fixed number x. (We already saw the CMF for discrete variables).

Clearly for cookies, X takes values between 0 and ∞, so

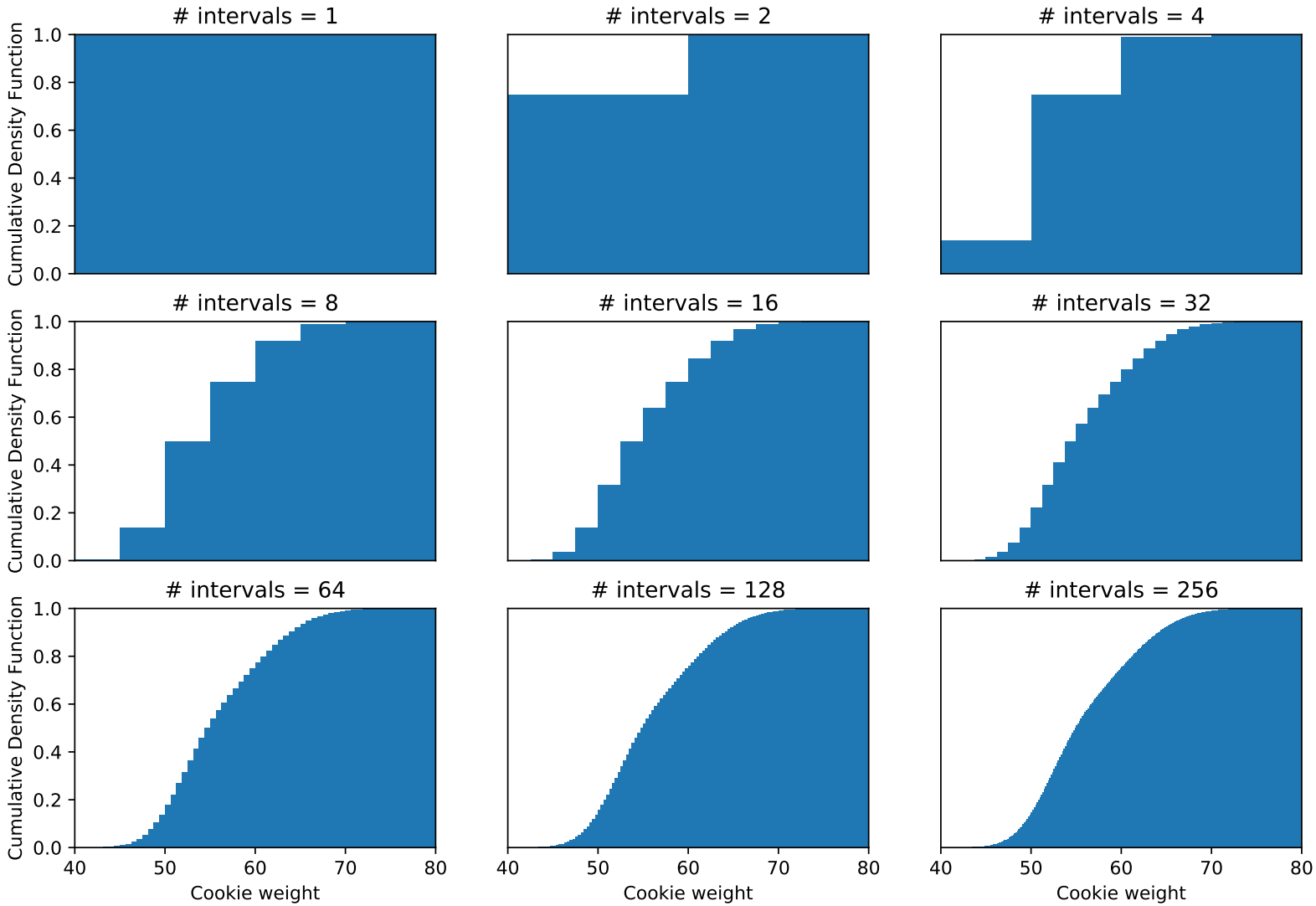
$$F(0) = 0, \qquad F(\infty) = 1$$

On a different sample space, say $X \in \mathbb{R}$, where X may be electronic charge

$$F(-\infty) = 0, \qquad F(\infty) = 1$$

Cumulative Density of Cookies

We also have that $F(x) \leq F(y) \implies x \leq y$, which is called **monotonicity**.



Probability density functions

Probability density function

The *probability density function* (PDF) $p(x)$ is the derivative of the CDF at x.

$$\boxed{p(x) := \frac{\mathrm{d}F}{\mathrm{d}x}}$$

Note that we use a small p for probability *density* and big P for probability *mass*.
We have $p(x) \geq 0$ due to the monotonicity, but this is NOT a probability.

$p(x)$ can be larger than 1.

Exponential distribution

e.g.

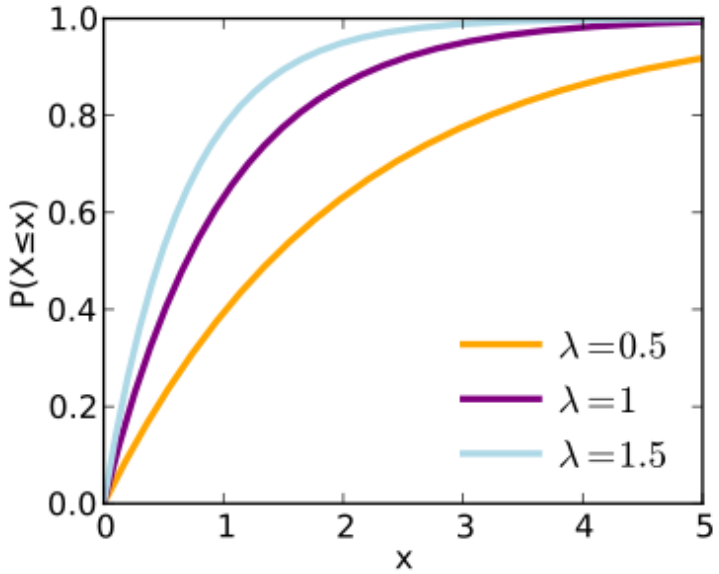
The CDF of the exponential distribution is

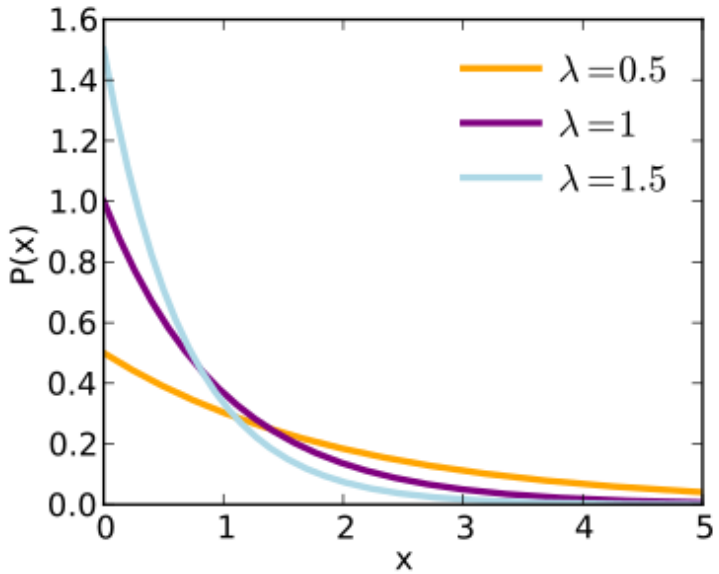
$$F(x) = 1 - e^{-\lambda x} \qquad x \leq 0, \lambda > 0$$

The PDF is

$$\begin{aligned} p(x;\lambda) &= \frac{\mathrm{d}F(x)}{\mathrm{d}x} \\ &= \frac{\mathrm{d}}{\mathrm{d}x}\left[1 - e^{-\lambda x}\right] \\ &= \lambda e^{-\lambda x} \qquad x \leq 0, \lambda > 0 \end{aligned}$$

This is a popular distribution for modelling waiting times. It has a parameter λ. It controls the shape of the distribution





Probability density functions

From PDF to CDF

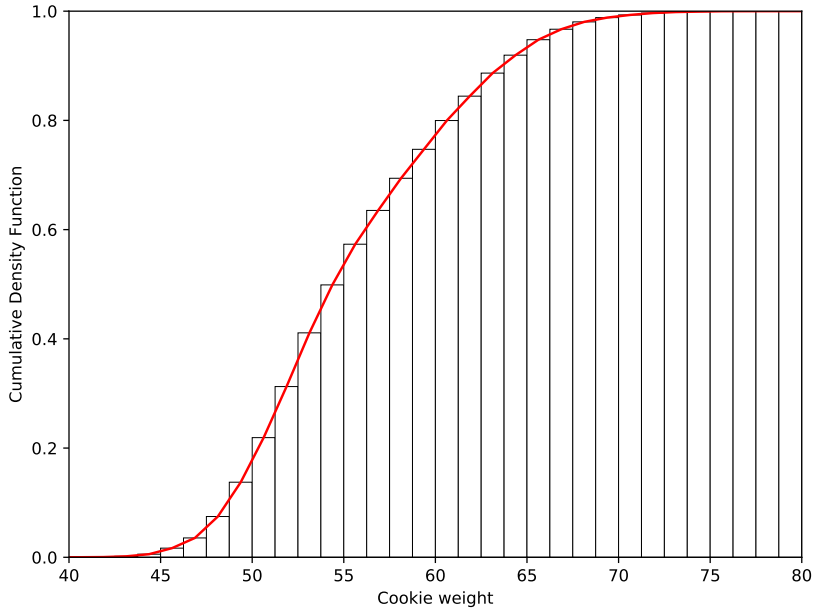
We can regain the CDF from the PDF as

$$F(x) = \int_{-\infty}^{x} p(x') \, \mathrm{d}x'$$

Integrals return the area under a curve.

It is the infinite sum of infinitesimally wide columns of height $p(x)$ and width $\mathrm{d}x$.

The probability of each column is $p(x) \, \mathrm{d}x$.



More densities

We can use the PDF to measure the probability of x being in some interval $[a, b]$

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a) \\ &= \int_{-\infty}^{b} p(x)\,\mathrm{d}x - \int_{-\infty}^{a} p(x)\,\mathrm{d}x \\ &= \int_{a}^{b} p(x)\,\mathrm{d}x \end{aligned}$$

PDFs can be used in joints; conditionals; marginals; Bayes' rule; and the sum and product rules, just like PMFs.

More densities: Normalization

A PDF $p(x)$ is not a probability, but

- $p(x) \geq 0$ for all x since CDFs are monotonic non-decreasing
- $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$ since $F(\infty) = 1$

e.g. If $\beta > 1, x \geq 1$, find the normalization constant Z for

$$p(x;\beta) = \frac{1}{Z} x^{-\beta}$$

Well

$$\int_1^{\infty} \frac{1}{Z} x^{-\beta}\,\mathrm{d}x = \frac{1}{Z}\left[\frac{1}{1-\beta} x^{1-\beta}\right]_1^{\infty} = \frac{1}{Z}\left[0 - \frac{1}{1-\beta}\right] = \frac{1}{Z(\beta-1)} = 1$$

$$\implies Z = \frac{1}{\beta - 1}$$

More densities: Normalization

e.g. If $x \in \mathbb{R}$, find the normalization constant Z for

$$p(x;\beta) = \frac{1}{Z} e^{-(\beta x + e^{-\beta x})}$$

Substitute $u = e^{-\beta x}$, so $\mathrm{d}u = -\beta e^{-\beta x}\,\mathrm{d}x$

$$\int_{-\infty}^{\infty} \frac{1}{Z} e^{-(\beta x + e^{-\beta x})}\,\mathrm{d}x = \frac{1}{Z}\int_{-\infty}^{\infty} e^{-e^{-\beta x}} e^{-\beta x}\,\mathrm{d}x = -\frac{1}{Z}\frac{1}{\beta}\int_{\infty}^{0} e^{-u}\,\mathrm{d}u$$

$$= \frac{1}{Z}\frac{1}{\beta}\left[-e^{-u}\right]_0^{\infty} = \frac{1}{Z}\frac{1}{\beta} = 1$$

$$\implies Z = \frac{1}{\beta}$$