ADA511 0.3 Data Science and AI prototyping

P.G.L. Porta Mana S. Mæland

2025-08-21

Table of contents

Dear student and aspiring data- & AI-engineer

If you can't join 'em,
beat 'em.
(J. Schwinger)

The goal of this course is not to help you learn how to tune the parameters of the latest kind of deep network, or how to choose a good prompt for a Large Language Model, or how to do cross-validation in the fastest way, or what is the latest improvement in random-forest algorithms.

The goal of this course is to help you learn the principles to build the machine-learning algorithms and AI devices *of the future*. And, as a side effect, you'll also learn how to concretely improve present-day algorithms, and also how to determine if any of them has already reached its maximal theoretical performance.

How can such a goal be achieved?

There is a small set of rules and one method that are *mathematically guaranteed* to output the optimal solution of any inference, prediction, classification, and decision-making problem. You can think of this set as defining an "unbeatable, optimal universal machine". Or, from an AI point of view, you can think of these rules and method as the "laws of robotics" that should govern any ideal AI designed to draw inferences, give answers, and make decisions.

These rules and method are quite easy to grasp and understand. You'll learn them very quickly, and they'll be the solid ground on which your data & AI engineering knowledge and skills are built.

$$\mathrm{P}(\neg Y \mid X) = 1 - \mathrm{P}(Y \mid X)$$

$$\mathrm{P}(Z \wedge Y \mid X) = \mathrm{P}(Z \mid Y \wedge X) \cdot \mathrm{P}(Y \mid X)$$

$$\mathrm{P}(Z \vee Y \mid X) = \mathrm{P}(Z \mid X) + \mathrm{P}(Y \mid X) - \mathrm{P}(Z \wedge Y \mid X)$$

$$\mathrm{P}(X \mid X \wedge Z) = 1$$

$$\texttt{choose} \quad \operatorname{argmax}_D \sum_Y \mathrm{U}(Y \wedge D \mid X) \cdot \mathrm{P}(Y \mid D \wedge X)$$

These rules and method are computationally extremely expensive; the more so, the more data points and data dimensions we need to deal with. Current machine-learning algorithms, from deep networks to large language models, are *approximations* to this ideal universal method; each one uses a different kind of approximation. The upside of these approximations is that they allow for much faster computations; their downside is that they generally give *sub-optimal* or *non-intelligent* results.[1]

But approximations can be improved with new technologies. The approximations used at any given time in history exploit the computational technologies then available. Deep networks, for instance, would have been a useless approximation 50 years ago, before the introduction of Graphical Processing Units.

Every new technological advance (think of possibly forthcoming quantum computers) opens up possibilities for new approximations that get us closer and closer to the ideal optimum. However, in order to *see* and *realize* these possibilities, or to judge whether they have already been realized, a data scientist needs at the very least:

⛰ to know the foundation of the maximally optimal method

🤓 to think outside the box

Without the first requirement, how do you know what is *the target* to approximate towards, and how far you are from it? You risk:
⚠ making an approximation that leads to worse results than before;
⚠ evaluating the approximation in the wrong way, so you don't even realize it's worse than before;
⚠ trying to improve an approximation that has already attained the theoretical optimum. Think about an engine that has already the maximal efficiency dictated by thermodynamics; and an engineer, ignorant of thermodynamics, who wastes effort in trying to improve it further.

Without the second requirement, you risk missing to take full advantage of the new technological possibilities. Consider the

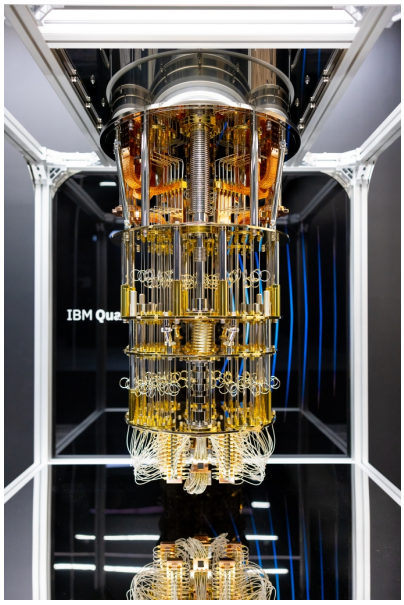
[1] Is a suboptimality of, say, just 0.1% important? In a life-or-death situation for 1 000 000 people, 0.1% means 1000 more deaths.

evolution of transportation: if you keep thinking in terms of how to improve a horse-carriage wooden wheels, you'll never conceive a combustion engine. If you keep thinking in terms of how to improve combustion fuel, you'll never conceive an electric motor. Existing approximations may of course be good starting points; but you need to clearly understand how they approximate the ideal optimum – so we're back to the first requirement.

If you want to make advances in machine learning and AI, you must know how the ideal universal algorithm looks like, and you must not limit yourself to thinking of "training sets", "cross-validation", "supervised learning", "overfitting", "models", and similar notions. In this course you'll see for yourself that such notions are anchored to the box of present-day approximations.

And we want to think outside that box.

This course will not only prepare you for the future. With the knowledge and insights acquired, you will be able to devise and *implement concrete improvements* to present-day methods as well, or calculate whether they can't be improved further.



What new ways of doing data science will quantum computers lead to?

Your role in the course Bugs & features

This course is still in an experimental, "alpha" version. So you will not only learn something from it (hopefully), but also test it together with us, and help improving it for future students. Thank you for this in advance!

For this reason it's good to clarify some goals and guidelines of this course:

Undergraduate maths requirements We believe that the fundamental rules and methods can be understood and also used (at least in not too complex applications) without complex mathematics. Indeed the basic laws of inference and decision-making involve only the four basic operations $+ - \times /$. So this course only requires maths at a beginning first-year undergraduate level.

Informal style The course notes are written in an informal style; for example they are not developed along "definitions", "lemmata", "theorems". This does not mean that they are inexact. We will warn you about parts that are oversimplified or that only cover special contexts.

Names don't constitute knowledge

https://vimeo.com/852936507?share=copy

In these course notes you'll often stumble upon **terms in blue bold** and *definitions in blue Italics*. This typographic emphasis does ***not*** mean that those terms and definitions should be ***memorized***: rather, it means that there are important ***ideas*** around there which you must try to ***understand*** and ***use***. In fact we don't care which terminology you adopt. Instead of the term **statistical population**, feel free to use the term **pink apple** if you like, as long you explain the terms you use by means of a discussion and examples.[2] What's important is that you know, can recognize, and can correctly use the ideas behind technical terms.

Memorizing terms, definitions, and where to use them, is how large language models (like chatGPT) operate. If your study is just memorization of terms, you'll have difficulties finding jobs in the future, because there will be algorithms that can do that better and at a cheaper cost than you.

Diverse textbooks This course does not have only one textbook: it refers to and merges together parts from several books and articles. As you read these works, you will notice that they adopt quite different terminologies,

[2]Some standard technical terms are no better. The common term *random variable*, for instance, often denotes something that is actually *not "random"* and *not variable*. Go figure. Using the term *green banana* would be less misleading!

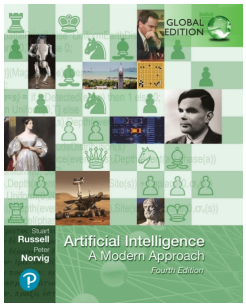
employ different symbolic notations, give different definitions for similar ideas, and sometimes even contradict each other.

These differences and contradictions are a feature, not a bug!

You might think that this makes studying more difficult; but it actually helps you to really understand an idea and acquire real knowledge, because it forces you to go *beyond* words, symbols, and specific points of view and examples. This point connects with the previous point, "names don't constitute knowledge". The present course notes will help you build comprehension bridges across those books.

Artificial intelligence In order to grasp and use the fundamental laws of inference and decision-making, we shall use notions that are also at the foundations of Artificial Intelligence (and less common in present-day machine learning). So you'll also get a light introduction to AI for free. Indeed, a textbook that we'll draw frequently from is Russell & Norvig's *Artificial Intelligence: A Modern Approach* (we'll avoid its part V on machine learning, however, because it's poorly explained and written).

Concrete examples Some students find it easier to grasp an idea by starting from an abstract description and then examining concrete examples; some find it easier the other way around. We try to make both happy by alternating between the two approaches. Ideas and notions are always accompanied by examples that we try to keep simple yet realistic, drawing from scenarios ranging from glass forensics to hotel booking.



Code We shall perform inferences on concrete datasets, also comparing different methodologies. Most of these can be performed with any specific programming language, so you can use your favourite one – remember that

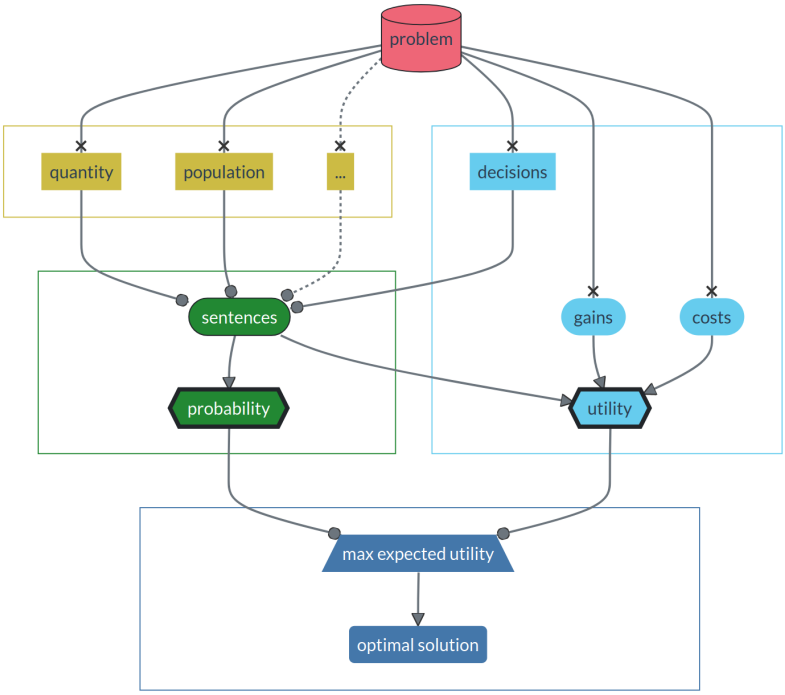
we want to try to think outside the box of present-day technologies, and that includes present-day programming languages. Most examples in class and in exercises will be given in `R` and sometimes in `Python`, but are easily translated into other languages.

Extra material The course has strong connections with many other disciplines, such as **formal logic**, **proof theory**, **psychology**, **philosophy**, **physics**. We have tried to provide a lot of extra reading material in "For the extra curious" side boxes, for those who want to deepen their understanding of topics covered or just connected to the present course. Maybe you'll stumble into a new passion or even into your life call?

For the extra curious

Course structure

The course structure reflects the way in which the ideal universal decision-making machine works. It can be roughly divided into three or four parts, illustrated as follows (this is just a caricature, don't take this diagram too literally):



- **Data** parts (top-left, yellow box) develop the language in which a **problem** can be fed into the decision-making machine. Here you will also learn about important pitfalls in handling data.
- **Inference** parts (left-centre, green box) develop the "inference engine" of the machine. Here you will learn ideas at the foundation of AI; and you will also meet probability, but from a point of view that may be quite novel to you – and much more fun.

These two parts will alternate so that their development proceeds almost in parallel.

- The **utility** part (top-right, light-blue box) develops the "decision engine" of the machine. Here you will meet several ideas that will probably be quite new to you – but also very simple and intuitive.
- The **solution** part (bottom, dark-blue box) simply shows how the inference and utility engines combine together to

yield the optimal solution to the problem. This part is simple, short, intuitive; it will be a breeze.

We shall start with a quick preview of the **solution** part in chapters 1–4, because it is very simple to understand, and it shows why the *inference* and the *utility* parts are necessary.

Then we shall continue with the **inference** part in chapters 5–10 and 14–18, alternating it with the **data** part in chapters 12–13 and 20–23, and with interludes about **present-day machine-learning algorithms** and their approximations in chapters 4, 11, 19, and 24–26.

As soon as the inference and data parts are complete, you will be able to apply the machine to real, albeit not too complex, inference problems. This **application** will be made in chapters 32–35.

We finally round up with the **utility** part in chapters 36–37, extending our concrete application to it in chapter 38. Final connections with present-day machine learning are made in chapters 39–40.

You should be able to see this timeline in the index tab on the side.

Preface

Science is built up with facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house. (H. Poincaré)

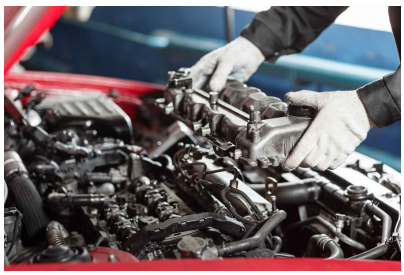
Mechanics and engineers

What is the difference between a *car mechanic* and an *automotive engineer*?

Both have knowledge about cars, but their knowledge domains are different and focus on different goals.

A car mechanic can keep your car in top-notch condition; can do different kinds of easy and difficult repairs if problems arise with it; knows whether a particular brand of valve can be used as a replacement for another brand; can recommend the optimal kind of tyres to use in a given season for different brands of cars. But a car mechanic would face difficulties in calculating the theoretical maximal efficiency of an engine; or predicting the temperature increase caused by a new kind of fuel; or exploiting the phase transition of a new kind of foam to design a safer airbag system; or calculating the optimal surface curvature for a spoiler. A car mechanic typically possesses a large amount of case-specific knowledge, and doesn't need to know in depth the principles of electromechanics and thermochemistry, or the laws of balance of momentum, energy, entropy.

Vice versa, an automotive engineer can assess how to use the electromechanical properties of a new material in order to design a more efficient and environmentally friendly engine; can



calculate how a new material-surface handling would affect air drag and speed; and ultimately can research how to exploit new physical phenomena to build completely new means of transportation. Yet, an automotive engineer could be completely incapable of changing a pipe in your car, or tell you whether it can use a particular brand of lubricant oil. An automotive engineer typically possesses knowledge about the principles of electromagnetism, mechanics, or thermochemistry; is acquainted with relevant physical laws; and doesn't need to have in-depth case-specific kinds of knowledge.

Note that the differences just sketched *do not imply a judgement of value*. Both professions, kinds of knowledge, and goals are necessary, interesting, and couldn't exist without each other. Choice between them is a subjective matter of personal passions and aspirations.

In fact there isn't a clear divide between these two kinds of knowledge, but rather a continuum between two vague extremities. A car mechanic can have knowledge and insight about new technologies, and an automotive engineer can know how to fix a carburettor. The two sketches above are meant to expose and emphasize the existence of such a continuum of knowledge and of goals.

Data mechanics and data engineers

A continuum with two similar extremities can also be drawn in **data science**.

Some data scientists have in-depth knowledge on, for instance, how to optimally store and read large amounts data; what kind of machine-learning algorithm to use in a given task; how to fine-tune an algorithm's parameters, and the currently best software for this purpose. Their particular knowledge is fundamental for the working of today's technological infrastructure.

At the same time, these data scientists typically face difficulties, for instance, in:

- calculating the theoretical maximal accuracy or performance achievable – by *any* possible algorithm – in a given inference problem
- explaining how the fundamental rules of inference and decision-making are implemented in a particular machine-learning algorithm
- identifying which sub-optimal approximations to the fundamental rules are made by popular machine-learning algorithms
- exploiting new technologies to build new algorithms that do calculations closer to the exact theoretical ones, thereby achieving a performance closer to the theoretical optimum

And it is also possible that they are not aware of, and maybe would be surprised by, some basic facts of data science. For instance:

- there is an optimal, universal inference & decision algorithm, of which all machine-learning algorithms (from support vector machines and deep networks to random forests and large language models), are an approximation
- there are only five or six fundamental laws upon which any inference, prediction, classification, regression, decision task is (or ought to be) based upon
- splittings of data into "training set", "validation set", and similar sets, are not part of the exact application of the laws of inference and decision-making; such splittings arise as coarse approximations of the exact method.
- cross-validation and related techniques are not part of the exact method either; they also arise as approximations
- overfitting, underfitting and related notions are not problems that appear in the exact method (which takes care of them automatically); they also arise from approximations

- it is possible to calculate, within probable bounds, the maximal accuracy (or other performance metric) achievable by *any* classification or regression algorithm for a given application
- some evaluation metrics, such as precision or the area under the curve of the receiver operating characteristic (AUC), have intrinsic flaws and may attribute higher values to worse-performing algorithms

…because this is a kind of general and principled knowledge that these data scientists don't need in their jobs. Their knowledge is more case-specific.

Drawing a parallel with the car example, a data scientist with this kind of case-specific knowledge is like a "data mechanic".

A "data engineer", on the other hand, is the kind of data scientist who has no difficulties with the knowledge and skills implicit in the bullet points above; but at the same time might not know what software to use for tuning parameters of a particular class of deep networks, or the best format to store particular kinds of data.

Just like in the case of the automotive industry, the difference just sketched does *not* imply any judgement of value. Both kinds of knowledge and goals are important and can't exist without each other.

Goals of this course

There is a plethora of academic courses, in all kinds of format, that target knowledge and goals for the "data mechanic". Those courses are usually inadequate to cover the knowledge and goals for the "data engineer". Some courses, misleadingly, even present approximations and recipes that are only valid for particular situations as if they were universal rules or methods instead.

Courses that target the "data engineer" seem to be more rare. One possible reason is that this kind of knowledge is actually hidden in courses on probability, statistics, and risk analysis,

presented with a language which makes only opaque and confusing connections with fields in data science and their goals; or, worse, with a language which emphasizes connections that are actually superficial and misleading.

We believe that it is important to teach and keep alive the less "mechanic" and more "engineer" side of data science:

- Continuous advances in computational technology – think of quantum computers – will offer completely novel and superior ways to approximate the exact method of inference and decision. Only the data scientist who knows the exact method and theory, and understands how present-day algorithms approximate it, will be able to exploit new technologies.
- Even without looking at the future, several present-day machine-learning algorithms could already be greatly optimized by any data engineer who is acquainted with the basic principles underlying data science.
- The foundations of data science are the bridge to the sibling discipline of Artificial Intelligence.

The present course aspires to give an introduction to the "data engineer" side, rather than "data mechanic" one, of data science, but using a point of view more familiar to data scientists than to, say, statisticians.

More details about its aims, structure, and features are already given in the *Dear student* introduction.

Part I

An invitation

1 Accept or discard?

Let's start with a question that could arise in an engineering problem:

> A particular kind of electronic component is produced on an assembly line. At the end of the line there is an *automated inspection device* that works as follows on each new component:
>
> - The inspection device first performs some tests on the component. The tests give an uncertain forecast of whether that component will fail **within its first year of use**, or **after**.
> - Then the device decides whether the component is **accepted** and packaged for sale, or **discarded** and thrown away.
>
> Consider also the following context. When a new electronic component is sold, the manufacturer has a net gain of 1$. That's the net gain if the component works *for at least a year*. But if the component instead *fails within a year* of use, the manufacturer incurs a *net loss* of 11$ (12$ loss, minus the 1$ gained at first), owing to warranty refunds and damage costs to be paid to the buyer. When a new electronic component is *discarded*, the manufacturer has 0$ net gain.
>
> Now we have a new electronic component, just come out of the assembly line. The tests of the automated inspection device indicate that there is a **10% probability** that the component will **fail within its first year** of use.

Should the inspection device accept the new component? or discard it?

Try to give and motivate an answer:

Very first exercise!

- Should the inspection device accept or discard the new component?

It doesn't matter if you don't get the correct answer; not even if you don't manage to get an answer at all. The purpose here is for you to do some introspection about your own reasoning.
Then examine and discuss the following points:

- Which numerical elements in the problem seem to affect the answer?
- Can these numerical elements be clearly separated? How would you separate them?
- How would the answer change, if these numerical elements were changed? Feel free to change them, also in extreme ways, and see how the answer would change.
- Could we solve the problem if we didn't have the probabilities? Why?
- Could we solve the problem if we didn't know the various gains and losses? Why?
- Can this problem be somehow abstracted, and then transformed into another one with completely different details? For instance, consider translating along these lines:
  - inspection device → computer pilot of self-driving car
  - tests → camera image
  - fail within a year → pedestrian in front of car
  - accept/discard → keep on going/ break

2 Framework

2.1 What does the intro problem tell us?

Let's approach the "accept or discard?" problem of the previous chapter 1 in an intuitive way.

First, what happens if we `accept` the component?

We're jumping the gun here, because we haven't learned the method to solve this problem yet!

We must try to make sense of the 10% probability that the component fails within a year. For the moment let's use an imagination trick: imagine that the present situation is repeated 100 times. In 10 of these repetitions the accepted electronic component is sold and fails within a year after selling. In the remaining 90 repetitions, the component is sold and works fine for at least a year. Later on we'll approach this in a more rigorous way, where the idea of "imaginary repetitions" is not needed.

In each of the 10 imaginary repetitions where the component fails early, the manufacturer loses 11\$. That's a total loss of $10 \cdot 11\$ = 110\$$. In each of the 90 imaginary repetitions in which the component doesn't fail early, the manufacturer gains 1\$. That's a total gain of $90 \cdot 1\$ = 90\$$. So over all 100 imaginary repetitions the manufacturer gains

$$10 \cdot (-11\$) + 90 \cdot 1\$ = -20\$$$

that is, the manufacturer has not gained, but *lost* 20\$! That's an average of 0.2\$ *lost* per repetition.

Now let's examine the second choice: what happens if we `discard` the component instead?

In this case it's clear that the manufacturer doesn't gain or lose anything. That is, the "gain" is 0$ (this is for sure, so we don't need to imagine any "repetitions").

The conclusion is this: If in a situation like the present one we accept the component, then we'll lose 0.2$ on average. Whereas if we discard it, then we'll lose 0$ on average.

Obviously the best, or "least worst", decision to make is to **discard** the component.

Exercises

1. Now that we have an idea of the general reasoning, check what happens with different values of the probability of failure and different values of the cost of failure. Is it still best to discard? For instance, try with

   a. failure probability `10%` and failure cost `5$`;
   b. failure probability `5%` and failure cost `11$`;
   c. failure probability `10%`, failure cost `11$`, non-failure gain `2$`.

   Feel free to get wild and do plots.

2. Identify the probability of failure for which there is no loss or gain, on average, if we accept the component (so it doesn't matter whether we discard or accept). You can solve this as you prefer: analytically with an equation, visually with a plot, by trial & error on several cases, or whatnot.

3. Consider the special case with failure probability `0%` and failure cost `10$`. That probability means that no new component will ever fail. It's clear what's the optimal decision in this limit case, without any calculations or imaginary repetitions. Yet, **confirm mathematically** that we arrive at this obvious conclusions if we perform a mathematical analysis like before.

4. Consider this completely different problem:

A patient is examined by a brand-new medical diagnostics AI system.

First, the AI performs some clinical tests on the patient. The tests give an uncertain forecast on whether the patient has a particular disease or not.

Then the AI decides whether the patient should be dismissed without treatment, or treated with a particular medicine.

If the patient is dismissed, then their life expectancy doesn't increase or decrease if the disease is not present, but it decreases by 10 years if the disease is actually present. If the patient is treated, then their life expectancy decreases by 1 year if the disease is not present (owing to treatment side-effects), but also if the disease is present (because it cures the disease, so the life expectancy doesn't decrease by 10 years; but it still decreases by 1 year owing to the side effects).

For this patient, the clinical tests indicate that there is a 10% probability that they have the disease.

Should the diagnostic AI dismiss or treat the patient? Find differences and similarities, even numerical, with the assembly-line problem.

From the solution of the problem and from the exploring exercises, we gather some instructive points:

- Is it enough if we simply know that the component is less likely to fail than not? In other words, is it enough to know that the probability of failure is *less than 50% without knowing its precise value*?

  Obviously not. We found that if the failure probability

is 10% then it's best to discard. But we also found that if it's 5% then it's best to accept. In either case the probability of failure was less than 50%, but the decision was different.

On top of that, we also found that the probability value determines the average amount of loss when the non-optimal decision is made. Therefore:

👉 **Knowledge of *precise* probabilities is absolutely necessary for making the best decision.**

- Is it enough if we simply know that failure leads to a loss, and non-failure leads to a gain, without knowing the precise amounts of loss and gain?

  Obviously not. In the exercise we found that if the cost of failure is 11$, then it's best to discard. But we also found that if it's 5$, then it's best to accept (given the same probability of failure). And we also found that it's best to accept when the cost of failure is 11$ but the gain from non-failure is 2$. Therefore:

  👉 **Knowledge of the precise gains and losses is absolutely necessary for making the best decision.**

- Is this kind of decision situation only relevant to assembly lines and sales?

  By all means not. We examined a clinical problem that's exactly analogous: there's uncertainty and probability, there are gains and losses (of lifetime rather than money), and the best decision depends on both probabilities and costs.

2.2 Our focus: decision-making, inference, and data science

Every data-driven engineering problem is unique, with unique difficulties, questions, issues. But there are some general aspects that are common to all engineering problems.

In the scenarios that we explored above, we found an extremely important problem-pattern:
- There is a decision or choice to make (and "not deciding" is not an option, or it's just another kind of choice).
- Making a particular decision will lead to some consequences. Some consequences are desirable, others are undesirable.
- The decision is difficult to make, because its consequences are not known with certainty, even considering the information and data available in the problem: we may lack information and data about past or present details, about future events and responses, and so on.

This is what we call a problem of **decision-making under uncertainty** or **under risk**[1]; or simply a "decision problem" for short.

This problem-pattern appears literally everywhere. Stop for a second, and think about all different situations in which you had to make a decision today. Do they show this pattern?

But our exploration of different scenarios also suggests something important: this problem-pattern seems to have a sort of systematic method of solution!

In this course we're going to focus on decision problems and their systematic solution method. We'll learn a framework and some general notions that allow us to frame and analyse this kind of problems. And we'll learn a universal set of principles to solve it. This set of principles goes under the name of **Decision Theory**.

But what do decision-making under uncertainty and Decision Theory have to do with *data* and *data science*? The three are profoundly, tightly connected on many different planes:

[1] We'll avoid the word "risk" because it has several different technical meanings in the literature, some even contradictory.

- *Data science* is based on the laws of *Decision Theory*. These laws are similar to what the laws of physics are to a rocket engineer. Failure to account for these fundamental laws leads to sub-optimal solutions – or to disasters.
- *Machine-learning* algorithms, in particular, are realizations or approximations of the rules of *Decision Theory*. This is clear, for instance, considering that a machine-learning classifier is actually *choosing* among possible output classes.
- The rules of *Decision Theory* are also the foundations upon which *artificial-intelligence agents* – which must perform optimal inferences and decisions – are built.
- We saw that *probability* values are essential to a decision problem. How do we find them? Obviously *data* play an important part in their calculation. In our introductory example, the failure probability must have come from observations or experiments on previous similar electronic components.
- We saw that the values of *gains and losses* are essential. *Data* play an important part in their calculation as well.

These five planes will constitute the major parts and motivations of the present course.

🚀 For the extra curious

Decision theory in expert systems and artificial intelligence

There are other important aspects in engineering problems, besides the one of making decisions under uncertainty. For instance the *discovery* or the *invention* of new technologies and solutions. Aspects such as these can barely be planned or decided. Their drive and direction, however, rest on a strive for improvement and optimization. But the fundamental laws of Decision Theory tell us what's optimal and what's not, so they play some part in these creative aspects as well.

Artificial intelligence is proving to be a valuable aid in these creative aspects. This kind of use of AI is outside the scope of the present notes. But some aspects of this creativity-assisting

use *do* fall within the domain of the present notes. A pattern-searching algorithm, for example, can be optimized by means of the method we are going to study.

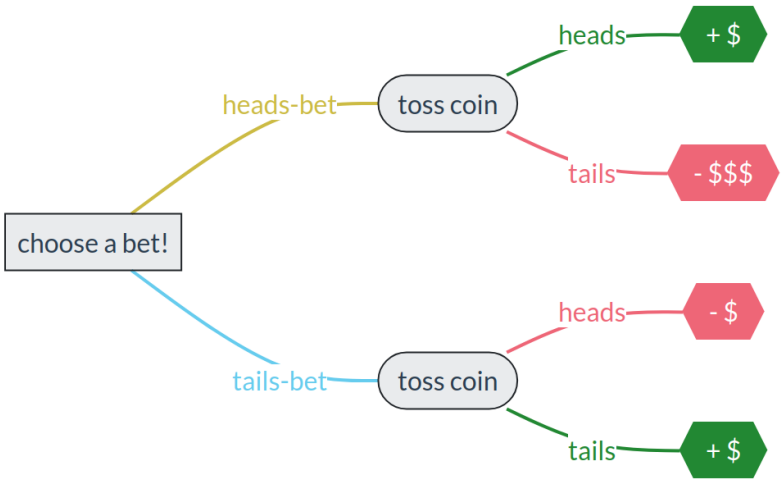
2.3 Our goal: optimality, not "success"

What should we demand from a systematic method for solving decision problems?

By definition, in a decision problem under *uncertainty* there is generally no method to determine the decision that *surely* leads to the desired consequence. If such a method existed, the problem would not have any uncertainty! Therefore, if there is a method to deal with decision problems, its goal cannot be the determination of the *successful* decision. Then what should be the goal of such a method?

Imagine two persons, Henry and Tina, who must choose between a "heads-bet" or a "tails-bet" before a coin is tossed. The bets are these:

- "heads-bet": If the coin lands heads, the person wins a *small* amount of money. But if it lands tails, they lose a *large* amount of money.
- "tails-bet": If the coin lands tails, the person *wins* a small amount of money. If it lands heads, they lose the same *small* amount of money.



Exercise

Which bet would you choose? why?

Now this happens: Henry chooses the heads-bet. Tina chooses the tails-bet. The coin comes down heads. So Henry wins the small amount of money, while Tina loses the same small amount.

What would we say about their decisions?

Henry's decision was lucky, and yet *irrational*: he risked losing much more money than he could win. Tina's decision was unlucky, and yet *rational*: she wasn't risking to lose more than she could win. Said otherwise, the heads-bet had higher risk of loss than the tails-bet, and not even an higher chance of gain. We expect that any person making Henry's decision in similar, future bets will eventually lose more money than any person making Tina's decision.

The method we're looking for is therefore one that, in the hypothetical situation above, would lead to the same decision as Tina's – even if Tina's decision was unlucky. That's the decision that we call rational or *optimal* in such an uncertain situation.

If you're thinking *"wouldn't it be best to have a method that works under uncertainty but that leads to Henry's decision, every time that decision is lucky?"* – then let's repeat: **such a method cannot logically exist**. If we know which decision is "lucky", then it means that we have no uncertainty. If we are uncertain, then it means that we don't know which decision is "lucky", and so it's impossible to choose it for sure.

Our discussion and the distinction between "successful" and "optimal" decisions also shows that *we cannot evaluate the efficacy of a method for decisions under uncertainty, by checking whether or how often that method leads to the desired, "successful" consequence*. This point is also easily illustrated with a variation on Henry and Tina's example:

Suppose the general context and the bets are exactly the same. But now imagine HENRY and TINA to be the names of two automated decision methods, say two machine-learning algorithms. Also, let's say that you first toss the coin in secret and see its outcome, then you offer the possible bets to HENRY and TINA, who are completely ignorant about the outcome (note that *no* cheating is involved).

You toss the coin and see that it lands heads. Then the choice of bets is offered to HENRY and TINA. HENRY chooses the heads-bet and TINA the tails-bet.

Now consider this: *you know* the "truth", you know what the successful decision would be. It turns out that HENRY made the choice corresponding to the truth. TINA didn't. Would you then evaluate the HENRY algorithm to be better than the TINA algorithm?

For exactly the same reasons already discussed, the TINA algorithm is the better one; it made the optimal decision. Yet

it didn't choose the "truth". You realize that *comparing algorithms is not as simple as checking which one yields the truth more often.*

We have then arrived at two conclusions:

- 👉 "Success" or "correspondence to truth" is generally **not** a good criterion to judge a decision under uncertainty or to evaluate an algorithm that makes such decisions.
- 👉 Even if there is no method to determine which decision is successful, there is nevertheless a method to determine which decision is **rational** or **optimal**, given the particular gains, losses, and uncertainties involved in the decision problem.

We had a glimpse of this method in our introductory scenarios with electronic components and their variations.

Let us emphasize, however, that we are not giving up on "success"; nor are we trading "success" for "optimality". We'll find out that *Decision Theory automatically leads to the* ***successful*** *decision* in problems where uncertainty is not present or is irrelevant. It's a win-win. Keep this point firmly in mind:

> Aiming to find the solution that is *successful* can make us *fail* to find the solution that is optimal, when the successful one cannot be determined.
> Aiming to find the solution that is *optimal* makes us automatically also find the solution that is *successful*, when this can be determined.

We shall later witness this fact with our own eyes. We will also take it up in the discussion of some misleading techniques to evaluate machine-learning algorithms.

2.4 Decision Theory

So far we have mentioned that Decision Theory has the following features:

- ✔ It tells us what's optimal and, when possible, what's successful.
- ✔ It takes into consideration decisions, consequences, costs and gains.
- ✔ It is able to deal with uncertainties.

What other kinds of features should we demand from it, in order to be applied to as many kinds of decision problems as possible, and to be relevant for data science? Here are two:

- If we find an optimal decision in regards to some problem, it may still happen that this decision leads to new, subsequent decision problems. For example, in the assembly-line scenario the decision `discard` could be carried out by burning, recycling, and so on. And each of these actions could have uncertain results and costs or gains. We thus face a decision within a decision. In general, a decision problem may involve several decision sub-problems, in turn involving decision sub-sub-problems, and so on.
- In data science, a common engineering goal is to design and build an automated or AI-based device capable of making an optimal decision, at least in specific kinds of uncertain situations. Think for instance of an aeronautic engineer designing an autopilot system; or a software company designing an image classifier.

Well, Decision Theory turns out to meet these two demands too, thanks to the following features:

- ✔ It is susceptible to recursive, sequential, and modular application.
- ✔ It can be used not only for human decision-makers, but also for AI or automated devices.

Decision Theory has a long history, going back to Leibniz in the 1600s and partly even to Aristotle in the −300s. It appeared in its present form around 1920–1960. What's remarkable about it is that it is not only **a** framework: it is **the** framework we must use. A logico-mathematical theorem shows that *any framework that does not break basic optimality and rationality criteria has to be equivalent to Decision Theory*. In other words, an "alternative" framework might use different terminology and apparently different mathematical operations, but it would boil down to the same notions and mathematical operations of Decision Theory. So if you wanted to invent and use another framework, then either (a) your framework would lead to some irrational or illogical consequences; or (b) your framework would lead to results identical to Decision Theory. Many frameworks that you are probably familiar with, such as optimization theory or Boolean logic, are just specific applications or particular cases of Decision Theory.

Thus we list one more important characteristic of Decision Theory:

- ✔ It is **normative**.

Normative contrasts with *descriptive*. The purpose of Decision Theory is not to describe, for example, how human decision-makers typically make decisions. Human decision-makers typically make irrational, sub-optimal, or biased decisions. That's exactly what we want to avoid! We want a theory, a *norm*, that human decision-makers should aspire to. That's what Decision Theory is.

For the extra curious

- *Judgment under uncertainty*
- *Heuristics and Biases*
- *Thinking, Fast and Slow*

Study reading

Who says that Decision Theory should be normative? – this is a respectable scientific question. If you found yourself wondering and doubting about this, then congratulations: that's how a scientist should think!
Later on we'll examine material and arguments about this point. If you like, feel free to already skim through the

following works, as a start in your investigations:

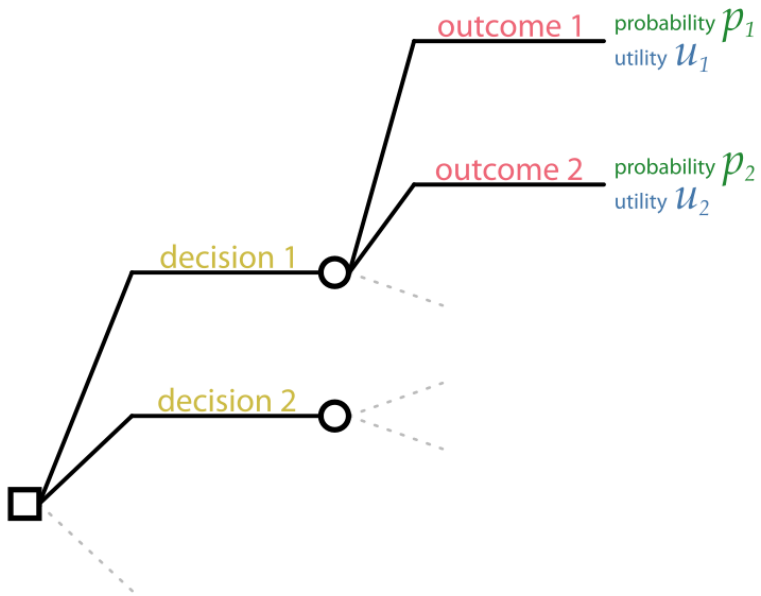
- Ch. 15, especially §15.1 and §"Bibliographical and Historical Notes" of *Artificial Intelligence*
- *Normative Theories of Rational Choice: Expected Utility*
- *Decision Theory*
- *Decision Analysis*

3 Basic decision problems

Decision Theory analyses any decision-making problem in terms of nested or sequential *basic* or *minimal* decision problems. The assembly-line scenario of the introduction 1 is an example.

3.1 Graphical representation and elements

A basic decision problem can be represented by a diagram like this:



It has one **decision node**, usually represented by a square ■, from which the available decisions depart as lines. Each decision leads to an **inference node**,[1] usually represented by

[1] also called *chance node* or *uncertainty node*

a circle ●, from which the possible outcomes depart as lines. Each outcome leads to a particular gain or loss, depending on the decision. The uncertainty of each outcome is quantified by a probability.

A basic decision problem is analysed in terms of the following elements:

- **Agent**, and **background** or **prior information**. The agent is the person or device that has to make the decision. An agent possesses (or has been programmed with) specific background information that is used and taken for granted in the decision-making process. This background information determines the probabilities, gains, and losses of the outcomes, together with other available data and information. Different agents typically have different background information.

> Remember: What matters is to be able to identify these elements in a concrete problem, understanding their role. Their technical names don't matter.

- **Data** and other **additional information**, sometimes called **evidence**. They differ from the background information in that they can change with every decision instance made by the same agent, while the background information stays the same. In the assembly-line scenario, for example, the test results could be different for every new electric component.

Agent means "conductor", "mover", and similar (from Latin *ago* = *to move* or *drive* and similar meanings).

We'll use the neutral pronouns *it*/*its* when referring to an agent, since an agent could be a person or a machine.

- **Decisions** available to the agent. They are assumed to be mutually exclusive and exhaustive; this can always be achieved by recombining them if necessary, as we'll discuss later.

- **Outcomes** of the possible decisions. Every decision can have a different set of outcomes, or some outcomes can appear for several or all decisions (in this case they are reported multiple times in the decision diagram). Note that even if an outcome can happen for two or more different decisions, its probabilities can still be different depending on the decision.

Decisions are called *courses of action* in some literature.

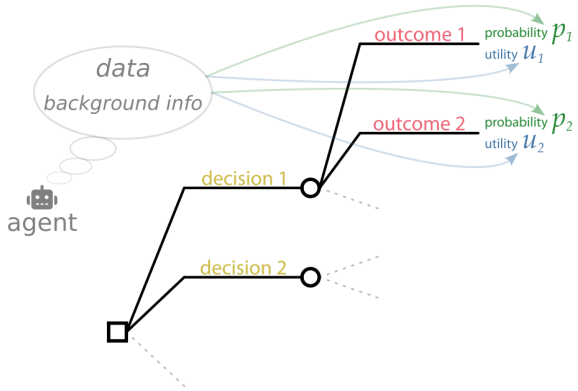
Many other terms instead of *outcome* are used in the literature, for instance *state* or *event*.

- **Probabilities** for each of the outcomes and for each decision. Their values typically depend also on the background information and the additional data.
- **Utilities**: the gains or losses associated with each possible outcome and each decision. We shall mainly use the term **utility**, instead of "gain", "loss", and similar, for several reasons:
  - gain and losses may involve not money, but time, or energy, or health, or emotional value, or other kinds of commodities and things that are important to us; or even a combination of them. The term "utility" is useful as a neutral term that doesn't mean "money", but depends on the context
  - we can just use one term instead of two: for example, when the utility is positive it's a "gain"; when it's negative it's a "loss"

  The particular numerical values of the utilities are always context-dependent: they may depend on the background information, the decisions, the outcomes, and the additional data.

  We shall sometimes use the generic currency sign ¤ to denote utilities, to make clear that gains and losses do not necessarily involve money, and not to reference any country in particular.

The relation between the elements above can be depicted as follows – but note that this is just an intuitive illustration:



⚠ Don't over-interpret the decision diagram

- The diagram above *doesn't have any temporal meaning*, that is, it doesn't mean that the decisions happen before the outcomes, or vice versa.

  In some situations the outcome can be realized after the decision is made; for instance, someone bets on heads or tails, and then a coin is tossed.

  In other situations, the outcome can be realized before the decision is made; for instance, sometimes a coin is tossed and covered, then one is asked to bet on what the outcome was. Another example is some research decision made by a archaeologist, the unknown being some detail about a dinosaur from millions of years ago.

  In yet other situations the outcome may have a complex nature, and it may be realized partly before the decision is made, and partly after; for instance, someone can bet on the outcome of two coin tosses; one coin is tossed before the decision is made, and the other after.

- The diagram above is not something that an agent *must* use in making decisions. It is not part of the theory. It's just a very convenient way to visualize and operate with the mathematics underlying the theory.

- It not always the case that the *outcomes* are unknown and the *data* are known. As we'll discuss later, in some situations we reason in hypothetical or counterfactual ways, using hypothetical data and considering outcomes which have already occurred. In such situations we can still use diagrams like the one above, because the help us doing the calculation, although the actual outcome is already known.

Study reading

- §1.1.4 in *Artificial Intelligence*
- *Skim through* Ch. 15 of *Artificial Intelligence*. No need to read thoroughly: just quickly glimpse whether there are ideas and notions that look familiar (a little like when you're in a large crowd and look quickly around to see if there are any familiar faces)

Exercise

- Identify the elements above in the assembly-line decision problem of the introduction 1.
- Sketch the decision diagram for the assembly-line decision problem.

Some of the decision-problem elements listed above may need to be in turn analysed by a decision sub-problem. For instance, the utilities could depend on uncertain factors: thus we have a decision sub-problem to determine the optimal values to be used for the utilities of the main problem. This is an example of the modular character of decision theory.

We shall soon see how to mathematically represent these elements.

The elements above must be identified unambiguously in every decision problem. The analysis into these elements greatly helps in making the problem and its solution well-defined.

An advantage of decision theory is that its application *forces* us

to make sense of an engineering problem. A useful procedure is to formulate the general problem in terms of the elements above, identifying them clearly. If the definition of any of the terms involves uncertainty of further decisions, then we analyse it in turn as a decision sub-problem, and so on.

> Suppose someone (probably a politician) says: "We must solve the energy crisis by reducing energy consumption or producing more energy". From a decision-making point of view, this person has effectively said *nothing whatsoever*. By definition the "energy crisis" is the problem that energy production doesn't meet demand. So this person has only said "we would like the problem to be solved", without specifying any solution. A decision-theory approach to this problem requires us to specify which concrete courses of action should be taken for reducing consumption or increasing productions, and what their probable outcomes, costs, and gains would be.

3.2 Setting up a basic decision problem

A basic decision problem can be set up along the following steps (which we illustrate afterwards with a couple of examples):

Setup of a basic decision problem

1. **List all available decisions**
2. **For each decision, list its possible outcomes**
3. **Pool together all outcomes of all decisions, counting the common ones only once**
4. **Prepare two tables: in each, display the decisions as rows, and the pooled outcomes as columns** (or you can do the opposite: decisions as columns and outcomes as rows)

For the extra curious

See MacKay's options-vs-costs rational analysis in Sustainable Energy – without the hot air

5. **In one table, report the probabilities for all decision-outcome pairs. If an outcome is not available for that decision, give it a 0% probability**

6. **In the other table, report the utilities for all decision-outcome pairs. If an outcome is not available for that decision, give it a 0 utility**

Example: the assembly-line problem

Let's apply the steps above in the assembly-line example of ch. 1:

1. List all available decisions

Easy: they are "accept the electronic component" and "discard it".

2. For each decision, list its possible outcomes

In general you will notice that some outcomes may be common to all decisions, while other outcomes can happen for some decisions only, or even for just one decision.

In the present example, the accept decision has two possible outcomes: "the component works with no faults for at least a year" and "the component fails within a year".

The discard cannot have those outcomes, because the component is discarded. It has indeed only one outcome: "component discarded".

3. Pool together all outcomes of all decisions, counting the common ones only once

In total we have *three* pooled outcomes:

- **no faults** (from the accept decision)

- **fails** (from the accept decision)
- **discarded** (from the discard decision)

4. Prepare two tables: in each, display the decisions as rows, and the pooled outcomes as columns (or you can do the opposite: decisions as columns and outcomes as rows)

In the present example each table looks like this:

	no faults for a year	fails within a year	discarded
accept			
discard			

5. In one table, report the probabilities for all decision-outcome pairs. If an outcome is not available for that decision, give it a 0% **probability**

Table 3.2: *Probability table*

	no faults for a year	fails within a year	discarded
accept	90%	10%	0%
discard	0%	0%	100%

Note how the outcomes that do not exist for a particular decision have been given a 0% probability (in grey). This is just a way of saying "this outcome can't happen, if this decision is made".

6. In the other table, report the utilities for all decision-outcome pairs. If an outcome is not available for that decision, give it a 0 **utility**

Table 3.3: *Utility table*

	no faults for a year	fails within a year	discarded
accept	+1$	−11$	0$
discard	0$	0$	0$

Note how the outcomes that do not exist for a particular decision have been given a 0$ utility (in grey). We shall see later that it actually doesn't matter which utilities we give to these impossible outcomes.

Exercises

Apply the steps above to the following basic decision problems (you only need to set them up with their probability & utility tables, but feel free to solve them as well, if you like):

- The "heads-bet" vs "tails-bet" example of § 2.3. Assume that the "small amount" of money is 10$, the "large amount" is 1000$, and the two outcomes' probabilities are 50% each.
- Peter must reach a particular destination, and is undecided between two alternatives: *go by car*, or *ride a bus*, or *go on foot*. If he goes by car, he could *arrive without problems*, with a probability of 80% and a utility of 10 ¤, or he could *get stuck in a traffic jam* and arrive late, with a probability of 20% and a utility of −10 ¤. If he rides a bus, he could *arrive without problems*, with a probability of 95% and a utility of 15 ¤, or arrive in time but *travelling in a fully-packed bus*, with a probability of 5% and a utility of −10 ¤. If he goes on foot, he could *arrive without problems*, with a probability of 20% and a utility of 20 ¤, or he could *get soaked from rain*, with a probability of 80% and a utility of −5 ¤.

(We are using the symbol "¤" because Peter's utilities are a combination of money savings, time of arrival, and comfort.)

3.3 How to make a basic decision?

Up to now we have seen what are the elements of a basic decision problem, and how to arrange them in a diagram and with tables. *But how do we determine what's the optimal decision?*

Decision Theory says that the optimal decision is determined by the "**principle of maximal expected utility**".

We shall study this principle more in detail toward the end of the course, although you already know its basic idea, because you intuitively used this very principle in solving all decision problems we met so far, starting from the assembly-line one.

However, let's quickly describe already now the basic procedure for this principle:

Principle of maximal expected utility

1. For each decision, multiply the probability and the utility of each of its outcomes, and then sum up these products. This way you obtain the *expected utility* of the decision.
2. Choose the decision that has the largest expected utility; if several decisions are maximal, choose any of them unsystematically.

This procedure can also be described in terms of the probability and utility tables introduced in the previous section:

a. Multiply element-by-element the probability table and the utility table, obtaining a new table with the same number of rows and columns

b. Sum up the elements of each row of the new table (this sum is the expected utility); remember that every row corresponds to a decision
c. Choose the decision corresponding to the largest of the sums above; if there are several maximal ones, choose among them unsystematically

Example: the assembly-line problem

Multiplying the *Probability table* and the *Utility table* above, element-by-element, we obtain the following table, where we also indicate the sum of each row:

Table 3.4: *Probability* × *Utility table*

	no faults for a year	fails within a year	discarded	*sum*
accept	+0.9$	−1.1$	0$	−0.2$
discard	0$	0$	0$	0$

and, as we already knew, discarding the electronic component is the decision with the maximal expected utility.

Exercise

Feel free to sketch some code (in your preferred programming language) that chooses the optimal decision according to the principle above. The code should take two inputs: the table or matrix of probabilities, and the table or matrix of utilities; and should give one output: the row-number of the optimal decision.

3.4 Plan for the next chapters

The **expected-utility maximization** above is intuitive and simple, and is the last stage in a basic decision problem.

But there are two stages which occur before, and which are the most difficult:

- **Inference** is the stage where the probabilities of the possible outcomes are calculated. Its rules are given by the **Probability Calculus**. Inference is independent from decision: in some situations we may simply wish to assess whether some hypotheses, conjectures, or outcomes are more or less plausible than others, without making any decision. This kind of assessment can be very important in problems of communication and storage, and it is specially considered by **Information Theory**.

The calculation of probabilities can be the part that demands most thinking, time, and computational resources in a decision problem. It is also the part that typically makes most use of data – and where data can be most easily misused.

Roughly half of this course will be devoted in understanding the laws of inference, their applications, uses, and misuses.

- **Utility assesment** is the stage where the gains or losses of the possible outcomes are calculated. Often this stage requires further inferences and further decision-making sub-problems. The theory underlying utility assessment is still much underdeveloped, compared to probability theory.

We shall now explore each of these two stages. We take up inference first because it is the most demanding and probably the one that can be optimized the most by new technologies.

4 Connection with machine learning and AI

4.1 Inferences with machine-learning algorithms

Some works in machine learning focus on "guessing the correct answer", and this focus is reflected in the way their machine-learning algorithms – especially classifiers – are trained and used.

In § 2.3 we emphasized that "guessing successfully" can be a misleading goal, however, because it can lead us away from guessing *optimally*. We shall now see two simple but concrete examples of this.

4.1.1 A "max-success" classifier vs an optimal classifier

> You find the code for this chapter and exercises also in this JupyterLab notebook for R and (courtesy of Viktor Karl Gravdal!) this JupyterLab notebook for python.

We shall compare the results obtained in some numerical simulations by using

- a Machine-Learning Classifier trained to do most successful guesses
- a prototype "Optimal Predictor Machine" trained to make the optimal decision

For the moment we treat both as "black boxes", that is, we don't study yet how they're calculating their outputs (although you may already have a good guess at how the Optimal Predictor Machine works).

Their operation is implemented in this R script that we now load:

```
source('code/mlc_vs_opm.R')
```

This script simply defines the function `hitsvsgain()`:

```
hitsvsgain(
    ntrials,
    chooseAtrueA,
    chooseAtrueB,
    chooseBtrueB,
    chooseBtrueA,
    probsA
)
```

having six arguments:

- `ntrials`: how many simulations of guesses to make
- `chooseAtrueA`: utility gained by guessing `A` when the successful guess is indeed `A`
- `chooseAtrueB`: utility gained by guessing `A` when the successful guess is `B` instead
- `chooseBtrueB`: utility gained by guessing `B` when the successful guess is indeed `B`
- `chooseBtrueA`: utility gained by guessing `B` when the successful guess is `A` instead
- `probsA`: a tuple of probabilities (between `0` and `1`) to be used in the simulations (recycling it if necessary), for the successful guess being `A`; the corresponding probabilities for `B` are therefore `1-probsA`. If this argument is omitted it defaults to `0.5` (not very interesting)

4.1.2 Example 1: electronic component

Let's apply our two classifiers to the *Accept or discard?* problem of § 1. We call `A` the alternative in which the element won't fail before one year, and should therefore be accepted *if this alternative were known at the time of the decision.* We call `B` the alternative in which the element will fail within a year, and should therefore be discarded *if this alternative were known at the time of the decision.* Remember that the crucial point here is that the classifiers *don't* have this information at the moment of making the decision.

We simulate this decision for 100 000 components ("trials"), assuming that the probabilities of failure can be `0.05`, `0.20`, `0.80`, `0.95`. The values of the arguments should be clear:

```
hitsvsgain(
    ntrials = 100000,
    chooseAtrueA = +1,
    chooseAtrueB = -11,
    chooseBtrueB = 0,
    chooseBtrueA = 0,
    probsA = c(0.05, 0.20, 0.80, 0.95)
)
```

```
Trials: 100000
Machine-Learning Classifier: successes 87549 ( 87.5 %) | total gain -24681
Optimal Predictor Machine:   successes 72671 ( 72.7 %) | total gain 10571
```

Note how the machine-learning classifier is the one that *makes most successful guesses* (around 88%), **and yet it leads to a net loss!** If the utility were in *kroner*, this classifier would cause the company producing the components a net loss of more than 20 000 kr.

The optimal predictor machine, on the other hand, *makes fewer successful guesses* overall (around 72%), **and yet it leads to a net gain!** It would earn the company a net gain of around 10 000 kr.

Exercise

How is this possible? Try to understand what's happening; feel free to research this by modifying the `hitsvsgain()` function, so that it prints additional outputs.

4.1.3 Example 2: find Aladdin! (image recognition)

A typical use of machine-learning classifiers is for image recognition: for instance, the classifier guesses whether a particular subject is present in the image or not.

Intuitively one may think that "guessing successfully" should be the best goal here. But exceptions to this may be more common than one thinks. Consider the following scenario:

> Bianca has a computer folder with 10 000 photos. Some of these include her beloved cat Aladdin, who sadly passed away recently. She would like to select all photos that include Aladdin and save them in a separate "Aladdin" folder. Doing this by hand would take too long, if at all possible; so Bianca wants to employ a machine-learning classifier.
>
> For Bianca it's important that no photo with Aladdin goes missing, so she would be very sad if any photo with him weren't correctly recognized; on the other hand she doesn't mind if some photos without him end up in the "Aladdin" folder – she can delete them herself afterwards.

Let's apply and compare our two classifiers to this image-recognition problem, using again the `hitsvsgain()` function. We call `A` the case where Aladdin is present in a photo, and `B` where he isn't. To reflect Bianca's preferences, let's use these "emotional utilities":

- `chooseAisA = +2`: Aladdin is correctly recognized
- `chooseBisA = -2`: Aladdin is not recognized and photo goes missing

- `chooseBisB = +1`: absence of Aladding is correctly recognized
- `chooseAisB = -1`: photo without Aladding end up in "Aladding" folder

and let's say that the photos may have probabilities `0.3`, `0.4`, `0.6`, `0.7` of including Aladding:

```
hitsvsgain(ntrials = 10000, chooseAtrueA = +2, chooseAtrueB = -1, chooseBtrueB = 1, chooseBtru
```

```
Trials: 10000
Machine-Learning Classifier: successes 6517 ( 65.2 %) | total gain 4525
Optimal Predictor Machine:   successes 6000 ( 60 %) | total gain 5537
```

Again we see that the machine-learning classifier makes more successful guesses than the optimal predictor machine, but the latter yields a higher "emotional utility".

You may sensibly object that this result could depend on the peculiar utilities or probabilities chosen for this example. The next exercise helps answering your objection.

Exercise

- Is there any case in which the optimal predictor machine yields a strictly lower utility than the machine-learning classifier?
  - Try using different utilities, for instance using `±5` instead of `±2`, or whatever other values you please.
  - Try using different probabilities as well.
- As in the previous exercise, try to understand what's happening. Consider this question: *how many photos including Aladdin did each classifier miss?*

  Modify the `hitsvsgain()` function to output this result.
- Do the comparison using the following utilities: `chooseAtrueA = +1`, `chooseAtrueB = -1`,

`chooseBtrueB = 1`, `chooseBtrueA = -1`. What's the result? what does this tell you about the relationship between the machine-learning classifier and the optimal predictor machine?

4.2 What is "Artificial Intelligence"?

4.2.1 "AI" as opposed to what?

The field of Artificial Intelligence is vast, and its boundaries are not clear-cut. Different books give slightly different definitions of AI. In everyday parlance the term "AI" is moreover used in ways that are *not* technically correct – a bit like it happens with physics terms such as "energy" or "force". In this course we want to use *AI* in a technically more correct way.

The discussion of the possible definitions of AI could take several chapters. Let's try a shorter approach, by examining why the two words "artificial" and "intelligence" are used specifically.

Artificial as opposed to what? As opposed to *natural* for example. So it denotes something human-made, as opposed to something directly found in nature; say in an orangutan or in a dolphin.

Intelligence as opposed to what? As opposed to *stupidity*. The definition of "intelligence" itself, even natural intelligence, is still quite open. Generally we mean something that is *logical* or *rational*. Thus an agent that breaks some logical procedure, or that does not follow a procedure that it claims to follow, is not "intelligent".

Of course neither term is fully dichotomous: we can distinguish different degrees of artificiality and of intelligence.

4.2.2 "Intelligence" is not "human-likeness"

We can distinguish two distinct endeavours in the field of Artificial Intelligence, considered in its most general extension:

- achieving *human-like* behaviour;
- achieving *intelligent reasoning*, or we could say *logical* or *rational reasoning*.

It's important to recognize immediately that these two endeavours may *not be mutually compatible*. We often associate human behaviour with error-making and irrationality. We may say that a person is very irrational, yet we don't say that because of this the person is inhuman.

Given the incompatible character of the two endeavours above, we must be very clear and conscious about which goal we're trying to achieve; otherwise we won't achieve any goal at all. And in technical discussions we must be careful to adopt the correct terminology. In particular we should avoid the term "intelligent" when we instead mean "human-like", and vice versa.

An example of such confusion is with present-day *large language models* (LLMs), and in particular those with a Generative Pretraining Transformer (GPT) architecture. In many media they are referred to as "AI systems"; yet what they achieve is not *intelligence*, but rather *human-like* language processing – including non-intelligent processing.

If you have access to a large language model, you have surely witnessed examples of stupid output[1]. You can try a variation of the following experiment:

1. Ask the LLM to write down a short list of some set, for instance of all Norwegian counties.
2. Ask the LLM to select from the list only those item that have one or more letter "r" in their name. See the result.
3. Ask the LLM to give you a step-by-step procedure to achieve the selection required in the previous step.

[1] often euphemistically called "hallucination" because this term may increase sales, whereas "stupid" would risk decreasing sales.

Typically a LLM fails at task 2., even if it can give a completely sound procedure in task 3. Clearly it isn't internally following the logical procedure.

This is the reason why in this course we do *not* categorize LLMs as "artificial intelligence", but rather as human-mimicking machines. But we shall consider possible ways in which a true intelligence framework could be built into these machines.

Study reading

- Chapters 1–2 of *Artificial Intelligence*.

Part II

Inference I

5 What is an inference?

In the assembly-line decision problem of § 1, the probability of early failure was very important in determining the optimal decision. If the probability had been 5% instead of 10%, the optimal decision would have been different. Also, if the probability had been 100% or 0%, it would have meant that we knew *for sure* what was the successful decision.

In that decision problem, the probabilities of the outcomes were already given. But in real decision problems the probabilities of the outcomes almost always need to be calculated, and their calculation can be the most time- and resource-demanding stage in solving a decision problem.

We'll loosely refer to problems of calculating probabilities as "*inference* problems", and to their calculation as "drawing an inference". Drawing inferences is very often a goal or need in itself, without any underlying decision process.

Our purpose now is to learn how to draw inferences – that is, how to calculate probabilities. We'll proceed by facing the following questions, in order:

- What do we mean by "inference" and "probability", more precisely? What important aspects about inferences and probabilities should we keep in mind?
- What kind of mathematical notation do we use for inferences and probabilities?
- What are the rules for drawing inferences, that is, for calculating probabilities?

5.1 The wide scope and characteristics of inferences

Let's see a couple more informal examples of inference problems. For some of them an underlying decision-making problem is also alluded to:

A. Looking at the weather, we try to assess if it'll rain today, to decide whether to take an umbrella.

B. Considering a patient's symptoms, test results, and medical history, a clinician tries to assess which disease affects the patient, in order to decide on the optimal treatment.

C. Looking at the present game position  the X-player, which moves next, wonders whether placing the next **X** on the mid-right position leads to a win.

D. The computer of a self-driving car needs to assess, from the current set of camera frames, whether a particular patch of colours in the frames is a person, in order to slow down the car and stop if that's the case.

E. Given that $G = 6.67{\cdot}10^{-11}\,\mathrm{m^3\,s^{-2}\,kg^{-1}}$, $M = 5.97{\cdot}10^{24}\,\mathrm{kg}$ (mass of the Earth), and $r = 6.37 \cdot 10^6\,\mathrm{m}$ (radius of the Earth), a rocket engineer needs to know how much is $\sqrt{2\,G\,M/r}$.

F. We'd like to know whether the rolled die is going to show ⚅.

G. An aircraft's autopilot system needs to assess how much the aircraft's roll will change, if the right wing's angle of attack is increased by 0.1 rad.

H. By looking at the dimensions, shape, texture of a newly dug-out fossil bone, an archaeologist wonders whether it belonged to a Tyrannosaurus rex.

I. A voltage test on a newly produced electronic component yields a value of 100 mV. The electronic component turns out to be defective. An engineer wants to assess whether

the voltage-test value could have been 100 mV even if the component had *not* been defective.

J. Same as above, but the engineer wants to assess whether the voltage-test value could have been 80 mV if the component had not been defective.

K. From measurements of the Sun's energy output, measurements of concentrations of various substances in the Earth's atmosphere over the past 500 000 years, and measurements of the emission rates of various substances in the years 1900–2022, climatologists and geophysicists try to assess the rate of mean-temperature increase in the years 2023–2100.

For the extra curious

Ch. 10 in *A Survival Guide to the Misinformation Age*.

Exercises

5. For each example above, pinpoint what has to be inferred, and also the *agent* interested in the inference.
6. Point out which of the examples above *explicitly* give data or information that should be used for the inference.
7. For the examples that do not give explicit data or information, speculate what information could be implicitly assumed. For those that do give explicit data, speculate which other additional information could be implicitly assumed.
8. Can any of the inferences above be done with full certainty (that is, to know which decision is successful), based the data given explicitly and implicitly?
9. Find the examples that explicitly involve a decision. In which of them does the decision affect the results of the inference? In which it does not?

10. Are any of the inferences "*one-time only*"? That is, has their object or the data on which they are based never happened before and will never happen again?

11. Are any of the inferences above based on data and information that come chronologically *after* the object of the inference?

12. Are any of the inferences above about something that is actually already known to the agent that's making the inference?

13. Are any of the inferences about something that actually did not happen?

14. Do any of the inferences use "data" or "information" that are actually known (within the scenario itself) to be fictive, that is, *not* real?

From the examples and from your answers to the exercise we observe some very important characteristics of inferences:

- Some inferences can be made exactly, that is, *without uncertainty*: it is possible to say for sure whether the object of the inference is true or false. Other inferences, instead, involve an uncertainty.
- *All inferences are based on some data and information*, which may be explicitly expressed or only implicitly understood.
- An inference can be about something *past*, but based on *present or future* data and information. In other words, inferences can show *all sorts of temporal relations*.
- An inference can be *essentially unrepeatable*, because it's about something unrepeatable or based on unrepeatable data and information.
- The data and information on which an inference is based can actually be unknown; that is, they can be only momentarily contemplated as real. Such an inference is said to be based on **hypothetical reasoning**.

- The object of an inference can actually be something already known to be false or not real: the inference tries to assess it in the case that some data or information had been different. Such an inference is said to be based on **counterfactual reasoning**.

5.2 Where are inferences drawn from?

This question is far from trivial. In fact it has connections with the earth-shaking development and theorems in the foundations of mathematics that originated in the 1900s.

The proper answer to this question will take up the next sections. But a central point can be emphasized now:

> **Inferences can only be drawn from other inferences.**

In order to draw an inference – calculate a probability – we usually go up a chain: we must first draw other inferences, and for drawing those we must draw yet other inferences, and so on.

At some point we must stop at *inferences that we take for granted without further proof*. These typically concern direct experiences and observations. For instance, you see a tree in front of you, so you can take "there's a tree here" as a true fact. Yet, notice that the situation is not so clear-cut: how do you know that you aren't hallucinating, and there's actually no tree there? That is taken for granted. If you analyse the possibility of hallucination, you realize that you are taking other things for granted, and so on.

Probably most philosophical research in the history of humanity has been about grappling with this runaway process – which is also a continuous source of sci-fi films. In logic and mathematical logic, this corresponds to the fact that in order to prove some *theorem*, we must always start from some *axioms*. There are "inferences", called *tautologies*, that can be drawn without

For the extra curious

Mathematics: The Loss of Certainty.

requiring others, but they are all trivial: for example "this component failed early, or it didn't". These tautologies are of little use in a real problem, although they have a deep theoretical importance. Useful inferences, on the other hand, must always start from some axioms.

In concrete applications, we start from many inferences upon which everyone, luckily, agrees. But sometimes we must also use starting inferences that are more dubious or not agreed upon by everyone. In this case the final inference has a somewhat contingent character. We accept it (as well as the solution of any underlying decision problem) as the best available one for the moment. This is partly the origin of the term "**model**".

Figure 5.1: Sci-fi films like *The Matrix* ultimat
Proposal is Johnson's (1924) terminology; Keynes (1921) uses "conclusion"; modern textbooks do not seem to use any specialized term. *Conditional* is modern terminology; other terms used: "evidence", "premise", "supposal". The *vertical bar*, originally a solidus, was introduced by Keynes (1921).
Fact (axiom) must take some infere
without further proof.

5.3 Basic elements of an inference

Let us introduce some mathematical notation and more precise terminology for inferences.

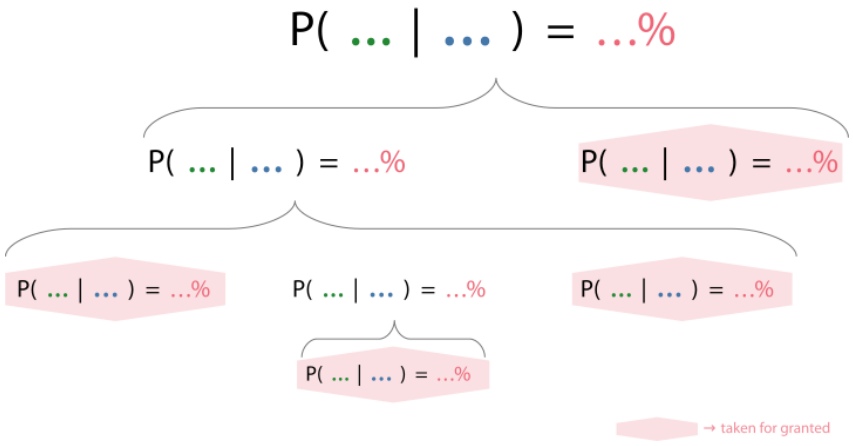
- Every inference has an "object": what is to be assessed or guessed. We call **proposal** the object of the inference.
- Every inference also has data, information, hypotheses, or hypothetical scenarios on which it is based. We call **conditional** what the inference is based upon.
- We separate *proposal* and *conditional* with a vertical bar " | ", which can be pronounced "**given**" or "**conditional on**".
- Finally, we put parentheses around this and a "P" in front, short for "probability":

$$\mathrm{P}(\text{␣}1px]\cdots_{proposal} \mid \text{␣}1px]\cdots_{conditional}) = \cdots\%$$

this means "the probability that *proposal*, supposing *conditional*, is . . . %". Or also: "supposing *conditional*, we can infer *proposal* with . . . % probability".

We have remarked that in order to calculate the probability for an inference, we must use the probabilities of other inferences,

which in turn are calculated by using the probabilities of other inferences, and so on, until we arrive at probabilities that are taken for granted. A basic inference process could therefore be schematized like this:



The next important task ahead of us is to introduce a flexible and enough general mathematical representation for the objects of an inference. Thereafter we shall study the rules for drawing correct inferences.

6 Sentences

We have seen that an inference involves, at the very least, two things: the object of the inference (*proposal*), and the data, information, or hypotheses on which the inference is based (*conditional*).

We also observed that wildly different "things" can be the object of an inference or the information on which the inference is based: measurement results, decision outcomes, hypotheses, not-real events, assumptions, data and information of all kinds (for example, images). In fact, such variety in some cases can make it difficult to pinpoint what an inference is about or what it is based upon.

Is there a general, flexible, yet precise way of representing all these kinds of "things"?

6.1 The central components of knowledge representation

When speaking of "data", what comes to mind to many people is numbers or collections of numbers. Maybe numbers, then, could be used to represent all the variety of "things" exemplified above? Well, this option turns out to be too restrictive.

I give you this number: "8", saying that it is "data". But what is it about? You, as an agent, can hardly call this number a piece of information, because you have no clue what to do with it.

Instead, if I tell you: "The number of official planets in the solar system is 8", then we can say that I've given you data. You can do different things with this piece of information. For instance, if you had decided to send one probe to each official planet,

now you know you have to build eight probes. Or maybe you can win at a pub quiz with it.

"Data" is therefore not just numbers. A number is not "data" unless there's an additional verbal and non-numeric context accompanying it – even if only implicitly. Sure, we could represent this meta-data information as numbers too; but this move would only shift the problem one level up: we would need an auxiliary verbal context explaining what the meta-data numbers are about.

Data can, moreover, be completely non-numeric. A clinician saying "The patient has fully recovered from the disease" is giving us a piece of information that we could further use, for instance, to make prognoses about other, similar patients. The clinician's statement surely is "data", but is essentially non-numeric data. Sure, in some situations we could represent this data with numbers, say "1" for "recovered" and "0" for "not recovered". But the opposite or some other convention could also be used: "0" for "recovered" and "1" for "not recovered", or the numbers "0.3" and "174". These numbers have intrinsically nothing to do with the clinician's "recovery" data.

The examples above, however, actually reveal the answer to our needs! In the examples we expressed the data by means of *sentences*. Clearly any measurement result, decision outcome, hypothesis, not-real event, assumption, data, and any piece of information can be expressed by a sentence.

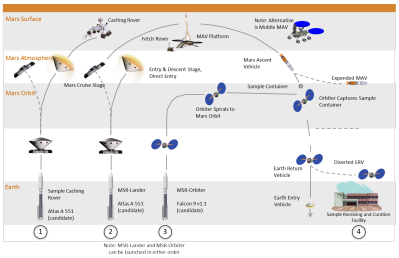
We shall therefore use **sentences**, also called **propositions** or **statements**,[1] to represent and communicate all the kinds of "things" that can be the proposal or the conditional of an inference. In some cases we can of course summarize a sentence by a number, as a shorthand, when the full meaning of the sentence is understood.

[1] These three terms are not always equivalent in formal logic, but here we'll use them as synonyms.

Sentences are the central components of knowledge representation in AI agents. For example they appear at the heart of automated control programs and fault-management systems in NASA spacecrafts.

Study reading

- §7.1 in *Artificial Intelligence*.
- Take a *quick look* at these:
  - *SMART: A propositional logic-based trade analysis and risk assessment tool for a complex mission*
  - around p. 22 in *No More Band-Aids: Integrating FM into the Onboard Execution Architecture*
  - part IV in *Model-based programming of intelligent embedded systems and robotic space explorers*



(From the *SMART* paper)

6.2 Identifying and working with sentences

But what is a sentence, more exactly? The everyday meaning of this word will work for us, even though there are more precise definitions – and still a lot of research in logic an artificial intelligence on how to define and use sentences. We shall adopt this useful definition:

*A **sentence** is a verbal message for which an agent can determine, at least in principle, whether it is `true` or `false`.*

Let's make this definition clearer with some remarks:

- 👉 **A sentence doesn't have to contain only words**. It can contain pictures, sounds, and other non-verbal items. For example, the following:

  "This:  is an animated picture of Saitama."

For the extra curious

Propositions

is a sentence, even if it contains animated graphics, because we can say that it is `true`. Likewise, the following:

"This link leads to a song by Pink Floyd."

is also a sentence, even if it contains links and audio, because we can say that it is `false` (that's a song by Monty Python).

- 👉 **A meaningful phrase may not be a sentence**. For instance, a phrase like "Apples are much tastier than pears" may not be a sentence, because it's a matter of personal taste whether it's `true` or `false`. Moreover, an agent's opinion about apples and pears might change from time to time.

  The phrase "Jenny right now finds apples tastier than pears", on the other hand, could be a sentence; its truth being found by asking Jenny at that very moment.

  In an engineering context, the phrase "This valve will operate for at least two months" is a sentence, even if its truth is unknown at the moment: one has to wait two months, and then its truth will be unambiguously known.

- 👉 **An expression involving technical terms may not be a sentence (and not meaningful either)**. For instance, in a data-science context the phrase "This neural-network algorithm has better performance than that random-forest one" is *not* a sentence unless we have objectively specified what "better" means (higher accuracy? higher true-positive rate? faster?), for example by adopting a particular comparison metric.

  Some expressions involving technical terms may appear to be sentences at first; but a deeper analysis then reveals that they are not. A famous example is the sentence "The two events (at different spatial locations) are simultaneous". Einstein showed that there's no physical way to determine whether such an expression is true or false. Its truth turns out to be a matter of convention. The Theory of Relativity was born from this observation.

🚀 For the extra curious

On the electrodynamics of moving bodies.

Important

❗ Be particularly careful when reading scientific and engineering papers with a lot of technical terms and phrases. Technical jargon often makes it especially difficult to see whether something true or at least meaningful is being said, or not!

- 👉 **A sentence can be expressed in different ways** by different phrases and in different languages. For instance, "The temperature is 248.15 K", "Temperaturen ligger på minus 25 grader", and "25 °C is the value of the temperature" all represent the *same* sentence.

There are many advantages in working with sentences (rather than just numbers), and in keeping in mind that every inference is about sentences:

First, this point of view leads to **clarity** in engineering problems, and makes them more **goal-oriented**. A data engineer must acquire information and convey information. "Acquiring information" does not simply consist in making measurements or counting something: the engineer must understand *what* is being measured and *why*. If data is gathered from third parties, the engineer must ask what exactly the data mean and how they were acquired. In designing a solution, it is important to understand what information or outcomes the end user exactly wants. These "what", "why", "how" are expressed by sentences. A data engineer will often ask "*wait, what do you mean by that?*". This question is not just an unofficial parenthesis in the official data-transfer workflow between the engineer and someone else. It is an integral part of that workflow: it means that some information has not been completely transferred yet.

Second, this point of view is extremely important in AI and machine-learning design. A (human) engineer may proceed informally when drawing inferences, without worrying about "sentences" unless a need for disambiguation arises. A data

engineer who's *designing* or *programming* an algorithm that will do inferences automatically, must instead be unambiguous and cover beforehand all possible cases that the algorithm will face.

We therefore agree that *the proposal and the conditional of an inference have to be sentences*. This means that the proposal of the inference must be something that can be true or false.

Many inferences, especially when they concern numerical measurements, involve more than one sentence. For example, an inference about the result of rolling a die actually consists of the probabilities for six separate proposals:

'The result of the roll is 1'

'The result of the roll is 2'

...

'The result of the roll is 6'

Later on we shall see how to work with more complex inferences of this kind. In real applications it can be useful, on some occasions, to pause and reduce an inference to its basic set of `true`/`false` sentences. This analysis may reveal contradictions in our inference problem. A simple way to do this is to reduce the complex inference into a set of yes/no questions.

This kind of analysis is also important in information-theoretic situations: the **information content** provided by an inference, when measured in *Shannons*, is related to the minimal amount of yes/no questions that the inference answers.

Exercise

Rewrite each inference scenario of § 5.1 in a formal way, as one or more inferences

$$[\textit{proposal}] \;\Big|\; [\textit{conditional}]$$

where proposal and conditional are well-defined sentences.

> In ambiguous cases, use your judgement and motivate your choices.

6.3 Notation and abbreviations

Writing full sentences would take up *a lot* of space. Even an expression such as "The speed is 10 m/s" is not a sentence, strictly speaking, because it leaves unspecified the speed of what, when it was measured and in which frame of reference, what we mean by "speed", how the unit "m/s" is defined, and so on.

Typically we leave the full content of a sentence to be understood from the context, and we denote the sentence by a simple expression. Example:

$$\mathsf{The\ speed\ is\ 10\,m/s}$$

or even more compactly introducing physical symbols:

$$v{=}10\,\mathrm{m/s}$$

where v is a physical variable denoting the speed. Sometimes we may simply write

$$10\,\mathrm{m/s}$$

In some problems it's useful to introduce symbols to denote sentences. As mentioned before, in these notes we'll use sans-serif italic letters: $\mathsf{A}, \mathsf{B}, \mathsf{a}, \mathsf{b}, ...$, possibly with sub- or superscripts. For instance, the sentence "The speed is 10 m/s" could be denoted by the symbol S_{10}. We express such a definition like this:

$$\mathsf{S}_{10} := \text{`}\mathsf{The\ speed\ is\ 10\,m/s}\text{'}$$

which means that the symbol S_{10} is defined to be the sentence 'The speed is 10 m/s'.

⚠ We must be wary of how much we shorten sentences

Consider these three sentences:

> 'The speed is measured to be 10 m/s'
>
> 'The speed is set to 10 m/s'
>
> 'The speed is reported, by a third party, to be 10 m/s'

The quantity "10 m/s" is the same in all three sentences, but their meanings are very different. They represent different kinds of data. The difference greatly affect any inference about or from these data. For instance, in the third case an engineer may not take the indirectly-reported speed "10 m/s" at face value, unlike in the first case. In a scenario where all three sentences can occur, it would be ambiguous to simply write "$v = 10\,\mathrm{m/s}$": would the equal-sign mean "measured", "set", or "indirectly reported"?

Exercise

How would you denote the three sentences above, to make their differences clear?

Get familiar with abbreviations of sentences

To summarize, a sentence like

'The temperature T has value x'

could be abbreviated in these different ways:

- A symbol for the sentence (note the sans-serif font):

$$\mathsf{S}$$

- Some key word appearing in the sentence:

temperature

- An equality:

$$T{=}x$$

- The quantity appearing in the sentence:

$$T$$

- The value appearing in the sentence:

$$x$$

Get familiar with these kinds of abbreviations because they are all very common. Some texts may even jump from one abbreviation to another in the same page or paragraph!

6.4 Connecting sentences

6.4.1 Atomic sentences

In analysing the measurement results, decision outcomes, hypotheses, assumptions, data and information that enter into an inference problem, it is convenient to find a collection of **basic sentences** or, using a more technical term, **atomic sentences**, out of which all other sentences of interest can be constructed. These atomic sentences often represent elementary pieces of information in the problem.

Consider for instance the following composite sentence, which could appear in our assembly-line scenario:

> "The electronic component is still whole after the shock test and the subsequent heating test. The voltage reported in the final power test is either 90 mV or 110 mV."

In this statement we can identify at least four atomic sentences, which we denote by these symbols:

$$
\begin{aligned}
s &:= \text{‘The component is whole after the shock test’} \\
h &:= \text{‘The component is whole after the heating test’} \\
v_{90} &:= \text{‘The power-test voltage reading is 90 mV’} \\
v_{110} &:= \text{‘The power-test voltage reading is 110 mV’}
\end{aligned}
$$

The inference may actually require additional atomic sentences. For example it might become necessary to consider atomic sentences with other values for the reported voltage, such as

$$
\begin{aligned}
v_{110} &:= \text{‘The power-test voltage reading is 100 mV’} \\
v_{80} &:= \text{‘The power-test voltage reading is 80 mV’}
\end{aligned}
$$

and so on.

6.4.2 Connectives

How do we construct composite sentences, like the one above, out of atomic sentences?

We consider three ways: one operation to change a sentence into another related to it, and two operations to combine two or more sentences together. These operations are called **connectives**. You may have already encountered them in Boolean algebra. Our natural language offers many more operations to combine sentences, but these three connectives turn out to be all we need in virtually all engineering and data-science problems:

Not (symbol $\neg$) example:

$$
\begin{aligned}
s &:= \text{‘The component is whole after the shock test’} \\
\neg s &= \text{‘The component is broken after the shock test’}
\end{aligned}
$$

And (symbols $\wedge$ also ,) example:

s := 'The component is whole after the shock test'

h := 'The component is whole after the heating test'

$s \wedge h$ = 'The component is whole after the shock and heating tests'

$s\,, h$ = 'The component is whole after the shock and heating tests'

Or (symbol $\vee$) example:

v_{90} := 'The power-test voltage reading is 90 mV'

v_{110} := 'The power-test voltage reading is 110 mV'

$v_{90} \vee v_{110}$ = 'The power-test voltage reading is 90 mV, or 110 mV, or both'

These connectives can be applied multiple times, to form increasingly more complex composite sentences.

The `and` connective appears very frequently in probability formulae. Using its standard symbol "$\wedge$" would consume a lot of horizontal space. For this reason a comma "," is often used as an alternative symbol. So the expressions $s \wedge h$ and $s\,, h$ **are completely equivalent**.

⚠ Important subtleties of the connectives:

- There is *no strict correspondence* between the words "not", "and", "or" in natural language and the three connectives. The `and` connective may for instance correspond to the words "but" or "whereas", or just to a comma " , ".
- `Not` means not some kind of complementary quality, but the denial. For instance, $\neg$'The chair is black' generally does not mean 'The chair is white', (although in some situations these two sentences could amount to the same thing).

  It's always best to *declare explicitly what the* ***`not`*** *of a sentence concretely means*. In our example we

take

¬'The component is whole' := 'The component is broken'

But in other examples the negation of "being whole" could comprise several different conditions. A good guideline is to always state the `not` of a sentence in *positive* terms.

- `Or` does not exclude that the sentences it connects can be both true. So in our example $v_{90} \vee v_{110}$ does not exclude, a priori, that the reported voltage could be both 90 mV and 110 mV. (There is a connective for that: "exclusive-or", but it can be constructed out of the three we already have.)

From the last remark we see that the sentence

'The power-test voltage reading is 90 mV or 110 mV'

does *not* correspond to $v_{90} \vee v_{110}$. It is implicitly understood that a voltage reading cannot yield two different values at the same time. Convince yourself that the correct way to write that sentence is this:

$$(v_{90} \vee v_{110}) \wedge \neg(v_{90} \wedge v_{110})$$

Finally, the full composite sentence of the present example can be written in symbols as follows:

> "The electronic component is still whole after the shock test and the subsequent heating test. The voltage reported in the final power test is either 90 mV or 110 mV."

$$s \wedge h \wedge (v_{90} \vee v_{110}) \wedge \neg(v_{90} \wedge v_{110})$$

Study reading

Take a quick look at §7.4.1 in *Artificial Intelligence* and note the similarities with what we've just learned. In these notes we follow a faster approach leading directly to probability logic.

6.5 "If... then..."

Sentences expressing data and information in natural language also appear connected with *if... then...*. For instance: "If the voltage reading is 200 mV, then the component is defective". This kind of expression actually indicates that the following inference

$$\text{'The component is defective'} \,\big|\, \text{'The voltage reading is 200 mV'}$$

is `true`.

This kind of information is very important because it is often the starting point of our inferences. We shall discuss this point in more detail in the next sections.

Careful

There is a connective in logic, called "material conditional", which is also often translated as "if... then...". But it is not the same as the inference relation discussed above. "If... then..." in natural language usually denotes an inference rather than a material conditional.
Research is still ongoing on these topics. If you are curious and in for a headache, look over *The logic of conditionals*.

We are now equipped with all the notions and symbolic notation to deal with our next task: learning the rules for drawing correct inferences.

@@ TODO: add connections to impossibility of large language models to learn maths (Gödel & Co.).

7 Truth inference

Some inferences can be drawn with absolute certainty, that is, we can ascertain for sure the truth or falsity of their proposal. We call this particular "sure" kind of inferences *truth inferences*. Mathematical inferences are a typical example of this kind. You probably have some acquaintance with rules for drawing truth inferences, so we start from these.

7.1 A trivial inference

Consider again the assembly-line scenario of § 1, and suppose that an inspector has the following information about an electric component:

> This electric component had an early failure (within a year of use). If an electric component fails early, then at production it didn't pass either the shock test or the heating test. This component passed the shock test.

The inspector wants to assess whether the component did not pass the heating test.

From the data and information given, the conclusion is that the component *for sure* did not pass the heating test. This conclusion is certain and somewhat trivial. But how did we obtain it? Which rules did we follow to arrive at it from the given data?

Formal logic, with its *deduction systems*, is the huge field that formalizes and makes rigorous the rules that a rational person or an artificial intelligence should use in drawing *sure* inferences like the one above. We'll now get a glimpse of it, as a trampoline for jumping towards more general and *uncertain* inferences.

7.2 Analysis and representation of the problem

First let's analyse our simple problem and represent it with compact symbols.

7.2.1 Atomic sentences

We can introduce the following atomic sentences and symbols:

$h :=$ 'The component passed the heating test'

$s :=$ 'The component passed the shock test'

$f :=$ 'The component had an early failure'

$I :=$ (all other implicit background information)

7.2.2 Proposal

The proposal is $\neg h$, but in the present case we could also have chosen h.

7.2.3 Conditional

The bases for the inference are two known facts in the present case: s and f. There may also be other obvious facts implicitly assumed in the inference, which we denote by I.

7.2.4 Starting inferences

Let us emphasize again that any inference is drawn from other inferences, which are either taken for granted, or drawn in turn from others. In the present case we are told that if an electric component fails early, then at production it didn't pass either the shock test or the heating test. We write this as

$$\neg s \vee \neg h \,|\, f \wedge I$$

and we shall take this to be `true` (that is, to have probability 100%).

But our scenario actually has at least one more, hidden, inference. We said that the component failed early, and that it did pass the shock test. This means, in particular, that it must be possible for the component to pass the shock test, even if it fails early. This means that

$$s \mid f \wedge I$$

can*not* be `false`.

7.2.5 Target inference

The inference that the inspector wants to draw can be compactly written:

$$\neg h \mid s \wedge f \wedge I$$

7.3 Truth-inference rules

7.3.1 Deduction systems; a specific choice

Formal logic gives us a set of rules for correctly drawing sure inferences, *when sure inferences are possible*. These rules can be formulated in different ways, leading to a wide variety of deduction systems (each one with a wide variety of possible notations). These systems are all equivalent, of course. The picture on the margin, for instance, shows how a proof of how our inference would look like, using the so-called sequent calculus, which consists of a dozen or so inference rules.

We choose to compactly encode all truth-inference rules in the following way.

First, represent `true` by the number **1**, and `false` by **0**.

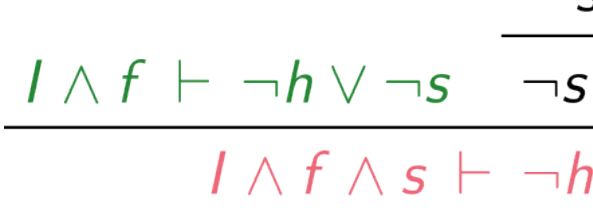


Figure 7.1: The bottom formula is the target
line denotes the application of a
from one or more inferences above
below the line. The two formulae w
are our starting inference, and a ta

Second, symbolically write that a proposal Y is `true`, given a conditional X, as follows:

$$\mathrm{T}(Y \mid X) = 1$$

or "$= 0$" if it's `false`.

The rules of truth-inference are then encoded by the following equations, which must always hold for any atomic or composite sentences X, Y, Z:

Rule for "not":

$$\mathrm{T}(\neg X \mid Z) + \mathrm{T}(X \mid Z) = 1 \tag{7.1}$$

Rule for "and":

$$\mathrm{T}(X \wedge Y \mid Z) = \mathrm{T}(X \mid Y \wedge Z) \cdot \mathrm{T}(Y \mid Z) = \mathrm{T}(Y \mid X \wedge Z) \cdot \mathrm{T}(X \mid Z) \tag{7.2}$$

Rule for "or":

$$\mathrm{T}(X \vee Y \mid Z) = \mathrm{T}(X \mid Z) + \mathrm{T}(Y \mid Z) - \mathrm{T}(X \wedge Y \mid Z) \tag{7.3}$$

Rule for truth:

$$\mathrm{T}(X \mid X \wedge Z) = 1 \tag{7.4}$$

How to use the rules: Each equality can be rewritten in different ways according to the usual rules of algebra. Then the resulting left side can be replaced by the right side, and vice versa. The numerical values of starting inferences can be replaced in the corresponding expressions.

Let's see two examples:

- from one rule for "and" we can obtain the equality

$$\mathrm{T}(X \mid Y \wedge Z) = \frac{\mathrm{T}(X \wedge Y \mid Z)}{\mathrm{T}(Y \mid Z)}$$

  provided that $\mathrm{T}(Y \mid Z) \neq 0$. Then wherever we see the left side, we can replace it with the fraction on the right side, and vice versa.

- from the rule for "or" we can obtain the equality

$$\mathrm{T}(X \mid Z) - \mathrm{T}(X \wedge Y \mid Z) = \mathrm{T}(X \vee Y \mid Z) - \mathrm{T}(Y \mid Z)$$

  Again wherever we see the left side, we can replace it with the sum on the right side, and vice versa.

7.3.2 Target inference in our scenario

Let's see how these rules allow us to arrive at our target inference,

$$\mathrm{T}(\neg h \mid s \wedge f \wedge l)$$

starting from the given ones

$$\mathrm{T}(\neg s \vee \neg h \mid f \wedge l) = 1 \ , \qquad \mathrm{T}(s \mid f \wedge l) \neq 0$$

One possibility is to work backwards from the target inference:

Therefore $\mathrm{T}(\neg h \mid s \wedge f \wedge l) = 1$. We find that, indeed, the electronic component must for sure have failed the heating test!

> **Exercise**
>
> Retrace the proof above step by step. At each step, how was its particular rule (indicated on the right) used?

The way in which the rules can be applied to arrive at the target inference is not unique. In fact, in some concrete applications

$$
\begin{aligned}
&\mathrm{T}(\neg h \mid s \wedge f \wedge l) \\
&\quad = \frac{\mathrm{T}(\neg h \wedge s \mid f \wedge l)}{[1pt]\mathrm{T}(s \mid f \wedge l)_{\neq 0}} && \text{-rule and starting inference} \\
&\quad = \frac{\mathrm{T}(s \mid \neg h \wedge f \wedge l) \cdot \mathrm{T}(\neg h \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \text{-rule} \\
&\quad = \frac{\left[1 - \mathrm{T}(\neg s \mid \neg h \wedge f \wedge l)\right] \cdot \mathrm{T}(\neg h \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \neg\text{-rule} \\
&\quad = \frac{\mathrm{T}(\neg h \mid f \wedge l) - \mathrm{T}(\neg s \mid \neg h \wedge f \wedge l) \cdot \mathrm{T}(\neg h \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \text{algebra} \\
&\quad = \frac{\mathrm{T}(\neg h \mid f \wedge l) - \mathrm{T}(\neg s \wedge \neg h \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \text{-rule} \\
&\quad = \frac{\mathrm{T}(\neg s \vee \neg h \mid f \wedge l) - \mathrm{T}(\neg s \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \text{-rule} \\
&\quad = \frac{1 - \mathrm{T}(\neg s \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \text{starting inference} \\
&\quad = \frac{\mathrm{T}(s \mid f \wedge l)}{\mathrm{T}(s \mid f \wedge l)} && \neg\text{-rule} \\
&\quad = 1 && \text{algebra}
\end{aligned}
$$

it can require a lot of work to find how to connect target inference with starting ones via the rules. The result, however, will always be the same:

The rules of truth-inference are self-consistent: even if applied in different sequences of steps, they always lead to the same final result.

Exercise

Prove the target inference $\mathrm{T}(\neg h \mid s \wedge f \wedge l) = 1$ using the rules of truth-inference, but beginning from the starting inference $\mathrm{T}(\neg s \wedge \neg h \mid f \wedge l) = 1$.

7.3.3 [*Optional*] Equivalence with truth-tables

If you have studied Boolean algebra, you may be familiar with truth-tables; for instance the one for "and" displayed on the side. The truth-inference rules (7.1)–(7.4) contain the truth-tables that you already know as special cases.

X	Y	$X \wedge Y$
1	1	1
1	0	0
0	1	0
0	0	0

Exercise

Use the truth-inference rules for "or" and "and" to build the truth-table for "or". Check if it matches the one you already knew.

The truth-inference rules (7.1)–(7.4) are more complicated than truth-tables, but have two important advantages. First, they allow us to work with conditionals, and to move sentences between proposals and conditionals. Second, they provide a smoother transition to the rules for probability-inference.

7.4 Logical AI agents and their limitations

The truth-inference discussed in this section are also the rules that a *logical AI agent* should follow. For example, the automated control and fault-management programs in NASA spacecrafts, mentioned in § 6.1, are programmed according to these rules.

Study reading

Look over Ch. 7 in *Artificial Intelligence*.

Many – if not most – inference problems that human and AI agents must face are, however, of the *uncertain* kind: it is not possible to surely infer the truth of some outcome, and the truth of some initial data or initial inferences may not be known either. We shall now see how to generalize the truth-inference rules to uncertain situations.

For the extra curious

Our cursory visit of formal logic only showed a microscopic part of this vast field. The study of truth-inference rules continues still today, with many exciting developments and applications. Feel free to take a look at

- *Logic in Computer Science*
- *Mathematical Logic for Computer Science*
- *Natural Deduction Systems in Logic*

8 Probability inference

In most engineering and data-science problems we don't know the truth or falsity of outcomes and hypotheses that interest us. But this doesn't mean that nothing can be said or done in such situations. Now we shall finally see how to draw *uncertain* inferences, that is, how to calculate the *probability* of something that interests us, given particular data, information, and assumptions.

So far we have used the term "probability" somewhat informally and intuitively. It is time to make it more precise and to emphasize some of its most important aspects. Then we'll dive into the rules of probability-inference.

8.1 When truth isn't known: probability

When we cross a busy city street we look left and right to check whether any cars are approaching. We typically don't look *up* to check whether something is falling from the sky. Yet, couldn't it be `false` that cars are approaching at that moment? and couldn't it be `true` that some object is falling from the sky? Of course both events are possible. Then why do we look left and right, but not up?

The main reason is that we *believe strongly* that cars might be approaching, and *believe very weakly* that some object might be falling from the sky. In other words, we consider the first occurrence to be *very probable*, and the second extremely *improbable*.

We shall take the notion of **probability** as intuitively understood (just as we did with the notion of truth). Terms equivalent to "probability" are *degree of belief*, *plausibility*,

credibility[1], *certainty*.

Probabilities are quantified between 0 and 1, or equivalently between 0% and 100%. Assigning to a sentence a probability `1` is the same as saying that it is `true`; and a probability `0`, that it is `false`. A probability of 0.5 represents a belief completely symmetric with respect to truth and falsity.

Alternatively, if an agent assigns to a sentence a probability `1`, it means that the agent is completely certain that the sentence is `true`. If the agent assigns a probability `0`, it means that the agent is completely certain that the sentence is `false`. If the agent assigns a probability `0.5`, it means that the agent is equally uncertain about the truth as about the falsity of the sentence.

Let's emphasize and agree on some important facts about probabilities:

- 👉 **Probabilities are assigned to *sentences*.** We already discussed this point in § 6.3, but let's reiterate it. Consider an engineer working on a problem of electric-power distribution in a specific geographical region. At a given moment the engineer may believe with 75% probability that the measured average power output in the next hour will be 100 MW. The 75% probability is assigned not to the quantity "100 MW", but to the *sentence*

  'The measured average power output in the next hour will be 100 MW'

  This difference is extremely important. Consider the alternative sentence

  'The average power output in the next hour will be *set* to 100 MW'

[1] *credibility* literally means "believability" (from Latin *credo* = *to believe*).

the numerical quantity is the same, but the meaning is very different. The probability can therefore be very different. If the engineer is the person who decides how to set that output, and has decided to set it to 100 MW, then the probability is obviously 100% (or very close to), because the engineer already knows what the output will be. The probability depends not only on a number, but on what it's being done with that number: measuring, setting, third-party reporting, and so on. Often we write simply "O=10 W", provided that the full sentence behind this shorthand is understood.

- ☛ **Probabilities are agent- and knowledge-dependent**. A coin is tossed, comes down heads, and is quickly hidden from view. Alice sees that it landed heads-up. Bob instead doesn't manage to see the outcome and has no clue. Alice considers the sentence 'Coin came down heads' to be `true`, that is, to have 100% probability. Bob considers the same sentence to have 50% probability.

  Note how Alice and Bob assign two different probabilities to the same sentence; yet both assignments are completely rational. If Bob assigned 100% to 'heads', we would suspect that he had seen the outcome after all. If he assigned 0% to 'heads', we would consider it unreasonable (he didn't see the outcome, so why exclude 'heads'?). At the same time we would be baffled if Alice assigned only 50% to 'heads', because she actually saw that the outcome was heads; maybe we would wonder whether she feels unsure about what she saw.

  An omniscient agent would know the truth or falsity of every sentence, and assign only probabilities `0` or `1`. Some authors speak of "*actual* (but unknown) probabilities". But if there were "actual" probabilities, they would be all `0` or `1`, and it would be pointless to speak about probabilities at all – every inference would be a truth-inference.

- ☛ **Probabilities are not frequencies**. Consider the fraction of defective mechanical components to total components produced per year in some factory. This quantity can be physically measured and, once measured, would

be agreed upon by every agent. It is a *frequency*, not a degree of belief or probability.

It is important to understand the difference between *probability* and *frequency*: mixing them up may lead to suboptimal decisions. Later we shall say more about the difference and the precise relations between probability and frequency.

Frequencies can be unknown to some agents. Probabilities cannot be "unknown": they can only be difficult to calculate. Be careful when you read authors speaking of an "unknown probability": they actually mean either "unknown frequency", or a probability that has to be calculated (it's "unknown" in the same sense that the value of $1 - 0.7 \cdot 0.2/(1 - 0.3)$ is "unknown" to you right now).

- 👉 **Probabilities are not physical properties**. Whether a tossed coin lands heads up or tails up is fully determined by the initial conditions (position, orientation, momentum, rotational momentum) of the toss and the boundary conditions (air velocity and pressure) during the flight. The same is true for all macroscopic engineering phenomena (even quantum phenomena have never been proved to be non-deterministic, and there are deterministic and experimentally consistent mathematical representations of quantum theory). So we cannot measure a probability using some physical apparatus; and the mechanisms underlying any engineering problem boil down to physical laws, not to probabilities.

📖 Study reading

Dynamical Bias in the Coin Toss.

These points listed above are not just a matter of principle. They have important practical consequences. A data scientist who is not attentive to the source of the data (measured? set? reported, and so maybe less trustworthy?), or who does not carefully assess the context of a probability, or who mixes

a probability with a frequency, or who does not take advantage (when possible) of the physics involved in the a problem – such data scientist will design systems with sub-optimal performance[2] – or even cause deaths.

8.2 ⚠ The many uses of the word "probability"

In these notes we shall consistently use the term "probability" in the sense explained above. But beware that this term is used in many different and incompatible senses, depending on whom you're speaking with or which literature you're reading.

Some people use this term in the sense of "frequency": the number of times something happened in a series of repetitions. But a frequency is an objective, measurable quantity; it doesn't depend on the knowledge of an agent. To us is not useful, because it doesn't quantify the belief or certainty of an agent. Suppose a coin is tossed 100 times, and it comes up heads 80 times. The frequency of heads is 80/100. Now suppose that an agent *that does not know anything about the 100 tosses* is asked to predict whether the next toss will be heads or tails. What should the agent's degree of belief or of certainty be? Obviously it should be 50%/50%. If we were to program the agent so that it has a degree of belief of 80% for heads *in situations where nothing is know about previous tosses* (because that's the situation our agent was in), then such an agent would on average lose big time in dealing with new coins.

But frequency is *data*, and if a frequency is known, then obviously an agent should take it into account in quantifying its credibility. If an agent *knows* that the coin came up heads 80 times in 100, then it is reasonable that the agent's degree of belief for the next toss should be around 80% for heads. And we shall see that this is indeed what happens.

So the distinction between "frequency" and "probability" is crucial. Frequencies do not enable us to quantify an agent's beliefs in situations where data are missing.

[2]This fact can be mathematically proven.

Some people use "probability" in the sense of the number of *a-priori* successes over the number of possibilities. For instance, if you roll a die, and the die comes up ⚄, then this result was 1 among six possible results. This is fine, but this definition does not allow us to capture the difference between an agent who has seen the outcome of the roll was ⚄, and an agent who has *not* seen the outcome of the roll. We expect the two agents to behave differently. If we programmed the agent to have a degree of belief of 1/6 *even after seeing the outcome of a die roll*, we would have programmed an agent incapable of using new data (seeing the outcome). Would you be happy if a clinician made some medical tests, and then ignored the results of the tests, behaving as if they were still unknown?

All these different uses are just a matter of semantics, and in the end it doesn't matter which word we use, as long as we understand its meaning and as long as we're adopting the meaning that is *useful* for our present task.

> ⚠ Beware of *likelihood* as a synonym for *probability*
>
> In everyday language, "likelihood" is synonym with "probability". In technical writings about probability or statistics, however, "likelihood" means something different and is *not* a synonym of "probability", as we explain below (§ 8.8.1).

8.3 An unsure inference. Probability notation

Consider now the following variation of the trivial inference problem of § 7.1.

> This electric component had an early failure. If an electric component fails early, then at production ei-

ther it didn't pass the heating test or it didn't pass the shock test. The probability that it passed neither test (that is, *both* tests failed) is 10%. There's no reason to believe that the component passed the heating test, more than to believe that it passed the shock test.

Again the inspector wants to assess whether the component *did not pass* the heating test.

From the data and information given, what would you say is the probability that the component didn't pass the heating test?

Exercises

- Try to argue why a conclusion cannot be drawn with certainty in this case. One way to argue this is by presenting two different scenarios that fit the given data but have opposite conclusions.
- Try to reason intuitively and assess the probability that the component didn't pass the heating test. Should it be larger or smaller than 50%? Why?

For this inference problem we cannot find a `true` or `false` final value. The truth-inference rules (7.1)–(7.4) therefore cannot help us here. In fact even the "T(... |...)" notation is unsuitable, because it only admits the values 1 (`true`) and 0 (`false`).

Let us first generalize this notation in a straightforward way:

First, let's represent the probability or degree of belief of a sentence by a number in the range $[0, 1]$, that is, between **1** (certainty or `true`) and **0** (impossibility or `false`). The value 0.5 represents a belief in the truth of the sentence which is as strong as the belief in its falsity.

Second, let's symbolically write in the following way that the probability of a proposal Y, given a conditional X, is some number p:

$$\mathrm{P}(Y \mid X) = p$$

Note that this notation includes the notation for truth-values as a special case:

$$\mathrm{P}(Y \mid X) = 0 \text{ or } 1 \quad \iff \quad \mathrm{T}(Y \mid X) = 0 \text{ or } 1$$

8.4 Inference rules

Extending our truth-inference notation to probability-inference notation has been straightforward. But which rules should we use for drawing inferences when probabilities are involved?

The amazing result is that *the rules for truth-inference, formulae (7.1)–(7.4), extend also to probability-inference*. The only difference is that they now hold for all values in the range $[0, 1]$, rather than only for 0 and 1.

This important result was taken more or less for granted at least since Laplace in the 1700s, but was formally proven for the first time in 1946 by R. T. Cox. The proof has been refined since then. What kind of proof is it? It shows that if we don't follow the rules we are doomed to arrive at illogical conclusions; we'll show some examples later.

Finally, here are *the fundamental rules of all inference*. They are encoded by the following equations, which must always hold for any atomic or composite sentences X, Y, Z:

It is amazing that **ALL** inference is nothing else but a repeated application of these four rules – maybe billions of times or more. All machine-learning algorithms are just applications or approximations of these rules. Methods that you may have heard about in statistics are just specific applications of these rules. Truth inferences are also special applications of these rules. Most of this course is just a study of how to apply these rules to particular kinds of problems.

🏛 THE FUNDAMENTAL LAWS OF INFERENCE 🏛

¬ "Not" rule

$$\mathrm{P}(\neg X \mid Z) + \mathrm{P}(X \mid Z) = 1$$

∧ "And" rule

$$\mathrm{P}(X \wedge Y \mid Z) = \mathrm{P}(X \mid Y \wedge Z) \cdot \mathrm{P}(Y \mid Z) = \mathrm{P}(Y \mid X \wedge Z) \cdot \mathrm{P}(X \mid Z)$$

∨ "Or" rule

$$\mathrm{P}(X \vee Y \mid Z) = \mathrm{P}(X \mid Z) + \mathrm{P}(Y \mid Z) - \mathrm{P}(X \wedge Y \mid Z)$$

Truth rule

$$\mathrm{P}(X \mid X \wedge Z) = 1$$

How to use the rules:

Each equality can be rewritten in different ways according to the usual rules of algebra. Then the resulting left side can be replaced by the right side, and vice versa. The numerical values of starting inferences can be replaced in the corresponding expressions.

Study reading

- Skim through *Probability, Frequency and Reasonable Expectation*. Try to get the ideas behind the reasoning, even if you can't follow the mathematical details.
- Ch. 2 of *Bayesian Logical Data Analysis for the Physical Sciences*
- Ch. 1 of *Probability*
- §§1.0–1.2 of *Data Analysis*
- Skim through Chs 1–2 of *Probability Theory*

The fundamental inference rules are used in the same way as their truth-inference counterpart of [§@truth-inference-rules]: Each equality can be rewritten in different ways according to the usual rules of algebra. The left and right side of the equality thus obtained can replace each other in a proof.

8.5 Solution of the uncertain-inference example

Armed with the fundamental rules of inference, let's solve our earlier inference problem. As usual, we first analyse it and represent it in terms of atomic sentences; we find what are its proposal and conditional; and we find which initial inferences are given in the problem.

8.5.1 Atomic sentences

h := 'The component passed the heating test'

s := 'The component passed the shock test'

f := 'The component had an early failure'

J := (all other implicit background information)

The background information in this example is different from the previous, truth-inference one, so we use the different symbol J for it.

8.5.2 Proposal, conditional, and target inference

The proposal is $\neg h$, just like in the truth-inference example.

The conditional is different now. We know that the component failed early, but we don't know whether it passed the shock test. Hence the conditional is $f \wedge J$.

The target inference is therefore

$$\mathrm{P}(\neg h \mid f \wedge J)$$

8.5.3 Starting inferences

We are told that if an electric component fails early, then at production it didn't pass the heating test or the shock test (or neither). This is given as a sure fact. Let's write it as

$$\mathrm{P}(\neg h \vee \neg s \mid f \wedge J) = 1 \tag{8.1}$$

We are also told that there is a 10% probability that both tests fail:

$$\mathrm{P}(\neg h \wedge \neg s \mid f \wedge J) = 0.1 \tag{8.2}$$

Finally the problem says that there's no reason to believe that the component didn't pass the heating test, more than it didn't pass the shock test. This can be written as follows:

$$\mathrm{P}(h \mid f \wedge J) = \mathrm{P}(s \mid f \wedge J) \tag{8.3}$$

Note the interesting situation above: we are not given the numerical values of these two probabilities; we are only told that

they are equal. This is an example of application of the *principle of indifference*, which we'll discuss more in detail later.

8.5.4 Final inference

Also in this case there is no unique way of applying the rules to reach our target inference, but *all paths will lead to the same result*. Let's try to proceed backwards:

$$
\begin{aligned}
&\mathrm{P}(\neg h \mid f \wedge J) && \text{-rule}\\
&\quad = \mathrm{P}(\neg s \vee \neg h \mid f \wedge J) + \mathrm{P}(\neg s \wedge \neg h \mid f \wedge J) - \mathrm{P}(\neg s \mid f \wedge J) && \text{starting inferences (8.1–2)}\\
&\quad = 1 + 0.1 - \mathrm{P}(\neg s \mid f \wedge J) && \neg\text{-rule}\\
&\quad = 0.1 + \mathrm{P}(s \mid f \wedge J) && \text{starting inference (8.3)}\\
&\quad = 0.1 + \mathrm{P}(h \mid f \wedge J) && \neg\text{-rule}\\
&\quad = 0.1 + 1 - \mathrm{P}(\neg h \mid f \wedge J)
\end{aligned}
$$

The target probability appears on the left and right side with opposite signs. We can solve for it:

$$
\begin{aligned}
2\,\mathrm{P}(\neg h \mid f \wedge J) &= 0.1 + 1\\
\mathrm{P}(\neg h \mid f \wedge J) &= 0.55
\end{aligned}
$$

So the probability that the component didn't pass the heating test is 55%.

Exercises

- Try to find an intuitive explanation of why the probability is 55%, slightly larger than 50%. If your intuition says this probability is wrong, then:
  - Check the proof of the inference for mistakes, or try to find a proof with a different path.
  - Examine your intuition critically and educate it.
- Check how the target probability $\mathrm{P}(\neg h \mid f \wedge J)$

changes if we change the value of the probability $\mathrm{P}(\neg s \wedge \neg h \mid f \wedge J)$ from 0.1.

- What result do we obtain if $\mathrm{P}(\neg s \wedge \neg h \mid f \wedge J) = 0$? Can it be intuitively explained?
- What if $\mathrm{P}(\neg s \wedge \neg h \mid f \wedge J) = 1$? Does the result make sense?

8.6 How the inference rules are used

In the solution above you noticed that the equations of the fundamental rules are not only used to obtain some of the probabilities appearing in them from the remaining probabilities.

The rules represent, first of all, *constraints of logical consistency*[3] among probabilities. For instance, if we have probabilities $\mathrm{P}(Y \mid X \wedge Z) = 0.1$, $\mathrm{P}(X \mid Z) = 0.7$, and $\mathrm{P}(X \wedge Y \mid Z) = 0.2$, then there's an inconsistency somewhere, because these values violate the and-rule: $0.2 \neq 0.1 \cdot 0.7$. In this case we must find the inconsistency and solve it. However, since probabilities are quantified by real numbers, it's possible and acceptable to have slight discrepancies within numerical round-off errors.

The rules also imply more general constraints. For example we must *always* have

$$\mathrm{P}(X \wedge Y \mid Z) \leq \min\{\mathrm{P}(X \mid Z),\ \mathrm{P}(Y \mid Z)\}$$
$$\mathrm{P}(X \vee Y \mid Z) \geq \max\{\mathrm{P}(X \mid Z),\ \mathrm{P}(Y \mid Z)\}$$

Exercise

Try to prove the two constraints above.

[3] The technical term is **coherence**.

8.7 Consequences of not following the rules

The fundamental rules of inference guarantee that the agent's uncertain reasoning is self-consistent, and that it follows logic when there's no uncertainty. Breaking the rules means that the resulting inference has some logical or irrational inconsistencies.

There are many examples of inconsistencies that appear when the rules are broken. Imagine for instance an agent that gives an 80% probability that it rains[4] in the next hour; and it also gives a 90% probability that it rains *and* that the average wind is above 3 m/s in the next hour. This is clearly unreasonable, because the raining scenario alone would be true with wind above 3 m/s *and also* below 3 m/s – therefore it should be *more* probable than the scenario where the wind is above 3 m/s. And indeed the two given probabilities break the `and`-rule, showing that they are unreasonable or illogical.

Exercise

Prove that the two probabilities in the example above break the `and`-rule. (Hint: you must use the fact that probabilities are numbers between 0 and 1, and that multiplying a number by something between 0 and 1 can only yield a smaller number.)

Study reading

- §12.2.3 in *Artificial Intelligence*
- As you continue your studies, go through chapters 4–8 of *Rational Choice in an Uncertain World*, just to get the main messages and an overview of curious psychological phenomena.

[4] to be precise, let's say "it rains above 1 mm"

8.8 Remarks on terminology and notation

8.8.1 Likelihood

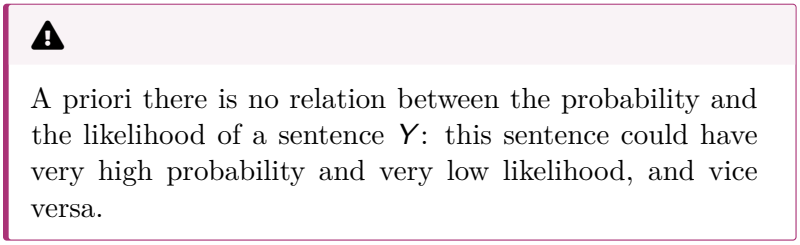
In everyday language, "likely" is often a synonym of "probable", and "likelihood" of "probability". But in technical writings about probability, inference, and decision-making, "likelihood" has a very different meaning. Beware of this important difference in definitions:

$\mathrm{P}(Y \mid X)$ is:

- the **probability of Y given X** (or **conditional on X**),
- the **likelihood of X in view of Y**.

We can also say:

- the **probability of Y** given X, is $\mathrm{P}(Y \mid X)$.
- the **likelihood of Y** in view of X, is $\mathrm{P}(X \mid Y)$.



A priori there is no relation between the probability and the likelihood of a sentence Y: this sentence could have very high probability and very low likelihood, and vice versa.

In these notes we'll avoid the possibly confusing term "likelihood". All we need to express can be phrased in terms of probability.

8.8.2 Omitting background information

In the analyses of the inference examples of § 7.1 and § 8.3 we defined sentences (I and J) expressing all background information, and always included these sentences in the conditionals of the inferences – because those inferences obviously depended on that specific background information.

In many concrete inference problems the background information usually stays in the conditional from beginning to end,

while the other sentences jump around between conditional and proposal as we apply the rules of inference. For this reason the background information is often omitted from the notation, being implicitly understood. For instance, if the background information is denoted I, one writes

- "$\mathrm{P}(Y \mid X)$" instead of $\mathrm{P}(Y \mid X \wedge I)$
- "$\mathrm{P}(Y)$" instead of $\mathrm{P}(Y \mid I)$

This is what's happening in books where you see "$P(x)$" without conditional.

Such practice may be convenient, but be wary of it, especially in particular situations:

- In some inference problems we suddenly realize that we must distinguish between cases that depend on hypotheses, say H_1 and H_2, that were buried in the background information I. If the background information I is explicitly reported in the notation, this is no problem: we can rewrite it as

  $$I = (H_1 \vee H_2) \wedge I'$$

  and then proceed as usual. If the background information was not explicitly written, this may lead to confusion and mistakes: there may suddenly appear two instances of $\mathrm{P}(X)$ with *different* values, just because one of them is invisibly conditional on I, the other on I'.
- In some inference problems we are considering *several different* instances of background information – for example because more than one agent is involved. It's then extremely important to write the background information explicitly, lest we mix up the degrees of belief of different agents.

This kind of confusion from poor notation happens more often than one thinks, and even appears in the scientific literature.

For the extra curious

A once-famous paper published in the quantum-theory literature arrived at completely wrong results simply by omitting background information, mixing up probabilities having different conditionals.

8.8.3 "Random variables"

Some texts speak of the probability of a "random variable", or more precisely of the probability "that a random variable takes on a particular value". As you notice, we have just expressed that idea by means of a *sentence*. The viewpoint and terminology of random variables is therefore a special case of that based on sentences, which we use here.

The dialect of "random variables" does not offer any advantages in concepts, notation, terminology, or calculations, and it has several shortcomings:

- As discussed in § 8.1, in concrete applications it is important to know how a quantity "takes on" a value: for example it could be directly measured, indirectly reported, or purposely set to that specific value. Thinking and working in terms of sentences, rather than of random variables, allows us to account for these important differences.
- We want a general AI agent to be able to deal with uncertainty and probability also in situations that do not involve mathematical sets.
- Very often the object (proposal) of a probability is not a "variable": it is actually a *constant* value that is simply unknown (simple example: we are uncertain about the mass of a particular block of concrete, so we speak of the probability of some mass value; this doesn't mean that the mass of the block of concrete is changing).
- What does "random" (or "chance") mean? Good luck finding an understandable and non-circular definition in texts that use that word. Strangely enough, texts that use that word never define it. In these notes, if the word "random" is ever used, it stands for "unpredictable" or "unsystematic".

It's a question for sociology of science why some people keep on using less flexible points of view or terminologies. Probably they just memorize them as students and then a fossilization process sets in.

Figure 8.1: James Clerk Maxwell is one of the r
statistical mechanics and kinetic t
tromagnetism). Yet he never used
dom" in his technical writings. M
for being very clear and meticulou
tions and terminology.

Finally, some texts speak of the probability of an "event". For all purposes, an "event" is just what's expressed in a sentence.

9 Shortcut rules

The fundamental rules introduced in chapter 8 are all we need, and all an AI needs, in order to draw inferences from other inferences and from initial data.

From them, however, it is possible to derive some "shortcut" rules than can make the inferences shorter and faster. The situation is similar to what happens with some rules in algebra: for instance, we know that whenever we find the expression

$$(a+b)\cdot(a-b)$$

then we can automatically substitute it with

$$a^2-b^2$$

no matter the values of a and b. The rule "$(a+b)\cdot(a-b)=a^2-b^2$" is not a new algebraic rule: it's simply the result of the application of the rules for addition $+$ and multiplication $\cdot$, and indeed we could just apply them directly:

$$\begin{aligned}(a+b)\cdot(a-b) &= a\cdot a+b\cdot a-a\cdot b-b\cdot b\\ &= a^2+b\cdot a-b\cdot a-b^2\\ &= a^2-b^2\end{aligned}$$

But if we remember that they always lead to the result a^2-b^2, then we can directly use the "shortcut" rule $(a+b)\cdot(a-b)=a^2-b^2$ and save ourselves some time.

Likewise with the four rules of inference. Some particular sequences of application of the rules occur very often. We can

then simply memorize the starting and final steps of these sequences, and use them directly, skipping all the steps in between. These shortcut rules are not only useful for saving time, however. We shall see that they reveal interesting and intuitive inference patterns, which are implicit in the four inference rules.

It is possible and legitimate to implement these shortcut rules in an AI agent, besides the four fundamental ones. Such an agent will arrive at the same results and decisions of an identical AI agent that doesn't use the shortcut rules – but a little faster.

Here are the shortcut rules we'll frequently use in the rest of the course:

9.1 Falsity and truth cannot be altered by additional knowledge

Suppose that sentence X is judged to be completely impossible, conditional on sentence Z:

$$\mathrm{P}(X \mid Z) = 0$$

It can then be proved, from the fundamental rules, that X is also completely impossible if we add information to Z. That is, for any sentence Y we'll also have

$$\mathrm{P}(X \mid Y \wedge Z) = 0$$

> **Exercise**
>
> Try to prove this. (Hint: try using the `and`-rule one or more times.)

What if we use $\neg X$ for Y, that is, what if we acquire knowledge that X is actually true? Then it can be proved that all probability calculations break down. The problem is that $\neg X$ and Z turn out to be mutually contradictory, so all inferences are starting from contradictory premises. You probably know

that in formal logic if we start from contradictory premises then we can obtain any conclusion whatsoever. The same happens with probability logic.

Note that this problem does not arise, however, if X is only extremely improbable conditional on Z, say with a probability of 10^{-100}, rather than flat-out impossible. In practical applications we often approximate extremely small probabilities by 0, or extremely large ones by 1. If the probability calculations break down, we must then step back and correct the approximation.

By using the `not`-rule it is possible to prove that full certainty about a sentence behaves in a similar manner. If sentence X is judged to be completely certain conditional on sentence Z:

$$\mathrm{P}(X \mid Z) = 1$$

then, from the fundamental rules, X is also completely certain if we add information to Z. That is, for any sentence Y we'll also have

$$\mathrm{P}(X \mid Y \wedge Z) = 1$$

Shortcut rules: permanence of truth and falsity

if $\mathrm{P}(X \mid Z) = 0$ or 1
then $\mathrm{P}(X \mid Y \wedge Z) = \mathrm{P}(X \mid Z)$ for any Y not contradicting Z

9.2 Boolean algebra

It is possible to show that all rules you may know from Boolean algebra *are a consequence of the fundamental rules* of § 8.4.

So we can always make the following convenient replacements anywhere in a probability expression:

Shortcut rules: Boolean algebra

$$\neg\neg X = X$$

$$X \wedge X = X$$

$$X \vee X = X$$

$$X \wedge Y = Y \wedge X$$

$$X \vee Y = Y \vee X$$

$$X \wedge (Y \vee Z) = (X \wedge Y) \vee (X \wedge Z)$$

$$X \vee (Y \wedge Z) = (X \vee Y) \wedge (X \vee Z)$$

$$\neg(X \wedge Y) = \neg X \vee \neg Y$$

$$\neg(X \vee Y) = \neg X \wedge \neg Y$$

For example, if we have the probability

$$\mathrm{P}[X \vee (Y \wedge Y) \mid (\neg\neg Z) \wedge I]$$

we can directly replace it with

$$\mathrm{P}[X \vee Y \mid Z \wedge I]$$

The derivation of the Boolean-algebra rules from the four inference rules is somewhat involved. As as example, a partial proof of the rule $X \wedge X = X$, called "`and`-idempotence" goes as follows:

$$\begin{aligned}
&\mathrm{P}(X \wedge X \mid Z) \\
&\quad = \mathrm{P}(X|X \wedge Z) \cdot \mathrm{P}(X \mid Z) && \text{-rule} \\
&\quad = 1 \cdot \mathrm{P}(X \mid Z) && \text{truth-rule} \\
&\quad = \mathrm{P}(X \mid Z)
\end{aligned}$$

and with a similar procedure it can be shown that $X \wedge X$ can be replaced with X no matter where it appears. The above proof shows that the `and`-idempotence rule is tightly connected with the `truth`-rule of inference.

9.3 Law of total probability or "extension of the conversation"

Suppose we have a set of n sentences $\{H_1, H_2, \ldots, H_n\}$ having these two properties:

- They are **mutually exclusive**, meaning that the "and" of any two of them is false, given some background knowledge Z:

$$\mathrm{P}(H_1 \wedge H_2 \mid Z) = 0 \ , \quad \mathrm{P}(H_1 \wedge H_3 \mid Z) = 0 \ , \quad \ldots \ , \quad \mathrm{P}(H_{n-1} \wedge H_n \mid Z) = 0$$

- They are **exhaustive**, meaning that the "or" of all of them is true, given the background knowledge Z:

$$\mathrm{P}(H_1 \vee H_2 \vee \cdots \vee H_n \mid Z) = 1$$

In other words, according to our background knowledge, one of those sentences *must* be true, but *only one*.

Then the probability of a sentence X, conditional on Z, is equal to a combination of probabilities conditional on $H_1, H_2, \ldots$:

Shortcut rule: extension of the conversation

$$\begin{aligned}&\mathrm{P}(X \mid Z) \\ &\quad = \mathrm{P}(X \mid H_1 \wedge Z) \cdot \mathrm{P}(H_1 \mid Z) + \mathrm{P}(X \mid H_2 \wedge Z) \cdot \mathrm{P}(H_2 \mid Z) + \cdots + \mathrm{P}(X \mid H_n \wedge Z) \cdot \mathrm{P}(H_n \mid Z)\end{aligned}$$

This rule is useful when it is difficult to assess the probability of a sentence conditional on the background information, but it is easier to assess the probability of that sentence conditional on

several auxiliary "scenarios" or hypotheses[1]. The name **extension of the conversation** for this shortcut rule comes from the fact that we are able to call these additional scenarios or hypotheses into play. This situation occurs very often in concrete applications.

9.4 Bayes's theorem

The probably most famous – or infamous – rule derived from the laws of inference is **Bayes's theorem**. It allows us to relate the probability of a proposal Y and a conditional X to the probability where their proposal-conditional roles are exchanged:

Shortcut rule: Bayes's theorem

$$\mathrm{P}(Y \mid X \wedge Z) = \frac{\mathrm{P}(X \mid Y \wedge Z) \cdot \mathrm{P}(Y \mid Z)}{\mathrm{P}(X \mid Z)}$$

Obviously this rule can only be used if $\mathrm{P}(X \mid Z) > 0$, that is, if the sentence X is not false conditional on Z.

Bayes's theorem is extremely useful when we want to assess the probability of a hypothesis (the proposal) given some data (the conditional), and it is easy to assess the probability of the data conditional on the hypothesis. Note, however, that the sentences Y and X in the theorem can be about anything whatsoever: Y doesn't always need to be a "hypothesis", and X doesn't always need to be "data".

Exercise

Prove Bayes's theorem from the fundamental rules of inference.

Study reading

§8.8 of *Rational Choice in an Uncertain World*

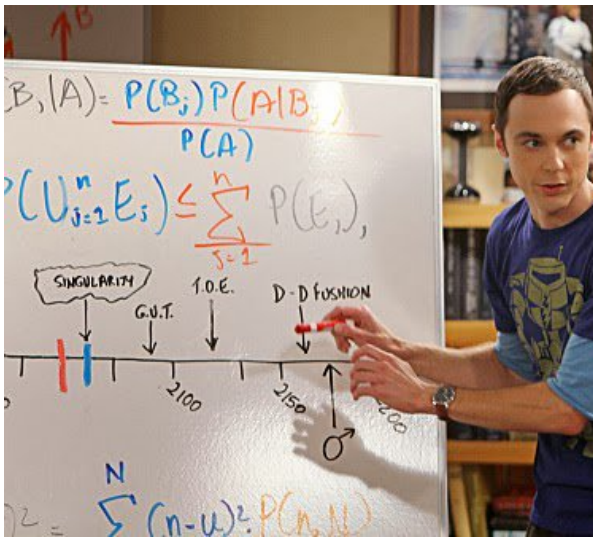


Figure 9.1: Bayes's theorem guest-starring in *Theory*

[1] this is why we used the symbol H for these sentences

9.5 Bayes's theorem & extension of the conversation

Bayes's theorem is often used with several sentences $\{Y_1, Y_2, \ldots, Y_n\}$ that are mutually exclusive and exhaustive. Typically these represent competing hypotheses. In this case the probability of the sentence X in the denominator can be expressed using the rule of extension of the conversation:

> Shortcut rule: Bayes's theorem with extension of the conversation
>
> $$\mathrm{P}(Y_1 \mid X \wedge Z) = \frac{\mathrm{P}(X \mid Y_1 \wedge Z) \cdot \mathrm{P}(Y_1 \mid Z)}{\mathrm{P}(X \mid Y_1 \wedge Z) \cdot \mathrm{P}(Y_1 \mid Z) + \cdots + \mathrm{P}(X \mid Y_n \wedge Z) \cdot \mathrm{P}(Y_n \mid Z)}$$
>
> and similarly for Y_2 and so on.

We will use this form of Bayes's theorem very frequently.

9.6 The many facets of Bayes's theorem

Bayes's theorem is a very general result of the fundamental rules of inference, valid for any sentences X, Y, Z. This generality leads to many uses and interpretations.

The theorem is often proclaimed to be the rule for "updating an agent's beliefs". The meaning of this proclamation is the following. Let's say that at some point Z represents all the agent's knowledge. The agent's degree of belief about some sentence Y is then (at least in theory) the value of $\mathrm{P}(Y \mid Z)$. At some later point, the agent gets to know – maybe thanks to an observation or measurement – that the sentence X is true. The agent's whole knowledge at that point is represented no longer by Z, but by $X \wedge Z$. The agent's degree of belief about Y is then given by the value of $\mathrm{P}(Y \mid X \wedge Z)$. Bayes's theorem allows us to find the agent's degree of belief about Y conditional on the new state of knowledge, from the one conditional on the old state of knowledge.

This chronological element, however, comes only from this particular way of using Bayes's theorem. The theorem can more

generally be used to connect any two states of knowledge Z and $X \wedge Z$, no matter their temporal order, even if they happen simultaneously, and even if they belong to two different agents.

Exercise

Using Bayes's theorem and the fundamental laws of inference, prove that if $\mathrm{P}(X \mid Z) = 1$, that is, if you already know that X is true in your current state of knowledge Z, then

$$\mathrm{P}(Y \mid X \wedge Z) = \mathrm{P}(Y \mid Z)$$

that is, your degree of belief about Y doesn't change (note that this is different from the rule of truth-permanence of § 9.1).
Is this result reasonable?

Study reading

- §§4.1–4.3 in *Medical Decision Making* give one more point of view on Bayes's theorem.
- Ch. 3 of *Probability*
- A graphical explanation of how Bayes's theorem works mathematically (using a specific interpretation of the theorem):

  https://vimeo.com/852937378?share=copy

9.7 Importance of seemingly trivial rules

Some of the fundamental or shortcut rules may seem obvious or unimportant, but are of extreme importance in data science. For instance, the `and`-idempotence rule $X \wedge X = X$ effectively asserts that *whenever we draw inferences, redundant information or data is* ***automatically*** *counted only once.*

This amazing feature saves us from a lot of headaches. Imagine

that an AI decision agent at the assembly line has been given the following background information: if an electronic component passes the heating test (h), then its probability of early failure (f) is only 10%:

$$\mathrm{P}(f \mid h \wedge Z) = 0.1$$

Now let's say that a new voltage test has also been devised, and if a component passes this test (v) then its probability of early failure is also 10%:

$$\mathrm{P}(f \mid v \wedge Z) = 0.1$$

However, it is discovered that the voltage test works in exactly the same way as the heating test – they're basically the same test! $v = h$. This means that if an element passes the heating test then it will automatically pass the voltage test, and vice versa (they're the same test!):[2]

$$\mathrm{P}(v \mid h \wedge Z) = 1$$

or equivalently $\quad v \wedge h = h \wedge h = h.$

Now suppose that inadvertently we give our AI agent the redundant information that an electronic component has passed the heating test and the voltage test. What will the agent say about the probability of early failure, given this duplicate information? will it count the test twice? Let's calculate:

[2]We are assuming that a test, if repeated, will always give the same result.

$$
\begin{aligned}
&\mathrm{P}(f \mid v \wedge h \wedge Z) \\
&\quad = \frac{\mathrm{P}(f \wedge v \mid h \wedge Z)}{\mathrm{P}(v \mid h \wedge Z)} && \text{-rule} \\
&\quad = \frac{\mathrm{P}(f \wedge v \mid h \wedge Z)}{1} = \mathrm{P}(f \wedge v \mid h \wedge Z) && \text{initial probability} \\
&\quad = \mathrm{P}(v \mid f \wedge h \wedge Z) \cdot \mathrm{P}(f \mid h \wedge Z) && \text{-rule} \\
&\quad = 1 \cdot \mathrm{P}(f \mid h \wedge Z) && \text{truth cannot be altered} \\
&\quad = 0.1 && \text{initial probability}
\end{aligned}
$$

The AI agent, thanks to the `truth`-rule or equivalently the `and`-idempotence rule, correctly detected the redundancy of the sentence v ("the element passed the voltage test") and automatically discarded it.

❗ This feature is of paramount importance in machine learning and data-driven engineering: the "features" that we give as an input to a machine-learning classifier could contain redundancies that we don't recognize, owing to the complexity of the data space. But if the classifier makes inferences according to the four fundamental rules, it will automatically discard any redundant features.

10 Monty Hall and related inference problems

10.1 Motivation: calculation vs intuition

The "Monty Hall problem", inspired by the TV show *Let's make a deal!* hosted by Monty Hall, was proposed in the *Parade* magazine in 1990 (the numbers of the doors are changed here):

> Suppose you are on a game show and given a choice of three doors. Behind one is a car; behind the others are goats. You pick door No. 1, and the host, who knows what is behind them [and wouldn't open the door with the car], opens No. 2, which has a goat. He then asks if you want to pick No. 3. Should you switch?

The web is full of insightful intuitive solutions and of informal probability discussions about this inference problem. Our purpose here is different: we want to solve it *mechanically*, by applying the fundamental rules of inference (§ 8.4) and the shortcut rules (§ 9) derived from them. No intuitive arguments. Our purpose is different because of two main reasons:

- We want to be able to implement or encode the procedure algorithmically in an AI agent.
- We generally cannot ground inferences on intuition. Intuition is shaky ground, and hopeless in data-science problems involving millions of data with thousands of numbers in abstract spaces of thousands of dimensions. To solve such complex problems we need to use a more mechanical procedure, a procedure *mathematically guaranteed* to be self-consistent. That's the probability calculus. Intuition is only useful for arriving at a method which we

can eventually prove, by mathematical and logical means, to be correct; or for approximately explaining a method that we already know, again by mathematical and logical means, to be correct.

Misleading intuition in high dimensions

As an example of our intuition can be completely astray in problems involving many data dimensions, consider the following fact.
Take a one-dimensional Gaussian distribution of probability. You probably know that the probability that a data point is within three standard deviations from the peak is approximately 99.73%. If we take a two-dimensional (symmetric) Gaussian distribution, the probability that a data point (two real numbers) is within three standard deviations from the peak is 98.89%, slightly less than the one-dimensional case. For a three-dimensional Gaussian, the analogous probability is 97.07%, slightly smaller yet.
Now try to answer this question: for a *100-dimensional* Gaussian, what is the probability that a data point is within three standard deviations from the peak? The answer is $(1.83 \cdot 10^{-32})\%$. This probability is so small that you would never observe a data point within three standard deviations from the peak, even if you checked one data point every second for the same duration as the present age of the universe – which is "only" around $4 \cdot 10^{17}$ seconds.

It is instructive, however, if you also check what your intuition told you about the problem:

Exercise

Examine what your intuition tells you the answer should be, without spending too much time thinking, just as if you were on the game show. Examine which kind of heuristics your intuition uses. If you already know the solution to this puzzle, try to remember what your intuition told you the first time you faced it. Keep your observations in mind for later on.

For the extra curious

For further examples of how our intuition leads us astray in high dimensions see

- Counterintuitive Properties of High Dimensional Space
- Exercise 2.20 (and its solution) in *Information Theory, Inference, and Learning Algorithms*

10.2 Which agent? whose knowledge?

A sentence can be assigned different probabilities by different agents having different background information, although in some cases different background information can still lead to numerically equal probabilities.

In the present case, who's the agent solving the inference problem? And what background information does it have?

From the problem statement it sounds like you (on the show) are the agent. But we can imagine that you have programmed an AI agent having your same background information, and ready to make the decision for you.

We must agree on which background information H to give to this agent. Let's define H as the knowledge you have right *before* picking door 1. We make this choice so that we can add your door pick as additional information.

10.3 Define the atomic sentences relevant to the problem

The following sentences seem sufficient:

H := [the background knowledge discussed in the previous section]

$car1$:= 'The car is behind door 1'

$you1$:= 'You initially pick door 1'

$host2$:= 'The host opens door 2'

and similarly for the other door numbers

We could have used other symbols for the sentences, for instance "C_1" instead of "$car1$". The specific symbol choice doesn't matter. We could also have stated the sentences slightly differently, for instance "You choose door 1 at the beginning of the game". What's important is that we understand and agree on the meaning of the atomic sentences above.

10.4 Specify the desired inference

We want the probabilities of the sentences $car1$, $car2$, $car3$, given the knowledge that you picked door 1 ($you1$), that the host opened door 2 ($host2$), and the remaining background knowledge (H). So in symbols we want the values of the following probabilities:

$$\begin{aligned}&\mathrm{P}(car1 \mid host2\,, you1\,, H)\\ &\mathrm{P}(car2 \mid host2\,, you1\,, H)\\ &\mathrm{P}(car3 \mid host2\,, you1\,, H)\end{aligned}$$

You may object: "but we already know that there's no car behind door 2, the one opened by the host; so that probability is 0%". That's correct, but how did you arrive at that probability value? Remember our goal: to solve this inference *mechanically*. Your intuitive probability must therefore either appear as an initial probability, or be derived via the inference rules. No intuitive shortcuts.

10.5 Specify all initial probabilities

As discussed in § 5.2, any inference – logical or uncertain – can only be derived from other inferences, or taken for granted as a starting point ("initial probability", or "axiom" in logic). The only inferences that don't need any initial probabilities are tautologies. We must explicitly write down the initial probabilities implicit in the present inference problem:

- The car is for sure behind one of the three doors, and cannot be behind more than one door:

$$\mathrm{P}(car1 \vee car2 \vee car3 \mid H) = 1$$

$$\mathrm{P}(car1\,, car2 \mid H) = \mathrm{P}(car1\,, car3 \mid H) = \mathrm{P}(car2\,, car3 \mid H) = 0$$

Remember from the shortcut rule for the permanence of truth and falsity (§ 9.1) that the 1 and 0 probabilities

above do not change if we `and` additional information to H.

- The host cannot open the door you picked or the door with the car. This translates in several initial probabilities. Here are some:

$$\mathrm{P}(host2 \mid car2\,, you1\,, H) = 0$$

$$\mathrm{P}(host1 \mid car3\,, you1\,, H) = \mathrm{P}(host3 \mid car3\,, you1\,, H) = 0$$

- The host must open one door, and cannot open more than one door:

$$\mathrm{P}(host1 \vee host2 \vee host3 \mid H) = 1$$

$$\mathrm{P}(host1\,, host2 \mid H) = \mathrm{P}(host1\,, host3 \mid H) = \mathrm{P}(host2\,, host3 \mid H) = 0$$

The probabilities above are all quite clear from the description of the puzzle. But implicit in that description are some more probabilities that will be needed in our inference. The values of these probabilities can be more open to debate, because the problem, as stated, provides ambiguous information. You shall later explore possible alternative values for these probabilities.

- It is equally probable that the car is behind any of the three doors, and your initial pick doesn't change this uncertainty:

$$\mathrm{P}(car1 \mid H) = \mathrm{P}(car1 \mid you1\,, H) = 1/3$$
$$\mathrm{P}(car2 \mid H) = \mathrm{P}(car2 \mid you1\,, H) = 1/3$$
$$\mathrm{P}(car3 \mid H) = \mathrm{P}(car3 \mid you1\,, H) = 1/3$$

  Remember that a probability is not a physical property. We aren't saying that the car should appear behind each door with a given frequency, or something similar. The values $1/3$ are simply saying that in the present situation you have no reason to *believe* the car to be behind one specific door more than behind another.

- If the host can choose between two doors (because the car is behind the door you picked initially), we are equally uncertain about the choice:

$$\mathrm{P}(\mathit{host}2 \,|\, \mathit{car}1, \mathit{you}1, H) = \mathrm{P}(\mathit{host}3 \,|\, \mathit{car}1, \mathit{you}1, H) = 1/2$$

This probability could be analysed into further hypotheses. Maybe the host, out of laziness, could more probably open the door that's closer. But from the problem it isn't fully clear which one is closer. The host could also more probably open the door that's further from the one you choose. The host could have a predetermined scheme on which door to open. The hypotheses are endless. We can imagine some hypotheses that make $\mathit{host}2$ more probable, and some that make $\mathit{host}3$ more probable, conditional on $\mathit{you}1 \wedge \mathit{car}1 \wedge H$. The probability of 50% seems like a good compromise. You shall later examine the effects of changing this probability.

10.5.1 Some peculiar probabilities

We defined the background knowledge H as the one you have right *before* choosing door 1. In this way the sentence $\mathit{you}1$, expressing your door pick, can be added as additional information: $\mathit{you}1 \wedge H$.

It is legitimate to ask: what is the probability that you pick door 1, given only the background information H:

$$\mathrm{P}(\mathit{you}1 \,|\, H) \ ?$$

To answer this question we would need to specify H more in detail. It is possible, for instance, that you planned to pick door 1 already the day before. In this case we would have $\mathrm{P}(\mathit{you}1 \,|\, H) = 1$ or very nearly so. Or you could pick door 1 right on the spot, with no clear conscious thought process behind your choice. In this case we would have $\mathrm{P}(\mathit{you}1 \,|\, H) = 1/3$ or a similar value.

Luckily in the present problem these probabilities are not needed. If they are used, their numerical values turn out not to matter: they will "cancel out" of the computation.

> ⚠ Silly literature
>
> Some texts on probability say that if you have decided something and therefore know for certain it in advance, then the probability of that something is undefined "because it is not random". Obviously this is nonsense. If you already know something, then the probability of that something is well-defined and its value is 100% – or something short of this value, if you want to make allowance for the occurrence of unplanned events.

10.6 Solution

Let's try first to calculate $\mathrm{P}(\mathit{car}1 \mid \mathit{host}2\,, \mathit{you}1\,, H)$, that is, the probability that the car is behind the door you picked.

Seeing that we have several initial probabilities of the "$\mathrm{P}(\mathit{host} \mid \mathit{you}\,, \mathit{car}\,, H)$" form, we can use Bayes's theorem together with the "extension of the conversation" (§ 9.5) to swap the positions of "car" and "host" sentences between proposal and conditional. In the present case the exhaustive and mutually exclusive sentences are $\mathit{car}1$, $\mathit{car}2$, $\mathit{car}3$:

$$\begin{aligned}
&\mathrm{P}(\mathit{car}1 \mid \mathit{host}2\,, \mathit{you}1\,, H) \\
&\quad = \frac{\mathrm{P}(\mathit{host}2 \mid \mathit{car}1\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}1 \mid \mathit{you}1\,, H)}{\left[\begin{array}{c} \mathrm{P}(\mathit{host}2 \mid \mathit{car}1\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}1 \mid \mathit{you}1\,, H) + \\ \mathrm{P}(\mathit{host}2 \mid \mathit{car}2\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}2 \mid \mathit{you}1\,, H) + \\ \mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}3 \mid \mathit{you}1\,, H) \end{array}\right]} \\
&\quad = \ldots
\end{aligned}$$

All probabilities in green are initial probabilities discussed in the previous steps. Let's substitute their values:

$$= \frac{1/2 \cdot 1/3}{\left[\begin{array}{c} 1/2 \cdot 1/3 + \\ 0 \cdot 1/3 + \\ \mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \cdot 1/3 \end{array}\right]}$$

$$= \frac{1/6}{1/6 + \mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \cdot 1/3}$$

$$= \ldots$$

All that's left is to find $\mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H)$. It's intuitively clear that this probability is 100%, because the host is forced to choose door 2 if you picked door 1 and the car is behind door 3. But our purpose is to make a fully mechanical derivation, starting from the initial probabilities only. We can find this probability by applying the `or`-rule and the `and`-rule to the probabilities that the host opens at least one door and cannot open more than one:

$$\begin{aligned}
&\mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\quad = \mathrm{P}(\mathit{host}2 \vee \mathit{host}1 \vee \mathit{host}3 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\qquad - \mathrm{P}(\mathit{host}1 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\qquad - \mathrm{P}(\mathit{host}3 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\qquad + \mathrm{P}(\mathit{host}1\,, \mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\qquad + \mathrm{P}(\mathit{host}1\,, \mathit{host}3 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\qquad + \mathrm{P}(\mathit{host}2\,, \mathit{host}3 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\qquad - \mathrm{P}(\mathit{host}1\,, \mathit{host}2\,, \mathit{host}3 \mid \mathit{car}3\,, \mathit{you}1\,, H) \\
&\quad = 1 - 0 - 0 + 0 + 0 + 0 - 0 = 1
\end{aligned}$$

as expected.

Finally, using this probability in our previous calculation we find

$$
\begin{aligned}
&\mathrm{P}(\mathit{car}1 \mid \mathit{host}2\,, \mathit{you}1\,, H) \\
&\qquad = \frac{1/6}{1/6 + \mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \cdot 1/3} \\
&\qquad = \frac{1/6}{1/6 + 1 \cdot 1/3} = \frac{1/6}{3/6} = \frac{1}{3}
\end{aligned}
$$

that is, there's a $1/3$ probability that the car is behind the door we picked!

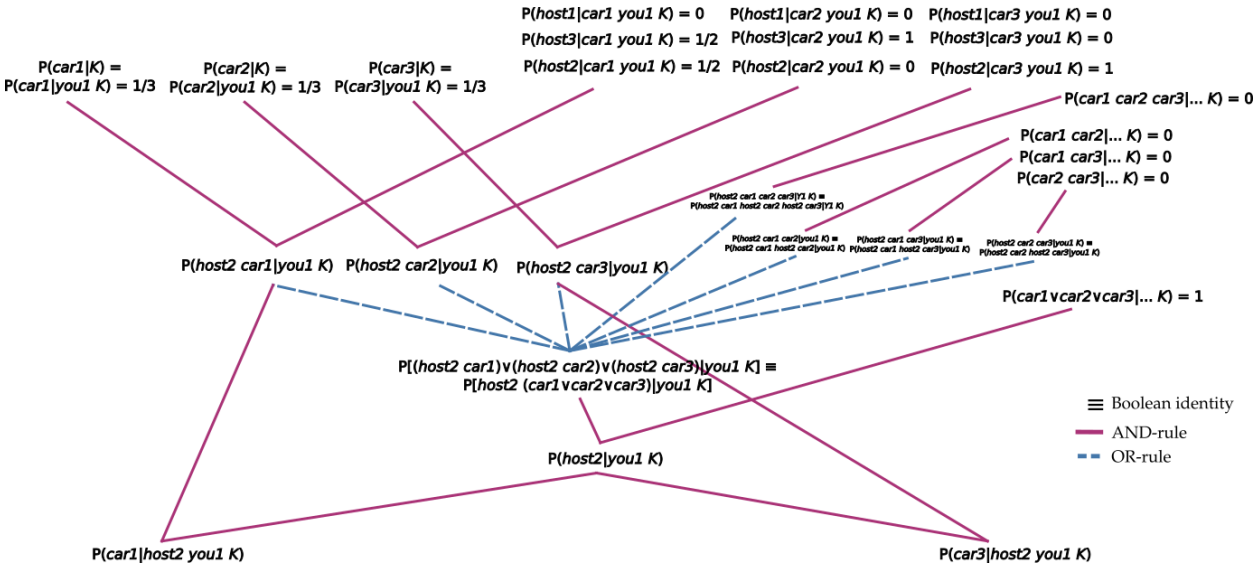
What about door 3, that is, the probability $\mathrm{P}(\mathit{car}1 \mid \mathit{host}2\,, \mathit{you}1\,, H)$? Also in this case we can use Bayes's theorem with the extension of the conversation. The calculation is immediate, because we have already calculated all the relevant pieces:

$$
\begin{aligned}
&\mathrm{P}(\mathit{car}3 \mid \mathit{host}2\,, \mathit{you}1\,, H) \\
&\qquad = \frac{\mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}3 \mid \mathit{you}1\,, H)}{\left[\begin{array}{c} \mathrm{P}(\mathit{host}2 \mid \mathit{car}1\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}1 \mid \mathit{you}1\,, H) + \\ \mathrm{P}(\mathit{host}2 \mid \mathit{car}2\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}2 \mid \mathit{you}1\,, H) + \\ \mathrm{P}(\mathit{host}2 \mid \mathit{car}3\,, \mathit{you}1\,, H) \cdot \mathrm{P}(\mathit{car}3 \mid \mathit{you}1\,, H) \end{array}\right]} \\
&\qquad = \frac{1 \cdot 1/3}{\left[\begin{array}{c} 1/2 \cdot 1/3 + \\ 0 \cdot 1/3 + \\ 1 \cdot 1/3 \end{array}\right]} \\
&\qquad = \frac{1/3}{1/2} = \frac{2}{3}
\end{aligned}
$$

that is, there's a $2/3$ probability that the car is behind door 3. If we'd like to win the car, then we should switch doors.

> **Exercise**
>
> Perform a similar calculation to find $\mathrm{P}(\mathit{car}2 \mid \mathit{host}2\,, \mathit{you}1\,, H)$

Note that we found these probabilities, and solved the Monty Hall problem, just by applying the fundamental rules of inference (§ 8.4), specifically the `and`-rule and `or`-rule, and the Boolean-algebra shortcut rules (§ 9), starting from given probabilities. Here is a depiction of how the fundamental and the shortcut rules connect the initial probabilities, at the top, to the final ones, at the bottom:



10.7 Remarks on the use of Bayes's theorem

You notice that at several points our calculations could have taken a different path. For instance, in order to find $\mathrm{P}(car1 \mid host2\,, you1\,, H)$ we applied Bayes's theorem to swap the sentences $car1$ and $host2$ in their proposal and conditional positions. Couldn't we have swapped $car1$ and $host2 \wedge you1$ instead? That is, couldn't we have made a calculation starting with

$$\mathrm{P}(car1|host2,you1,H) = \frac{\mathrm{P}(host2\,,you1\,|\,car1\,,H)\cdot\mathrm{P}(car1\,|\,H)}{\ldots} \quad ?$$

after all, this is also a legitimate application of Bayes's theorem.

The answer is: yes, we could have, **and the final result would have been the same.** The self-consistency of the probability calculus guarantees that there are no "wrong steps", as long as every step is an application of one of the four fundamental rules (or of their shortcuts). The worst that can happen is that we take a longer route – but to exactly the same result. In fact it's possible that there's a shorter calculation route to arrive at the probabilities that we found in the previous section. But it doesn't matter, because it would lead to the same result that we found.

10.8 Sensitivity analysis

In § 10.5 we briefly discussed possible interpretations or variations of the Monty Hall problem, for which the probability that the host chooses among the available doors 2 and 3 (if the car is behind the door you picked) is different from 50%.

When we want to know how an initial probability value can affect the final probabilities, we can leave its value as a variable, and check how the final probabilities change as we change this variable. This procedure is often called **sensitivity analysis**. Try to do a sensitivity analysis for the Monty Hall problem:

> Exercise
>
> Instead of assuming
>
> $\mathrm{P}(host2\,|\,car1\,,you1\,,H) = \mathrm{P}(host3\,|\,car1\,,you1\,,H) = 1/2$
>
> assign a generic variable value p

$\mathrm{P}(\mathit{host}2|\mathit{car}1,\mathit{you}1,H) = p \qquad \mathrm{P}(\mathit{host}3|\mathit{car}1,\mathit{you}1,H) = 1-p$

where p could be any value between 0 and 1.

- Calculate $\mathrm{P}(\mathit{car}1\,|\,\mathit{host}2,\mathit{you}1,H)$ as was done in the previous sections, but keeping p as a generic variable. This way you'll find a probability $\mathrm{P}(\mathit{car}1 \mid \mathit{host}2\,, \mathit{you}1\,, H)$ that depends numerically on p; it could be considered as a function of p.
- Plot how the value of $\mathrm{P}(\mathit{car}1 \mid \mathit{host}2\,, \mathit{you}1\,, H)$ depends on p, as the latter ranges from 0 to 1.
- For which range of values of p is it convenient to switch door, that is, $\mathrm{P}(\mathit{car}1\,|\,\mathit{host}2,\mathit{you}1,H) < 1/2$?
- Imagine and describe alternative scenarios or background information that would lead to values of p different from 0.5.

10.9 Variations and further exercises

Exercise: other variations

- In § 10.2 we decided that the agent in this inference was you, with the knowledge H right before you picked door 1. Try to change the agent: do you arrive at different probabilities?
  - Consider a person in the audience, right before you picked door 1, as the agent, and re-solve the problem, adjusting all initial probabilities as needed.
  - Consider the *host* as the agent, right before you picked door 1, and re-solve the problem, adjusting all initial probabilities as needed. Note that the host knows for certain where the car is, so you need to provide this additional, secret in-

formation. Consider the cases where the car is behind door 1 and behind door 3.

- Suppose a friend of yours, backstage, gave you partial information about the location of the car (you cheater!), which makes you believe that the car should be closer to door 1. Assign the probabilities

$$\begin{aligned}
\mathrm{P}(\mathit{car}1 \mid H') &= \mathrm{P}(\mathit{car}1 \mid \mathit{you}1\,, H') = 1/3 + q \\
\mathrm{P}(\mathit{car}2 \mid H') &= \mathrm{P}(\mathit{car}2 \mid \mathit{you}1\,, H') = 1/3 \\
\mathrm{P}(\mathit{car}3 \mid H') &= \mathrm{P}(\mathit{car}3 \mid \mathit{you}1\,, H') = 1/3 - q
\end{aligned}$$

with $0 \leq q \leq 1/3$ (this background information is different from the previous one, so we denote it H'). Re-solve the problem keeping the variable q, and find if there's any value for q for which it's best to keep door 1.

Exercise: making decisions

In this chapter we only solved the *inference* problem for the Monty Hall scenario. We calculated the probabilities of various outcomes. But no decision has been made yet.

- Assign utilities to winning the car or winning the goat from the point of view of an agent who values the car more. The available decisions are, of course, "keep door 1" vs "switch to door 3". Then solve the decision-making problem according to the procedure of § 3.3. What's the optimal decision?
- Now assign utilities from the point of view of an agent who values the *goat* more than the car. Then

solve the decision-making problem according to the usual procedure. What's the optimal decision?

Exercise: the Sleeping Beauty problem

Take a look at the inference problem presented in this video:
https://vimeo.com/893102563?share=copy
and try to solve it, not using intuition, but using the mechanical procedure and steps as in the Monty Hall solution above.
Note that the video asks "What do you believe is the probability that the coin came up heads?". Since probability and degree of belief are the same thing, that is like asking "What do you believe is your belief that the coin came up heads?" which is a redundant or quirky question. Instead, simply answer the question "What is your degree of belief (that is, probability) that the coin came up heads?".

11 Second connection with machine learning

In these first chapters we have been developing notions and methods about agents that draw inferences and make decisions, sentences expressing facts and information, and probabilities expressing uncertainty and certainty. Let's draw some first qualitative connections between these notions and notions typically used in machine learning.

A machine-learning algorithm is usually presented in textbooks as something that first "learns" from some training data, and thereafter performs some kind of task – typically it yields a response or outcome, for example a label, of some kind. More precisely, the training data are instances or examples of the task that the algorithm is expected to perform. These instances have a special status because their details are fully known, whereas new instances, where the algorithm will be applied, have some uncertain aspects. A new instance typically has an ideal or optimal outcome, for example "choosing the correct label", but this outcome is unknown beforehand. The response given by the algorithm in new instances depends on the algorithm's internal architecture and parameters (for brevity we shall just use "architecture" to mean both).

Let's try to rephrase this description from the point of view of the previous chapters. A machine-learning algorithm is given known pieces of information (the training data), and then forms some kind of connection with a new piece of information of similar kind (the outcome in a new application) that was not known beforehand. The connection depends on the algorithm's architecture.

11.1 "Learning" and "output" from the point of view of inference & decision

The remarks above reveal similarities with what an agent does when drawing an inference: it uses known pieces of information, expressed by sentences $D_1, D_2, \ldots, D_N$, together with some background or built-in information I, in order to calculate the probability of a new piece of information of a similar kind, expressed by a sentence D_{N+1}:

$$\mathrm{P}(D_{N+1} \mid D_N \wedge \cdots \wedge D_2 \wedge D_1 \wedge I)$$

We can thus consider a first *tentative* correspondence:

$$\mathrm{P}([\,0ex]D_{N+1} \mid [0ex]D_N \wedge \cdots \wedge D_2 \wedge D_1 \wedge [\,0ex]I)$$

outcome? training data? architecture?

This correspondence seems convincing for architecture and training data: in both cases we're speaking about the use of pre-existing or built-in information, combined with additional information.

But the correspondence is less convincing with regard to the outcome. The "agents" that we have envisioned find the probabilities for several possible "outcomes" or "outputs"; they don't yield only one output. This indicates that there must also be some **decision** involved among the possible outcomes.

We'll return to this tentative connection later.

11.2 Why different outputs?

In the previous chapters we have seen, over and over, what was claimed at the beginning of these lecture notes: that an inference & decision problem has only one optimal solution. Once we specify the utilities and the initial probabilities of the problem, the fundamental rules of inference and the principle of maximal expected utility lead to one unique answer (unless,

of course, there are several optimal ones with equal expected utilities).

Different machine-learning algorithms, trained with the same training data, often give different answers or outputs to the same problem. Where do these differences come from? From the point of view of decision theory there are three possibilities, which don't exclude one another:

- **The initial probabilities given to the algorithms are different**. Since the training data are the same, this means that the **background information built into** one machine-learning algorithm is different from those built into another.

  It is therefore important to *understand what are the built-in background information and initial probabilities* of different machine-learning algorithms. The built-in assumptions of an algorithm must match those of the real problem as closely as possible, in order to avoid sub-optimal or even disastrously wrong answers and outputs.

- **The utilities built into one machine-learning algorithm are different** from those built into another.

  It is therefore also important to *understand what are the built-in utilities* of different machine-learning algorithms. The built-in utilities must also match those of the real problem as closely as possible.

- **The calculations made by the algorithms are approximate**, and different algorithms use different approximations. This means that the algorithms don't arrive at the unique answer determined by decision theory, but to some other answers which may be approximately close to the optimal one – or not!

  It is therefore important to *understand what are the calculation approximations* made by different machine-learning algorithms. Some approximations may be too crude for some real problems, and may again lead to sub-optimal or even disastrously wrong answers and outputs.

11.3 Data pre-processing and the data-processing inequality

"Data pre-processing" is a collective name given to very different operations on data before they are used in some algorithm to solve a decision or inference problem. Some of these operations are often said to be essential for the solution of these problems. This statement in not completely true, and needs qualification.

We can divide pre-processing procedures in roughly three categories:

Inconsistency checks Procedures in this category make sure that the data are what they were intended to be. For instance, if data should consist of the power outputs of several engines, but one datapoint is the physical *weight* of an engine, then that "datapoint" is actually no data at all for the present problem. It's something included by mistake and should be removed. Such procedures are necessary and useful, but they are just consistency checks and do not change the information contained in the *proper* data.

In later chapters we shall say more about some often erroneous procedures, like "tail trimming", that actually remove *proper* data and lead to sub-optimal or completely erroneous solutions.

Formatting These procedures make sure that data are in the correct format to be inputted into the algorithm. They may also include rescaling of numerical values for avoiding numerical overflow or underflow errors during computation. Such procedures are often necessary and useful, but they just change the way data are encoded. They do not actually change the information contained in the data.

"Mutilation" or information-alteration Procedures of this kind *alter the content of data*. For instance, such a

procedure may replace, in a dataset of temperatures, a datapoint having value 20 °C with one having value 25 °C; this is not just a simple rescaling. Procedures of this kind include "de-noising", "de-biasing", "de-trending", "filtering", "dimensionality reduction" and similar ones (often having noble-sounding names). We must state, clearly and strongly, that *within Decision Theory and Probability Theory, such information-altering pre-processing is* ***not*** *necessary , and is in fact* ***detrimental***. This is why we call it "mutilation" here.

It is important that you understand that such data pre-processing is ***not*** *something that one, in principle, has to do in data science. Quite the opposite, in principle we should* ***not*** *do it, because it is a destructive procedure. Such pre-processing is done in order to correct deficiencies of the algorithms currently in use, as discussed below.*

If we build an "optimal predictor machine" that fully operates according to the four rules of inference (§ 8.4) and of maximization of expected utility, then the data fed into this machine *should not be pre-processed with any information-altering procedures*. The reason is that the four fundamental rules automatically take care, in an optimal way, of factors such as noise, bias, systematic errors, redundancy. We briefly discussed this fact in § 9.7 and saw a simple example of how redundancy is accounted for by the four inference rules.

If we have information about noise or other factors affecting the data, then we should include this information in the background information provided to the "optimal predictor machine", rather than altering the data given to it. The reason, in intuitive terms, is that the machine does the adjustments while fully exploring the data themselves; so it can "see" more deeply how to make optimal adjustments given the "inner structure" of the data. In the pre-processing phase – as the prefix "*pre-*" indicates – we don't have the full picture about the data, so any adjustment risks to eliminate actually useful information.

More formally, this is the content of the **data-processing inequality** from information theory:

Data-processing inequality

"No clever manipulation of the data can improve the inferences that can be made from the data"
(*Elements of Information Theory* §2.8)
or, from a complementary point of view:
"Data processing can only destroy information"
(*Information Theory, Inference, and Learning Algorithms* exercise 8.9)

Study reading

- Skim through §2.8 of *Elements of Information Theory*
- Take a look Exercise 8.9 and its solution in *Information Theory, Inference, and Learning Algorithms*

There are two main, partially connected reasons why one performs "mutilating" pre-processing of data:

- The algorithm used is *non-optimal*: it's only using approximations of the four fundamental rules, and therefore cannot remove noise, bias, redundancies, or similar factors in an optimal way, or at all. In this case, pre-processing is an approximate way of correcting the deficiency of the non-optimal algorithm.
- Full optimal processing is *computationally too expensive*. In this case we try to simplify the optimal calculation by doing, in advance and in a cruder, faster way, some of the "cleaning" that the full calculation would otherwise spend time doing in an optimal way.

Part III

Data I

12 Quantities and data types

Motivation for the "Data I" part

In the "Inference I" part we surveyed the four fundamental rules of inference, which determine how an agent's degrees of belief should propagate and be self-consistent. We explored some applications and consequences of the four fundamental rules. The rules can be used with any sentences whatsoever, so their application can be further developed and specialized in a wide spectrum of directions, with applications ranging from robotics to psychology. Each of these possible developments would require by itself a full university course!

We shall now restrict our attention to applications typical of "data science" and machine learning, like classification, forecast, prognosis, hypothesis testing, in situations that involve *quantifiable* and *measurable* phenomena. For this purpose we focus on sentences of particular kinds, which can express such quantification and measurement. In a sense, we develop a specialized "language" for this kind of situations.

Still, since we're nonetheless dealing with sentences (Ch. 6), the probability calculus and its inference rules apply without changes of any kind.

12.1 Quantities

12.1.1 Quantities, values, domains

Most decisions and inferences in engineering and data science involve things or properties of things that we can *measure*. We represent them by mathematical objects of different kinds. These objects have particular mathematical properties and can undergo particular operations.

The different mathematical properties of these things reflect the kind of activities that we can do with them. For instance, colours are represented by particular tuples of numbers. These tuples can be multiplied by some numeric weights and added, to obtain another tuple. This mathematical operation, called "convex combination", represents the fact that colours can be obtained by mixing other colours in different proportions.

$25\%\,\#\mathbf{FF0000} + 75\%\,\#\mathbf{0000FF} = \#\mathbf{4000C0}$

It's difficult to find a general term to denote any instance of such "things" and their mathematical representation. Yet it's convenient if we find one, so we can discuss the general theory without getting bogged down in individual cases. To this purpose we'll borrow the term **quantity** from physics and engineering.

> The definition of "quantity" we are using here is similar to the one having the *maximum specific level* as defined in §1.1 of the *International vocabulary of metrology* by the Joint Committee for Guides in Metrology.
> Using the word "quantity" this way is just a convention between us. Other texts and scientists may use other words – for example "variable", "event", "state". When you read a text or listen to a scientist, try to grasp the general idea *behind* their words.
> As a general term, we prefer the word "quantity" to a word like "variable", because the latter may give the idea of something changing in time, and that may very well *not* be the case (think of the mass of a block of concrete). Same goes with a word like "state", for the opposite reason.

We distinguish between a quantity and its **value**. For instance, a quantity could be:
"The temperature at the point having GPS coordinates 60.3775029, 5.3869233, 643, at time 1895-10-04T10:03:14Z";
and its value could be:
24 °C. To understand the difference between a quantity and its value, you may think of the quantity as a question, and of the value as the answer to that question:

(*quantity*:) "What was the temperature at the point having GPS coordinates 60.3775029, 5.3869233, 643, at time 1895-10-04T10:03:14Z?"
(*value*:) "It was 24 °C."

The distinction between a quantity and its value is important and necessary in inference and decision problems, because an agent may *not* know the value of a particular quantity, while still knowing what the quantity is. In this case the agent can consider every possible value that the quantity could have, and assign a probability to each. The set of possible values is called the **domain** of the quantity. Think of it as the collection of all meaningful answers that could be given to the question. In our temperature example, the domain is the set of all possible temperatures from 0 K and above.

Keep in mind that our definition of quantity is quite general. Here's another example:

- *Quantity*: the image taken by a particular camera at a particular time, represented by a specific collection of numbers (say $128 \times 128 \times 3$ integers between 0 and 255).
- One example *value* is this:  (corresponding to a grid of $128 \times 128 \times 3$ *specific* numbers). Another example value:  .

- *Domain*: the collection of $256^{3\times128\times128} \approx 10^{118\,370}$ possible images (corresponding to the collection of possible grids of numeric values).

Other examples of quantities and their domains:

1. The distance between two objects in the Solar System at a specific Barycentric Coordinate Time. The domain could be, say, all values from 0 m to $6 \cdot 10^{12}$ m (Pluto's average orbital distance).
2. The number of total views of a specific online video (at a specific time), with a domain, say, from 0 to 20 billions.
3. The force on an object at a specific time and place. The domain could be, say, 3D vectors with components in $[-100\,\text{N}, +100\,\text{N}]$.
4. The degree of satisfaction in a customer survey, with five possible values `Not at all satisfied`, `Slightly satisfied`, `Moderately satisfied`, `Very satisfied`, `Extremely satisfied`.
5. The graph representing a particular social network. Individuals are represented by nodes, and different kinds of relationships by directed or undirected links between nodes, possibly with numbers indicating their strength. The domain consists of all possible graphs with, say, 0 to 10 000 nodes and all possible combinations of links and weights between the nodes.
6. The relationship between the input voltage and output current of an electric component. The domain could be all possible continuous *curves* from $[0\,\text{V}, 10\,\text{V}]$ to $[0\,\text{A}, 1\,\text{A}]$. Note that the domain in this case is not made of *numbers*.
7. A 1-minute audio track recorded by a device with a sampling frequency of 48 kHz (that is, 48 000 audio samples per second). The domain could be all possible sequences of 2 880 000 numbers in $[0, 1]$.
8. The subject of an image, with domain of three possible values `cat`, `dog`, `something else`.

9. The roll, pitch, yaw of a rocket at a specific time and place, with domain $(-180°, +180°] \times (-90°, +90°] \times (-180°, +180°]$.

The vague term "data" typically means the values of a collection of quantities.

In these notes we agree that *a quantity has one, and only one, actual value.*

> ⚠ Quantity vs variate or variable
>
> We can consider something that changes with time, or with space, or from individual to individual, or from unit to unit. Then this "something" is not a quantity, according to our present terminology, but a *collection* of quantities: one for each time, or space, or individual. Later we shall call this collection a **variate**, especially when it refers to individuals or unit; or a **variable**, especially when it refers to time.
> For instance, your height at this exact moment is a *quantity*, but your height throughout your life is a *variable*, and the height (at this moment) across all Norwegian people is a *variate*.
> These are just terminological conventions adopted in the present notes. As mentioned before, different scientists often adopt different terms. What matters is not the terms, but that you have a clear understanding of the difference between the two **notions** that we here call "quantity" and "variate".

12.1.2 Notation

We shall denote quantities by italic letters, such as X, or U, or A. The sentences that appear in decision-making and inferences are therefore often of the kind:
"the quantity X was observed to have value x",

where "x" stands for a specific value. This kind of sentences are often abbreviated in the form "$X{=}x$".

❗ Keep in mind our discussion from § 6.3: we must make clear what "=" means. It could mean "observed", "set", "reported", and so on.
❗ Note the subtle difference between X, in italics, and X, in sans-serif. The first denotes a *quantity*, the second denotes a *sentence*. Usually we don't have to worry too much about these symbol differences, because the meaning of the symbol is clear from the context. But just in case, you know the convention.

12.2 Basic types of quantities

As the examples above show, quantities and data come in all sorts, and with various degrees of complexity. There is no clear-cut divide between different sorts of quantities. The same quantity can moreover be viewed and represented in many different ways, depending on the specific context, problem, purpose, and background information.

It is possible, however, to roughly differentiate between a handful of basic **types** of quantities, from which more complex types are built. Here is one kind of differentiation that is useful for inference problems about quantities:

12.2.1 Nominal

A **nominal** or **categorical** quantity has a domain with a discrete and usually finite number of values. The values *are not related by any mathematical property*, and *do not have any specific order*.

This means that when we speak of a nominal quantity, it does not make sense to say, for instance, that one value is "twice" or "1.5 times" another; or that one value is "larger" or "later" than another. Nor does it make sense to "add" two quantities. In

particular, *there is no notion of cumulative probability, quantile, median, average, or standard deviation for a nominal quantity*; these are notions that we'll discuss in Ch. 21.

Examples: the possible breeds of a dog, or the characters of a film.

It is of course possible to represent the values of a nominal quantity with numbers; say `1` for `Dachshund`, `2` for `Labrador`, `3` for `Dalmatian`, and so on. But that doesn't mean that
$\texttt{Dalmatian} - \texttt{Labrador} = \texttt{Labrador} - \texttt{Dachshund}$
just because $3 - 2 = 2 - 1$, or similar nonsense.

12.2.2 Ordinal

An **ordinal** quantity has a domain with a discrete and usually finite number of values. The values *are not related by any mathematical property*, but they *do have a specific order*.

This means that that when we speak of a nominal quantity, it does *not* make sense to say that one value is "twice" or "1.5 times" another, and we can*not* add or subtract two values. But it does make sense to say, for any two values, which one has higher rank, for example "stronger", or "later", or "larger", and similar. Owing to the ordering property, *it does make sense to speak of cumulative probability, quantile, and median of an ordinal quantity*; but *there is no notion of average or standard deviation for an ordinal quantity.*

Example: a pain-intensity scale. A patient can say whether some pain is more severe than another, but it isn't clear what a pain "twice as severe" as another would mean (although there's a lot of research on more precise quantification of pain). Another example: the "strength of friendship" in a social network. We can say that we have a "stronger friendship" with a person than with another; but it doesn't make sense to say that we are "four times stronger friends" with a person than with another.

It is possible to represent the values of an ordinal quantity with numbers which reflect the *order* of the values. But it's important to keep in mind that differences or averages of such

numbers do *not* make sense. For this reason the use of numbers to represent an ordinal quantity can be misleading. A less misleading possibility is to represent ordered values by alphabet letters.

12.2.3 Binary

A **binary** or **dichotomous** quantity has only two possible values. It can be seen as a special case of a nominal or ordinal quantity, but the fact of having only two values lends it some special properties in inference problems. This is why we list it separately.

Obviously it doesn't make much sense to speak of the difference or average of the two values; and their ranking is trivial even if it makes sense.

There's an abundance of examples of binary quantities: yes/no answers, presence/absence of something, and so on.

12.2.4 Interval

An **interval** quantity has a domain that can be discrete or continuous, finite or infinite. The values *do admit some mathematical operations*, at least *convex combination* and *subtraction*. They also admit an ordering.

This means that for such a quantity we can say, at the very least, whether the interval or "distance" between one pair of values is the same, or larger, or smaller than the interval between another pair. For this reason we can also say whether one value is larger than another. We can also take weighted sums of values, called *convex combinations* (keep in mind that simple *addition* of values may be meaningless for some quantities).

Owing to these mathematical properties, *it does make sense to speak of the cumulative probability, quantile, median, and also average and standard deviation for an interval quantity.*

The number of electronic components produced in a year by an assembly line is an example of a discrete interval quantity. The

power output of a nuclear plant at a given time is an example of a continuous interval quantity.

It is also possible to speak of *ratio* quantities, which are a special case of interval quantities, but we won't have use of this distinction in the present notes.

12.2.5 How to decide the basic type of a quantity?

To attribute a basic type to a quantity we must ultimately *check how that quantity is defined, obtained, and used*. In some cases the values of the quantity may give some clue. For example, if we see values "2.74", "8.23", "3.01", then the quantity is probably of the interval type. But if we see values "1", "2", "3", then it's unclear whether the quantity is interval, ordinal, nominal, or maybe of yet some other type.

The type of a quantity also depends on its use in the specific problem. A quantity of a more complex type can be treated as a simpler type if needed. For instance, the response time of some device is in principle an interval quantity: it could be measured, say, in seconds, as precisely as we want. But in a specific situation we could simply label its values as `slow`, `medium`, `fast`, thus turning it into an ordinal quantity.

@@ TODO: add examples for image spaces

Exercises

- For each example at the beginning of the present section, assess whether that quantity can be considered as being of a basic type, and which type.
- For each basic type discussed above, find two more concrete examples of that type of quantity.

For the extra curious

On the theory of scales of measurement

12.3 Other attributes of basic types

It is useful to consider other basic aspects of quantities that are somewhat transversal to "type". These aspects are also important when drawing inferences.

12.3.1 Discrete vs continuous

Nominal and ordinal quantities have discrete domains. The domain of an interval quantity can be discrete or continuous. Ultimately all domains are discrete, since we cannot observe, measure, report, or store values with infinite precision. In a modern computer, for example, a real number can "only" take on $2^{64} \approx 20\,000\,000\,000\,000\,000\,000$ possible values. But in practice, in many situations the available precision is so high that we can consider the quantity as continuous for all practical purposes. This can be convenient also because we can then use the mathematics of continuous sets – derivation, integration, and so on – to our advantage.

12.3.2 Bounded vs unbounded

Ordinal and interval quantities may have domains with no minimum value, or no maximum value, or neither. Typical terms for these situations are *lower-* or *upper-bounded*, or *left-* or *right-bounded*, and analogously with *unbounded*; or similar terms.

Whether to treat a quantity domain as bounded or unbounded depends on the quantity, the specific problem, and the computational resources. For example, the number of times a link on a webpage has been clicked can in principle be (upper-)unbounded. Another example is the distance between two objects: we can consider it unbounded, but in concrete problems might be bounded, say, by the size of a laboratory, or by Earth's circumference, or the Solar System's extension, and so on.

Exercises

- If you had to set a maximum number of times a web link can be clicked, what number would you choose? Try to find a reasonable number, considering factors such as how fast a person can repeatedly click on a link, how long a website (or the Earth?) can last, and how many people can live during such an extent of time.
- What about the age of a person? What upper bound would you set, if you had to treat it as a bounded quantity?

12.3.3 Finite vs infinite

The domain of a discrete quantity can consist of a finite or an infinite number (at least in theory) of possible values. The domain of a continuous quantity always has an infinite number of values. Note that a domain can be infinite and yet bounded: consider the numbers in the range $[0, 1]$.

Whether to treat a domain as finite or infinite depends on the quantity, the specific problem, and the computational resources. For example, the intensity of a base colour in a pixel of a particular image might really take on 256 discrete steps between 0 and 1: $0, 0.0039215686, 0.0078431373, \ldots, 1$. But in some situations we can treat this domain as practically infinite, with any possible value between 0 and 1.

12.3.4 Rounded

A continuous interval quantity may be rounded, because of the way it's measured. In this case the quantity could be considered discrete rather than continuous.

For instance, the famous *Iris* dataset consists of several lengths – continuous interval quantities – of parts of flowers. All values are rounded to the millimetre, even if in reality the lengths could of course have intermediate values. The age of a person is

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6	2·8	4·9	2·0
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7	2·8	6·7	2·0
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3	2·7	4·9	1·8
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7	3·3	5·7	2·1
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2	3·2	6·0	1·8
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2	2·8	4·8	1·8
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1	3·0	4·9	1·8
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4	2·8	5·6	2·1
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2	3·0	5·8	1·6
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4	2·8	6·1	1·9
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9	3·8	6·4	2·0
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4	2·8	5·6	2·2
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3	2·8	5·1	1·5
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1	2·6	5·6	1·4
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7	3·0	6·1	2·3
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3	3·4	5·6	2·4
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4	3·1	5·5	1·8
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0	3·0	4·8	1·8
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9	3·1	5·4	2·1
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7	3·1	5·6	2·4
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9	3·1	5·1	2·3
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8	2·7	5·1	1·9
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8	3·2	5·9	2·3
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7	3·3	5·7	2·5
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7	3·0	5·2	2·3
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3	2·5	5·0	1·9
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5	3·0	5·2	2·0
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2	3·4	5·4	2·3
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9	3·0	5·1	1·8

The *Iris* dataset from its original

another frequent example of an in-principle continuous quantity which is often rounded, say to the year or to the month.

Rounding can impact the way we do inferences about such a quantity. In some situations, rounding can lead to quantities with different unrounded values to take on identical rounded ones.

12.3.5 Censored

The measurement procedure of a quantity may have an artificial lower or upper bound. A clinical thermometer, for instance, could have a maximum reading of 45 °C. If we measure with it the temperature of a 50 °C-hot body, we'll read "45 °C", not the real temperature.

A quantity with this characteristic is called **censored**, more specifically *left-censored* or *right-censored* when there's only one artificial bound. The bound is called the *censoring value*.

A censoring value denotes an actual value that could also be greater or less. This is important when we draw inferences about this kind of quantities.

Exercises

Explore datasets in a database such as the UC Irvine Machine Learning Repository:

- Read the description of the quantities listed in the dataset (sometimes in a `readme` file included with the dataset download)
- Analyse the values of some of the quantities in the dataset: check if they can be considered continuous, discrete, or rounded; bounded or unbounded; uncensored or censored; and so on.

12.4 "True" vs "measured" values

A difference is often drawn, especially in physics and engineering, between the "true" value of a quantity and the value "observed" or "measured" with a particular measuring instrument. What's the difference? and how is the "true" value defined? There actually are deep philosophical questions and choices underlying this distinction, and it would take a whole university course to do them justice.

Intuitively we define the "true" value as the value that would be measured with an instrument that is perfectly calibrated and as precise as theoretically possible. If we make a distinction between such value and the currently measured value then we're implying that the current measurement is made with a less precise instrument, and that the "true" and "measured" values could be different.

> For the extra curious
>
> *The Logic of Modern Physics*

In some circumstances this distinction is unimportant and an agent can use the "measured" value without worries, and consider it as the "true" one. Typically this is the case when the possible discrepancy between measured and true value is enough small to have no consequences. In other circumstances the discrepancy is important: slightly different values might lead to quite different consequences. In such circumstances it is then necessary for the agent to try to *infer* – using the probability calculus – the true value, using the measured one as "data" or "evidence". Said otherwise, the agent doesn't use the measured value directly, but only as an intermediate step to guess the true value. The latter, in turn, can be used for further inferences.

From the point of view of inference and decision-making, the distinction between "true" and "measured" value doesn't lead to anything methodologically new. It just means that an agent has to do a chain of inferences instead of just one, using the four rules of inference as usual. This situation often requires the definition of two distinct quantities, the "true" and the "observed", which can have slightly different domains. For instance, we could have a voltage V_{obs} measured with rounding to $1\,\text{V}$ and therefore with discrete domain $\{10\,\text{V}, 11\,\text{V}, 12\,\text{V}, ...\}$; but we

need the "true" voltage V_{true} with a precision of at least 0.01 V, so this latter quantity could have a continuous domain.

In solving data-science and engineering problems it's important to make clear whether a particular quantity value can be considered "true" and used as-is, or only "observed" with insufficient precision and used as data to infer the true value.

12.5 Importance of metadata for inference and decision

The characteristics of quantities and domains that we have discussed so far are examples of **metadata**. As the name implies, *meta*data is information that typically *cannot* be found in the data.

As a simple example, consider this collection of numerical data values:
8 2 6 19 1 5 4 19 1 8 12 3 1 2 17
and suppose that some machine-learning algorithm has to generate a new number "similar" to the ones above, or guess what a new one could be. Consider the following possible guesses; would they be acceptable?:

- **13**. Note that this number does not appear among the values above. Could it be that it is impossible for some particular reason? For example, this value is omitted from the seat numbers of some airlines or street addresses, because of triskaidekaphobia.

- **21**. This guess would be impossible if the reported values are rolls of a 20-sided die, typical of fantasy roleplay games. But if the values are, say, the ages of some people, then **21** could be an admissible guess.

- **3.5**. The reported values are all integers. But they could be rounded values of say, temperature readings or ages. We don't know whether a more precise, non-rounded guess such as **3.5** could be acceptable.

- **-2**. If the reported values are people's ages or objects' weights, then a negative value would be impossible. But if the values were temperatures in degrees Celsius, then the guess **-2** could be acceptable.

The collection of values does *not* allow us to determine which of the possible scenarios above applies.

A simple piece of metadata can actually correspond to an infinite amount of datapoints. Even if we have a collection of one million positive numbers, they still don't tell us whether negative values are impossible or not. When we are given the information that the domain of the relevant quantity is positive, this effectively corresponds to knowing that all future data – possibly an infinity – will not be negative.

An AI agent or machine-learning algorithm will therefore make better guesses and better decisions if it is given full metadata, besides "training" data. For this reason metadata are extremely important for inference and decision-making, and an optimal agent should make use of metadata.

Exercises

Examine the adult-income dataset at the UC Irvine Machine Learning Repository. This set contains more than 30 000 datapoints.
Take a look at the quantity `native_country` (what type of variate is this?).

- Do you see the values `Norway` or `Finland` or `Sweden` listed among these values?
- Does this mean that in the adult USA population there are no people coming from these three countries?
- Should an AI agent or machine-learning algorithm exclude these three values from its future guesses?
- What would you do to give the algorithm this metadata information? (In a later chapter we shall see

how to do this in a rigorous way.)

13 Joint quantities and complex data types

Quantities of more complex types can often be viewed and represented as sets (that is, collections) of quantities of basic and possibly different types.

13.1 Joint quantities

A simple collection of quantities of basic types, for instance "age, sex, nationality", usually does not have any new mathematical properties appearing just because we're considering those quantities together. We shall call such a collection a **joint quantity**. Note that a "joint quantity" it is still a quantity, but not a quantity of a basic type.

The values of a joint quantity are just tuples of values of its basic component quantities. Its domain is the Cartesian product of the domains of the basic quantities.

Consider for instance the age, sex[1], and nationality of a particular individual. They can be represented as an interval-continuous quantity A, a binary one S, and a nominal one N. We can join them together to form the joint quantity "(age, sex, nationality)" which can be denoted by (A, S, N). One value of this joint quantity is, for example, $(25\,\mathrm{y}, \texttt{F}, \texttt{Norwegian})$. The domain could be

$$[0, +\infty) \times \{\texttt{F;}, \texttt{M;}\} \times \{\texttt{Afghan;}, \texttt{Albanian;}, \ldots, \texttt{Zimbabwean;}\}$$

[1]We define *sex* by the presence of at least one Y chromosome or not. It is different from *gender*, which involves how a person identifies.

13.1.1 Discreteness, boundedness, continuity

A joint quantity may not be simply characterized as "discrete", or "bounded", or "infinite", and so on. Usually we must specify these characteristics for each of its basic component quantities instead. Sometimes a joint quantity is called, for instance, "continuous" if all its basic components are continuous; but other conventions are also used.

> Exercises
>
> Consider again the examples of § 12.1.1. Do you find any examples of joint quantities?

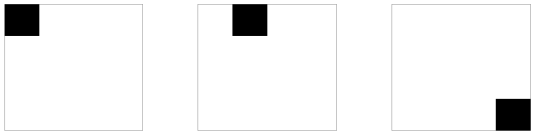
13.2 Complex quantities

We shall call "complex quantity" a quantity that is not of a basic type, nor a collection of quantities of basic type, that is, a joint quantity.

Familiar examples of complex quantities are vectorial quantities from physics and engineering, such as location, velocity, force, torque. Other examples are images, sounds, videos.

Note that a complex quantity may be *represented* as a collection of quantities of basic type. This collection, however, is "more than the sum of its parts", in the sense that it has new mathematical properties that do not apply or do not make sense for the single components.

Consider for example a 4×4 monochrome image, represented as a grid of 16 binary quantities 0 or 1. Three possible values could be these:



We can numerically represent these images as the matrices

$\begin{pmatrix} 1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0 \end{pmatrix}$, $\begin{pmatrix} 0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0 \end{pmatrix}$, $\begin{pmatrix} 0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1 \end{pmatrix}$.

With this representation, this quantity is made to correspond to 16 binary digits, or in other words 16 binary quantities.

From the point of view of the individual binary quantities, these three values are "equally different" from one another: where one of them has grid value 1, the others have 0. But properly considered as images, we can say that the first and the second are somewhat more "similar" or "closer" to each other than the first and the third. This similarity can be represented and quantified by a *metric* over the domain of all such images. This metric involves all basic binary quantities at once; it is a new mathematical property that does not belong to any of the 16 binary quantities individually.

More generally, complex quantities have additional, peculiar properties, represented by mathematical structures, which distinguish them from joint quantities; although there is not a clear separation between the two.

These properties and structures are very important for inference problems, and usually make them computationally very hard. Machine-learning methods are important because they allow us to do approximate inference on these kinds of complex data. The peculiar structures of these data, however, are often also the cause of striking failures of some machine-learning methods, for example the reason why they may classify incorrectly, or why they may classify correctly but for the wrong reason.

Part IV

Inference II

14 Probability distributions

Motivation for the "Inference II" part

In the "Data I" part we developed a language, that is, particular kinds of sentences, to approach inferences and probability calculations typical of data-science and engineering problems.

In the present part we focus on probability calculations that often occur with this kind of sentences and data. We also focus on how to visually represent such probabilities in useful ways.

Always keep in mind that at bottom we're just using the four fundamental rules of inference over and over again – nothing more than that!

14.1 Distribution of probability among values

When an agent is uncertain about the value of a quantity, its uncertainty is expressed and quantified by assigning a degree of belief, conditional on the agent's knowledge, to all the possible cases – all the possible values that could be the true one.

For a temperature measurement, for instance, the cases could be "The temperature is measured to have value 271 K", "The temperature is measured to have value 272 K", and so on up to 275 K. These cases are expressed by mutually exclusive and exhaustive sentences. Denoting the temperature with T, these sentences can be abbreviated as

$T = 271\,\mathrm{K}$, $T = 272\,\mathrm{K}$, $T = 273\,\mathrm{K}$, $T = 274\,\mathrm{K}$, $T = 275\,\mathrm{K}$.

The agent's belief about the quantity is then expressed by the probabilities about these five sentences, conditional on the agent's state of knowledge, which we may denote by the letter I. These probabilities could be, for instance,

$$\begin{aligned}
\mathrm{P}(T{=}271\,\mathrm{K} \mid I) &= 0.04 \\
\mathrm{P}(T{=}272\,\mathrm{K} \mid I) &= 0.10 \\
\mathrm{P}(T{=}273\,\mathrm{K} \mid I) &= 0.18 \\
\mathrm{P}(T{=}274\,\mathrm{K} \mid I) &= 0.28 \\
\mathrm{P}(T{=}275\,\mathrm{K} \mid I) &= 0.40
\end{aligned}$$

Note that they sum up to one:

$$\begin{aligned}
&\mathrm{P}(T{=}271\,\mathrm{K} \mid I) + \cdots + \mathrm{P}(T{=}275\,\mathrm{K} \mid I) \\
&= 0.04 + 0.10 + 0.18 + 0.28 + 0.40 \\
&= 1
\end{aligned}$$

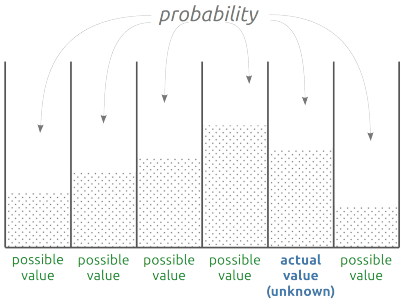
This collection of probabilities is called a **probability distribution**, because we are distributing the probability among the possible alternatives.

What's "distributed"?

The *probability* is distributed among the possible values, as illustrated in the side picture. The quantity cannot be "distributed": it has one, definite value, which is however unknown to the agent.

Exercise

Consider three sentences X_1, X_2, X_3 that are mutually exclusive and exhaustive on conditional I, that is:



$$\mathrm{P}(X_1 \wedge X_2 \mid I) = \mathrm{P}(X_1 \wedge X_3 \mid I) = \mathrm{P}(X_2 \wedge X_3 \mid I) = 0$$
$$\mathrm{P}(X_1 \vee X_2 \vee X_3 \mid I) = 1$$

Prove, using the fundamental rules of inferences and any derived rules from § 8, that we must then have

$$\mathrm{P}(X_1 \mid I) + \mathrm{P}(X_2 \mid I) + \mathrm{P}(X_3 \mid I) = 1$$

Let's see how probability distributions can be represented and visualized for the basic types of quantities discussed in § 12.

We start with probability distributions over discrete domains.

14.2 Discrete probability distributions

14.2.1 Tables and functions

A probability distribution over a discrete domain can obviously be displayed as a table of values and their probabilities. For instance

value	271 K	272 K	273 K	274 K	275 K
probability	0.04	0.10	0.18	0.28	0.40

In the case of ordinal or interval quantities it is sometimes possible to express the probability as a *function* of the value. For instance, the probability distribution above could be summarized by this function of the value t:

$$\mathrm{P}(T{=}t \mid I) = \frac{(t/\mathrm{K} - 269)^2}{90} \quad \text{(rounded to two decimals)}$$

A graphical representation is often helpful to detect features, peculiarities, and even inconsistencies in one or more probability distributions.

14.2.2 Histograms and area-based representations

A probability distribution for a nominal, ordinal, or discrete-interval quantity can be neatly represented by a **histogram**.

The possible values are placed on a line. For an ordinal or interval quantity, the sequence of values on the line should correspond to their natural order. For a nominal quantity the order is irrelevant.

A rectangle is then drawn above each value. The rectangles might be contiguous or not. The bases of the rectangles are all equal, and the *areas* of the rectangles are proportional to the probabilities. Since the bases are equal, this implies that the heights of the rectangles are also proportional to the probabilities.

Such kind of drawing can of course be horizontal, vertical, upside-down, and so on, depending on convenience.

Since the probabilities must sum to one, the total area of the rectangles represents an area equal to 1. So in principle there is no need of writing probability values on some vertical axis, or grid, or similar visual device, because the probability value can be visually read as the ratio of a rectangle area to the total area. An axis or grid can nevertheless be helpful. Alternatively the probabilities can be reported above or below each rectangle.

Nominal quantities do not have any specific order, so their values do not need to be ordered on a line. Other area-based representations, such as pie charts, can also be used for these quantities.

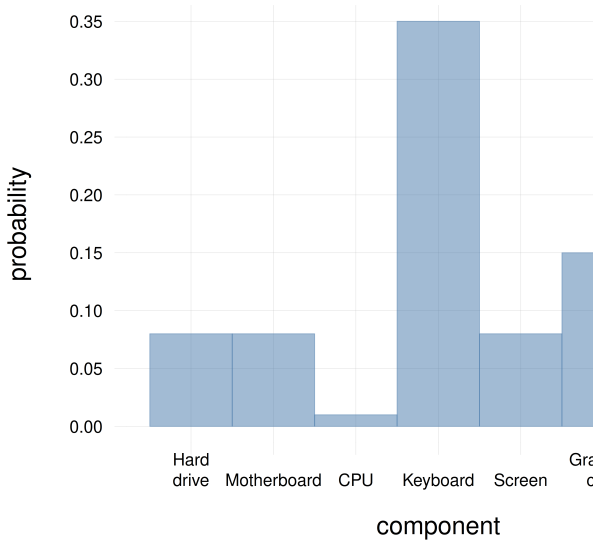


Figure 14.1: Histogram for the probability d
possible component failures

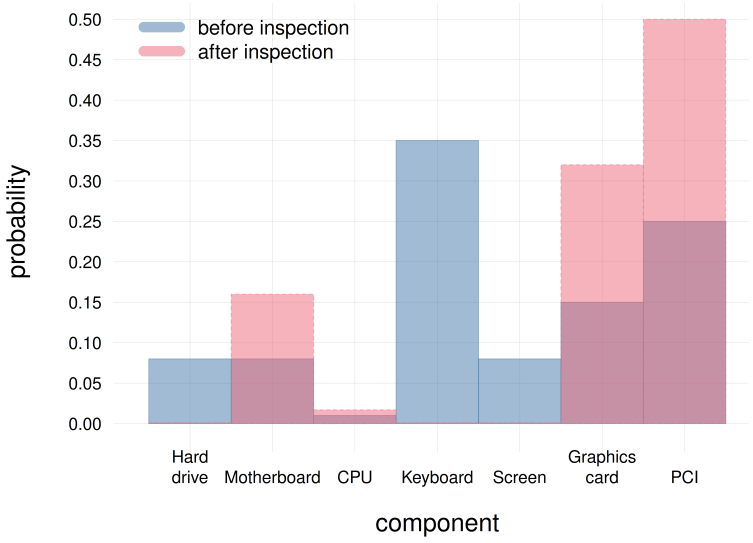
14.2.3 Line-based representations

Histograms give faithful representations of discrete probability distributions. Their graphical bulkiness, however, can be a disadvantage in some situations, for instance when we want to have a clearer idea of how the probability varies across values for ordinal or interval quantities; or when we want to compare several different probability distributions over the same values.

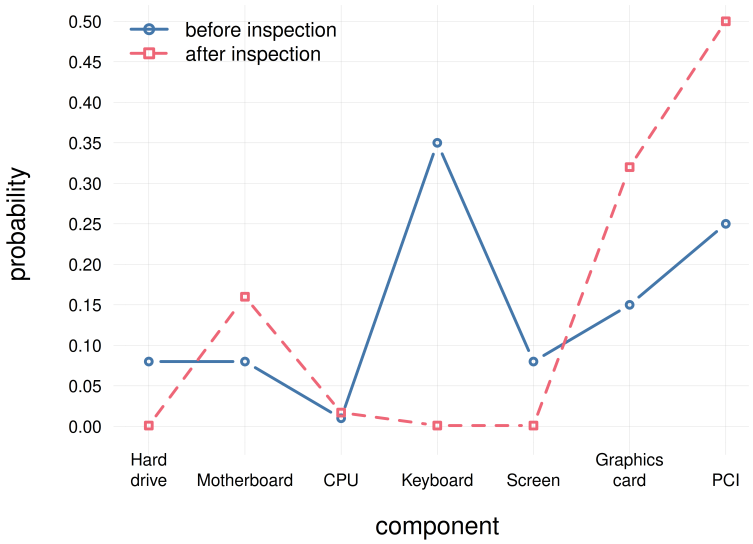
In these cases we can use standard line plots, or variations thereof. Compare the following example.

A technician wonders which component of a laptop failed first (only one can fail at a time), with seven possible alternatives: {`hard-drive;`, `motherboard;`, `CPU;`, `keyboard;`, `screen;`, `graphics-card;`, `PCI;`}. This is a nominal quantity.

Before examining the laptop, the technician's belief about which component failed first is distributed among the seven alternatives as shown by the blue histogram with solid borders. After a first inspection of the laptop, the technician's belief has a new distribution, shown by the red histogram with dashed borders:



It requires some concentration to tell the two probability distributions apart, for example to understand where their peaks are. Let us represent them by two line plots instead: solid blue with circles for the pre-inspection belief distribution, and dashed red with squares for the post-inspection one:



this line plot displays more cleanly the differences between the two distributions. We see that at first the technician most strongly believed the `keyboard` to be the faulty candidate, the second strongest belief being for the `PCI`. After the preliminary inspection, the technician most strongly believes the `PCI` to be the faulty candidate, followed by the `graphics card`.

14.3 Probability densities

Distributions of probability over continuous domains present several counter-intuitive aspects, which essentially arise because we are dealing with uncountable infinities – while often using linguistic expressions that make only sense for countable infinities. Here we follow a practical and realistic approach for working with such distributions.

Consider a quantity X with a continuous domain. When we say that this quantity has some value x we really mean that it has a value somewhere in the range $x-\epsilon/2$ to $x+\epsilon/2$, where the width ϵ is usually extremely small, because we never have infinite precision. For example, for double-precision values stored in a computer, the width[1] must be at least $\epsilon \approx 2 \cdot 10^{-16}$. You

[1] more precisely the relative width

can check indeed that your computer might not distinguish between two numbers that differ in their 16th decimal digit:

```
#### R code
## difference in 15th decimal digit
> 1.234567890123456 == 1.234567890123455
[1] FALSE

## difference in 16th decimal digit
> 1.2345678901234567 == 1.2345678901234566
[1] TRUE
```

The value `1.3` really represents a range between `1.29999999999999982236431605997495353221893310546875` and `1.30000000000000026645352591003756970167160034179687 5`, this range coming from the internal binary representation of `1.3`. Often the width ϵ is much larger than the computer's precision, and comes from the precision with which the value is *experimentally* measured.

When we consider a distribution of probability for a continuous quantity, the probabilities are therefore distributed among such small ranges, not among single values.

Since these ranges are very small, they are also very numerous. But the total probability assigned to all of them must still sum up to 1. This means that each small range receives an extremely small amount of probability. A standard Gaussian distribution for a real-valued quantity, for instance, assigns a probability of approximately $8 \cdot 10^{-17}$, or 0.000 000 000 000 000 08, to a range of width $2 \cdot 10^{-16}$ around the value 0. All other ranges are assigned even smaller probabilities.

In would be impractical to work with such small probabilities. We use **probability densities** instead. As implied by the term "density", a probability density is the amount of probability P assigned to a standard range of width ϵ, *divided* by that width. For example, if the probability assigned to a range of width $\epsilon = 2 \cdot 10^{-16}$ around the value 0 is $P = 7.97885 \cdot 10^{-17}$, then the *probability density* around 0 is

$$\frac{P}{\epsilon} = \frac{7.97885 \cdot 10^{-17}}{2 \cdot 10^{-16}} = 0.398942$$

which is a more convenient number to work with.

Probability densities are convenient because they usually do not depend on the range width ϵ, if it's small enough. Owing to physics reasons, we don't expect a situation where X is between 0.9999999999999999 and 1.0000000000000001 to be very different from one where X is between 1.0000000000000001 and 1.0000000000000003. The probabilities assigned to these two small ranges of width $\epsilon = 2 \cdot 10^{-16}$ will therefore be approximately equal, let's say P each. Now if we use a small range of width ϵ around $X = 1$, the probability is P, and the probability *density* is P/ϵ. If we consider a range of double width $2\,\epsilon$ around $X = 1$, then the probability is $P + P$ instead, but the probability density is still

$$\frac{P+P}{2\,\epsilon} = \frac{1.59577 \cdot 10^{-16}}{4 \cdot 10^{-16}} = 0.398942 \ .$$

As you see, even if we consider a range with double the width as before, the probability density is still the same.

In these notes we'll denote probability densities with a *lowercase* p, with the following notation:

$$\mathrm{p}(X{=}x \mid I) := \frac{\mathrm{P}(\text{“}X \text{ has value between } x-\epsilon/2 \text{ and } x+\epsilon/2\text{'} \mid I)}{\epsilon}$$

This definition works even if we don't specify the exact value of ϵ, as long as it's small enough.

> **Probability densities are not probabilities**
>
> If X is a continuous quantity, the expression "$\mathrm{p}(X{=}2.5 \mid I) = 0.3$" does *not* mean "There is a 0.3 probability that X=2.5". The probability that X=2.5 *exactly* is, if anything, zero.
> That expression instead means "There is a $0.3 \cdot \epsilon$ probability that X is between $2.5 - \epsilon/2$ and $2.5 + \epsilon/2$, for any

ϵ small enough".
In fact, **probability densities can be larger than 1**, because they are obtained by dividing by a number, the width, that is in principle arbitrary. This fact shows that they cannot be probabilities.
It is important not to mix up probabilities and probability *densities*. We shall see later that densities have very different properties, for example with respect to maxima and averages.

A helpful practice (though followed by few texts) is to always write a probability density as

$$\mathrm{p}(X{=}x \mid I)\,\mathrm{d}X$$

where "dX" stands for the width of a small range around x. This notation is also helpful with integrals. Unfortunately it becomes a little cumbersome when we are dealing with more than one quantity.

14.3.1 Physical dimensions and units

In the International System of Units (SI), "Degree of belief" is considered to be a dimensionless quantity, or more precisely a quantity of dimension "1". This is why we don't write units such as "metres" (m), "kilograms" (kg) or similar together with a probability value.[2]

A probability *density*, however, is defined as the ratio of a probability amount and an interval ϵ of some quantity. This latter quantity might well have physical dimensions, say "metres" m. Then the ratio, which is the probability density, has dimensions 1/m. So *probability densities in general have physical dimensions.*

As another example, suppose that an agent with background knowledge I assigns a degree of belief 0.00012 to an interval of

[2] See also the material at the International Bureau of Weights and Measures (BIPM)

temperature of width 0.0001 °C, around the temperature $T = 20\,°\mathrm{C}$. Then the probability *density* at 20 °C is equal to

$$\mathrm{p}(T{=}20°\mathrm{C} \mid I) = \frac{0.00012}{0.0001\,°\mathrm{C}} = 1.2\,°\mathrm{C}^{-1}$$

It is an error to report probability densities without their correct physical units. In fact, keeping track of these units is often useful for consistency checks and finding errors in calculations, just like in other engineering or physics calculations.

On the other hand, if we write probability densities as previously suggested, in this case as "$\mathrm{p}(T{=}20°\mathrm{C} \mid I)\,\mathrm{d}T$", then the density written this way does not need any units: the units "$°\mathrm{C}^{-1}$" disappear because multiplied by $\mathrm{d}T$, which has the inverse units "°C".

14.4 Representation of probability densities

14.4.1 Line-based representations

The histogram and the line representations become indistinguishable for a probability density.

If we represent the probability P assigned to a small range of width ϵ as the area of a rectangle, and the width of the rectangle is equal to ϵ, then the height P/ϵ of the rectangle is numerically equal to the probability density. The difference from histograms for discrete quantities lies in the values reported on the vertical axis: for discrete quantities the values are *probabilities* (the *areas* of the rectangles), but for continuous quantities they are probability *densities* (the *heights* of the rectangles). This is also evident from the fact that the values reported on the vertical axis can be larger than 1, as in the example plots shown in the margin.

The rectangles, however, are so thin (usually thinner than a pixel on a screen) that they appear just as vertical lines, and together they look just like a curve delimiting a coloured area. If we don't colour the area underneath the curve, then we just

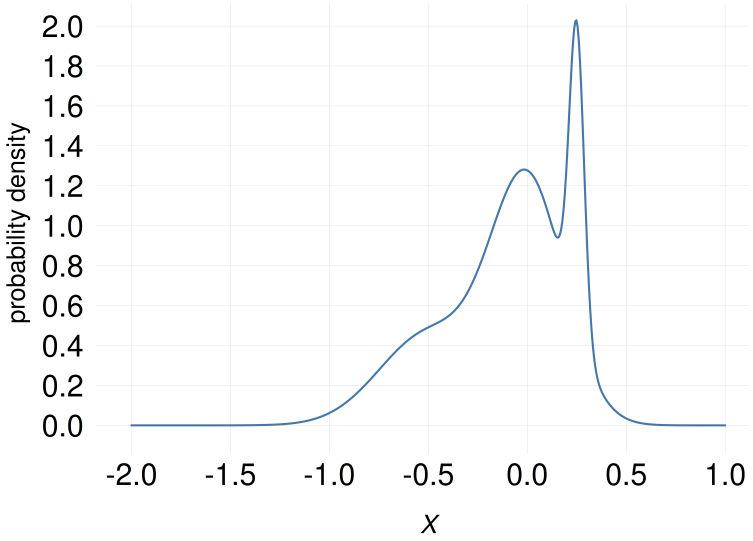
have a line-based, or rather curve-based, representation of the probability density.

Keep in mind that the curve representing the probability density is *not quite a function*. In fact it's best to call it a "density" or a "density function". There are important reasons for keeping this distinction, which have also consequences for probability calculations, but we shall not delve into them for the moment.

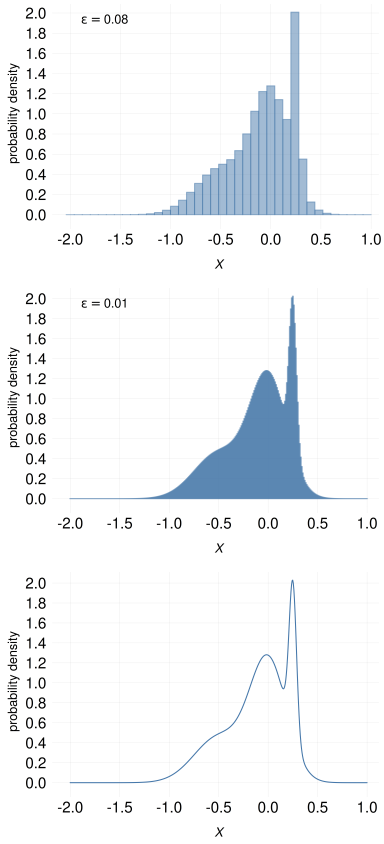
14.4.2 Scatter plots

Line plots of a probability density are very informative, but they can also be slightly deceiving. Try the following experiment.

Consider a continuous quantity X with the following probability density:



We want to represent the amount of probability in any small range, say between X=0 and X=0.1, by drawing in that range a number of short thin lines, the number being proportional to the probability. So a range containing 10 lines has twice the probability of a range containing 5 lines. The probability



As the width ϵ of the small ranges is decreased, a histogram based on these widths become indistinguishable from a line plot

density around a value is therefore roughly represented by the density of the lines around that value.

Suppose that we have 50 lines available to distribute this way. Where should we place them?

In a **scatter plot**, the probability density is (approximately) represented by density of lines, or points, or similar objects, as in the examples above (only one of the examples above, though, correctly matches the density represented by the curve).

As the experiment and exercise above may have demonstrated, line plots sometimes give us slightly misleading ideas of how the probability is distributed across the domain. For example, peaks at some values make us overestimate the probability density around those values. Scatter plots often give a less misleading representation of the probability density.

Scatter plots are also useful for representing probability densities in more than one dimension – sometimes even in infinite dimensions! They can moreover be easier to produce computationally than line plots.

@@ TODO Behaviour of representations under transformations of data.

Study reading

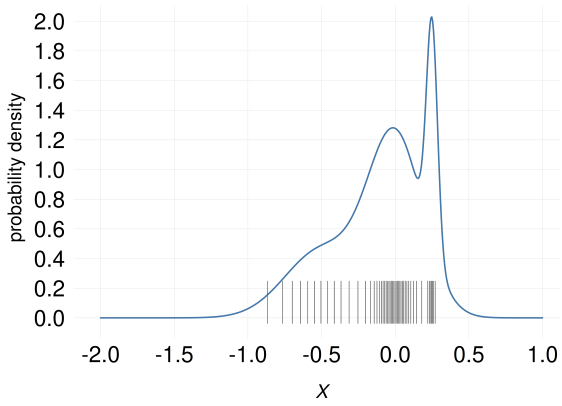
- §§5.3.0–5.3.1 of *Risk Assessment and Decision Analysis with Bayesian Networks*

14.5 Combined probabilities

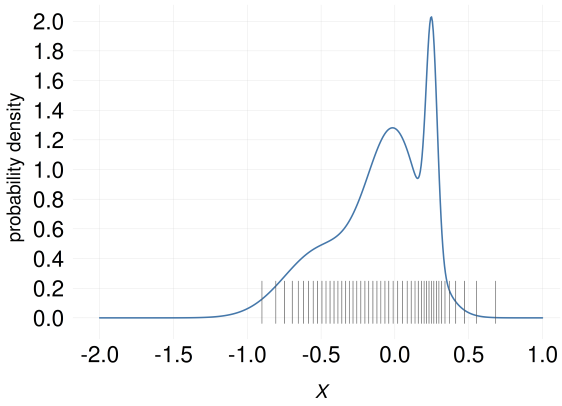
A probability distribution is defined over a set of mutually exclusive and exhaustive sentences. In some inference problems, however, we do not need the probability of those sentences, but of some other sentence that can be obtained from them by an `or` operation. The probability of this sentence can then be obtained by a sum, according to the `or`-rule of inference. We can

Exercise

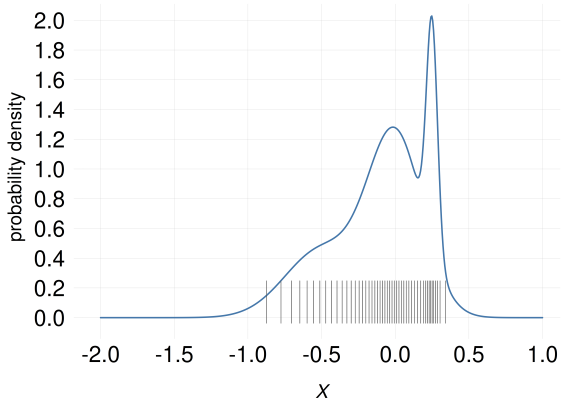
Which of these plots shows the correct placement of the 50 lines? (NB: the position of the correct answer is determined by a pseudorandom-number generator.)



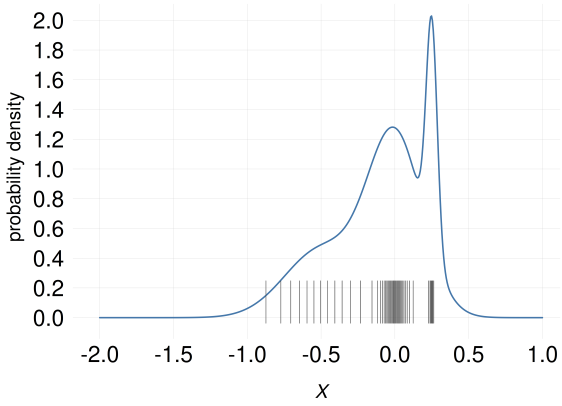
(a) (A)



(a) (B)



(a) (C)



(a) (D)

call this a *combined probability*. Let's explain this procedure with an example.

Back to our initial assembly-line scenario from Ch. 1, the inference problem was to predict whether a specific component would fail within a year or not. Consider the time when the component will fail (if it's sold), and represent it by the quantity T with the following 24 different values, where "mo" stands for "months":

'The component will fail during it 1st month of use'

'The component will fail during it 2nd month of use'

...

'The component will fail during it 23rd month of use'

'The component will fail during it 24th month of use or after'

which we can shorten to $T{=}1, T{=}2, \ldots, T{=}24$; note the slightly different meaning of the last value.

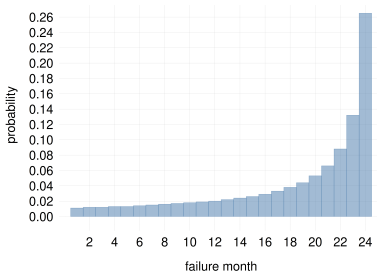
> **Exercise**
>
> What is the basic type of the quantity T? Which other characteristics does it have? for instance discrete? unbounded? rounded? uncensored?

Suppose that the inspection device – our agent – has internally calculated a probability distribution for T, conditional on its internal programming and the results of the tests on the component, collectively denoted I. The probabilities, compactly written, are

$$\mathrm{P}(T{=}1 \mid I), \quad \mathrm{P}(T{=}2 \mid I), \quad \ldots, \quad \mathrm{P}(T{=}24 \mid I)$$

Their values are stored in the file `failure_probability.csv` and plotted in the histogram on the side.

What's important for the agent's decision about rejecting or accepting the component, is not the exact time when it will fail, but only whether it will fail within the first year or not. That is, the agent needs the probability of the sentence



'The component will fail within a year of use'. But this sentence is just the `or` of the first 12 sentences expressing the values of T:

'The component will fail within a year of use'

$$\begin{aligned}
&\equiv \text{'The component will fail during it 1st month of use'} \vee \\
&\quad \text{'The component will fail during it 2nd month of use'} \vee \cdots \\
&\quad \cdots \vee \text{'The component will fail during it 12th month of use'} \\
&\equiv (T{=}1) \vee (T{=}2) \vee \cdots \vee (T{=}12)
\end{aligned}$$

The probability needed by the agent is therefore

$$\mathrm{P}(T{=}1 \vee T{=}2 \vee \cdots \vee T{=}12 \mid I)$$

which can be calculated using the `or`-rule, considering that the sentences involved are mutually exclusive:

$$\begin{aligned}
&\mathrm{P}(\text{'The component will fail within a year of use'} \mid I) \\
&\quad = \mathrm{P}(T{=}1 \vee T{=}2 \vee \cdots \vee T{=}12 \mid I) \\
&\quad = \mathrm{P}(T{=}1 \mid I) + \mathrm{P}(T{=}2 \mid I) + \cdots + \mathrm{P}(T{=}12 \mid I) \\
&\quad = \sum_{t=1}^{12} \mathrm{P}(T{=}t \mid I)
\end{aligned}$$

Sum notation

We shall often use the $\sum$-notation for sums, as in the example above. A notation like "$\sum_{i=5}^{20}$" means: write multiple copies of what's written on its right side, and in each copy replace the symbol "i" with values from 5 to 20, in turn; then sum up these copies. The symbol "i" is called the *index* of the sum. Sometimes the initial and final values, 5 and 20 in the example, are omitted if they are understood from the context, and the sum is written simply "$\sum_{i}$".

Exercise

Using your favourite programming language:

- Load the file `failure_probability.csv` containing the probabilities.
- Inspect this file, find the headers of its columns and so on.
- Calculate the probability that the component will fail within a year of use.
- Calculate the probability that the component will fail "within two months of use, or after a year of use".

15 Joint probability distributions

So far we have considered probability distributions for quantities of a basic (binary, nominal, ordinal, interval) type. These distributions have a sort of one-dimensional character and can be represented by ordinary histograms, line plots, and scatter plots. We now consider probability distributions for the kind of joint quantities that were discussed in § 13.1.

15.1 Joint probability distributions

A joint quantity is just a collection or set of quantities of basic types. Saying that a joint quantity has a particular value means that each basic component quantity has a particular value in its specific domain. This is expressed by an `and` of sentences.

Consider for instance the joint quantity X consisting of the age A and sex S of a specific person. The fact that X has a particular value is expressed by a composite sentence such as

'The person's age is 25 years and the person's sex is female'

which we can compactly write with an `and`:

$$A{=}25\,\mathrm{y} \wedge S{=}\mathrm{f}$$

All the possible composite sentences of this kind are *mutually exclusive* and *exhaustive*.

An agent's uncertainty about X's true value is therefore represented by a probability distribution over all `and`-ed sentences of this kind, representing all possible joint values:

$$\mathrm{P}(A{=}25\,\mathrm{y} \wedge S{=}\mathrm{f} \mid I)\ , \qquad \mathrm{P}(A{=}31\,\mathrm{y} \wedge S{=}\mathrm{m} \mid I)\ , \qquad \ldots$$

where I is the agent's state of knowledge, and the probabilities sum up to 1. We call each of these probabilities a **joint probability**, and their collection a **joint probability distribution**. Usually these probabilities are written in a much abbreviated form. A comma "," is typically used instead of "∧" (§ 6.4). You can commonly encounter the following notation:

$$\mathrm{P}(A{=}25\,, S{=}\mathrm{f} \mid I)$$

or even just

$$\mathrm{P}(25, \mathrm{f} \mid I)$$

15.2 Representation of joint probability distributions

There is a wide variety of ways of representing joint probability distributions, and new ways are invented (and rediscovered) all the time. In some cases, especially when the quantity has more than three component quantities, it can become impossible to graphically represent the probability distribution in a faithful way. Therefore one often tries to represent only some aspects or features of interest of the full distribution. Whenever you see a plot of a joint probability distribution, you should carefully read what the plot shows and how it was made. Here we only illustrate some examples and ideas for representations.

15.2.1 Tables

When a joint quantity consists of *two, discrete and finite* component quantities, the joint probabilities can be reported as a table, sometimes called a **contingency table**[1].

[1] this term is most often used for joint distributions of *frequencies* rather than probability

Example: Consider the next patient that will arrive at a particular hospital. There's the possibility of arrival by `ambulance`, `helicopter`, or `other` transportation means; and the possibility that the patient will need `urgent` or `non-urgent` care. We can represent these possibilities by two quantities T (nominal) and U (binary). Now suppose that an agent has the following joint probability distribution, conditional on the hospital's data I_{H}:

Table 15.1: Joint probability distribution for transportation and urgency

$\mathrm{P}(U{=}u, T{=}t \mid I_{\mathrm{H}})$		**transportation at arrival** T		
		ambulance	helicopter	other
urgency U	urgent	0.11	0.04	0.03
	non-urgent	0.17	0.01	0.64

From the table we see that the most probable possibility is that the next patient will arrive by other transportation means than ambulance and helicopter, and will not require urgent care:

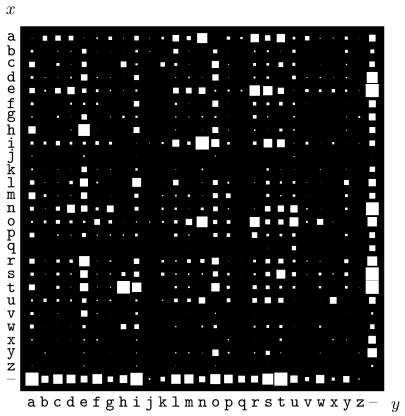
$$\mathrm{P}(U{=}\texttt{non-urgent}\,, T{=}\texttt{other} \mid I_{\mathrm{H}}) = 0.64$$

In this kind of tables it is also possible to replace the numerical probability values with graphical representations; for example as shades of a colour, or squares with different areas.

> Exercise – never forget the agent!
>
> Who could be the agent whose degrees of belief are represented in the table above? What could be the background information leading to such beliefs?

15.2.2 Scatter plots and similar

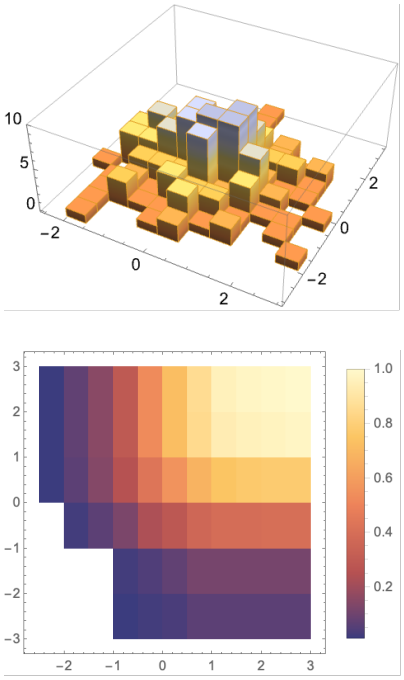
We saw in § 14.2 that probability distributions for nominal, ordinal, or discrete-interval quantities can be represented by histograms or line plots. Histograms could be generalized to



Probability distribution over the 27×27 possible bigrams xy in an English language document. Probabilities are represented by the areas of white squares. From MacKay's *Information Theory, Inference, and Learning Algorithms*

quantities consisting of *two* joint discrete quantities: a probability could be represented by a cuboid or rectangular prism, or cylinder, or similar. This representation, even if it can look flamboyant, is often inconvenient because some of the three-dimensional objects can be hidden from view, as in the example in the margin illustration.

Alternatively, one can replace the numerical values of the probabilities in the tabular representation of the previous section with some graphical encoding. An example is a colour scheme with `white` for probability 0, `black` for probability 1, and grey levels for intermediate probabilities. This is sometimes called a "density histogram"; see the example in the margin figure. This representation can be useful for qualitative or semi-quantitative assessments, for example for seeing which joint values have highest probabilities.



Examples of a density histogram and a generalized histogram (from Mathematica)

Another representation, similar to the scatter plot (§ 14.4.2), is to encode the probability values with a proportional number of points or other shapes, as illustrated here for the probabilities of table 15.1:

Figure 15.1: Scatter plot for the urgency-transportation joint probability distribution

the points do not need to be scattered in regular fashion as long as it's clear which quantity value they are associated with. The scatter plot above has 100 points, and therefore we can see for instance that $\mathrm{P}(U{=}\text{urgent}\,, T{=}\text{helicopter} \mid I_{\mathrm{H}}) = 0.03$, since the corresponding region has 3 points out of 100.

15.3 Joint probability densities

If a joint quantity consists in several continuous interval quantities, then its joint probability distribution is usually represented by a **joint probability density**, which generalizes the one-dimensional discussion of § 14.3 to several dimensions.

For instance, if X and Y are two continuous interval quantities, then the notation

$$\mathrm{p}(X{=}x\,, Y{=}y \mid I) = 0.001$$

means that the joint sentence "X has value between $x-\epsilon/2$ and $x+\epsilon/2$, and Y between $y-\delta/2$ and $y+\delta/2$", or in symbols

$$(x - \tfrac{\epsilon}{2}{<}X{<}x + \tfrac{\epsilon}{2}) \wedge (y - \tfrac{\delta}{2}{<}Y{<}y + \tfrac{\delta}{2})$$

in being given a degree of belief $0.001 \cdot \epsilon \cdot \delta$, conditional on the background knowledge I. Visually, the rectangular region of values around (x, y) with sides of lengths ϵ and δ is assigned a probability $0.001 \cdot \epsilon \cdot \delta$.

Remember that a density typically has physical units, as in the one-dimensional case (§ 14.3). For instance, if X above is a temperature measured in kelvin (K) and Y a resistance measured in ohm (Ω), then we should write

$\mathrm{p}(X{=}x\,, Y{=}y \mid I) = \frac{0.001}{\mathrm{K}\,\Omega}$.

15.4 Representation of joint probability densities

For one-dimensional densities we discussed line-based representations and scatter plots (§ 14.4). The first of these representations can be generalized to two-dimensional densities, leading to a **surface plot**. Below you see the surface density plot for the probability density given by the formula

$$\mathrm{p}(X{=}x\,,Y{=}y\,|\,I) = \tfrac{3}{8\,\pi}\,\mathrm{e}^{-\frac{1}{2}(x-1)^2-(y-1)^2} + \tfrac{3}{64\,\pi}\,\mathrm{e}^{-\frac{1}{32}(x-2)^2-\frac{1}{2}(y-4)^2} + \tfrac{1}{40\,\pi}\,\mathrm{e}^{-\frac{1}{8}(x-5)^2-\frac{1}{5}(y-2)^2}$$

This kind of representation can be neat, but it has three drawbacks: 1. It sometimes hides from view some features of the density (in the plot above, can you exclude that there's a small peak right behind the main one?). 2. It cannot be extended to three-dimensional densities. 3. Sometimes the analytical expression for the probability density (like the formula above) is not available.

The scatter plot overcomes the three drawbacks above. It does not hides features; it can also be used for three-dimensional densities; it can be generated in cases where the analytical formula of a probability distribution is not available or too complicated, but we can still obtain "representative" points from it. The representation of a scatter plot is, however, quantitatively more imprecise. Here is a scatter plot, using 10 000 points, for the probability density given above:

The probability of a small region is proportional to the density of points in that region. If we had a joint density for *three* continuous quantities, its scatter plot would consist of three-dimensional clouds of points instead.

Clearly both kinds of representation have advantages and disadvantages. The choice between them depends on the problem, on the probability density, and on what we wish to visually emphasize. It is also possible to use both, of course.

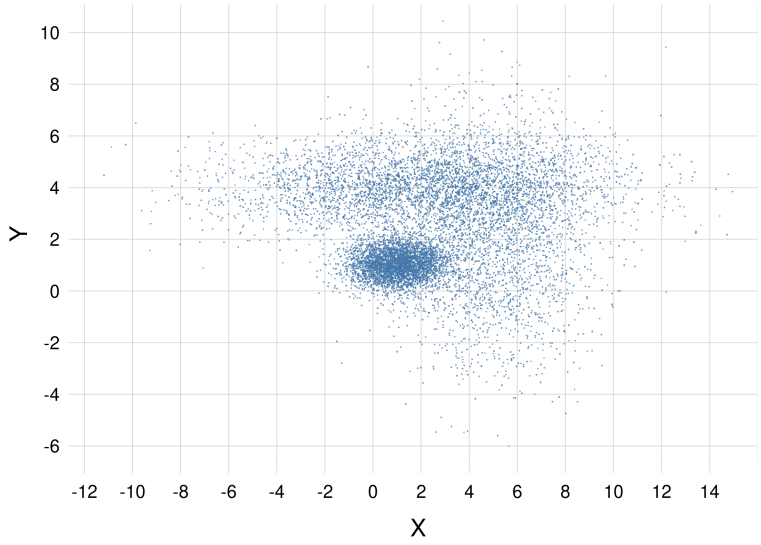


Figure 15.2: Scatter-plot representation of the joint probability density $\mathrm{p}(X{=}x\,, Y{=}y \mid I)$ above

15.5 Joint mixed discrete-continuous probability distributions

Frequently occurring in engineering and data-science problems are joint quantities composed by some discrete and some continuous quantities. Their joint probability distribution is a density with respect to the continuous component quantity.

Suppose for instance that Z is a binary quantity with domain $\{\texttt{low;}, \texttt{high;}\}$, and X a real-valued continuous quantity with domain $\mathbf{R}$. Together they form the joint quantity $(Z, X) \in \{\texttt{low;}, \texttt{high;}\} \times \mathbf{R}$. Then the probability expression

$$\mathrm{p}(Z{=}\texttt{low}\,, X{=}3 \mid I) = 0.07$$

means that the agent with background information I has a degree of belief equal to $0.07 \cdot \epsilon$ in the joint sentence "Z has value `low` and X has value between $3 - \epsilon/2$ and $3 + \epsilon/2$". As usual, this is only valid for any small ϵ, and if X has physical dimensions, say metres m, then the probability density above has value $0.07\,\mathrm{m}^{-1}$.

15.6 Representation of mixed probability distributions

Mixed discrete-continuous probability distributions can be somewhat tricky to represent graphically. Here we consider line-based representations and scatter plots. We take as example the probability that the next patient who arrives at a particular hospital has a given age (positive continuous quantity) and arrives by `ambulance`, `helicopter`, or `other` transportation means (table 15.1).

15.6.1 Multi-line plots

A line plot can be used to represent the probability density for the continuous quantity and each specific value of the discrete quantity:

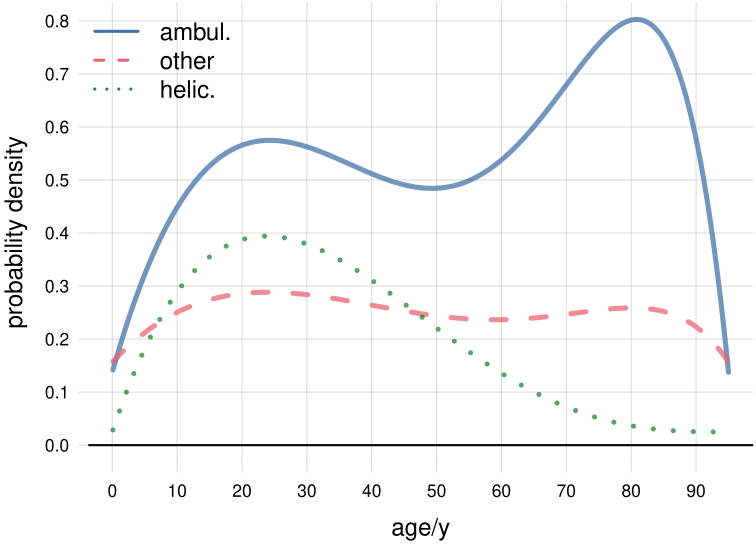


Figure 15.3: Line plot for the age-transportation joint probability distribution (table 15.1)

With the plot above it's important to keep in mind that the three curves are three pieces of the *same* probability density, not three different densities. This is also clear from the fact

that the three areas under them (which partly overlap) cannot each be equal to 1, as would instead be required for a probability density. The probability density is separated into three curves owing to the presence of the discrete quantity, which has three possible values.

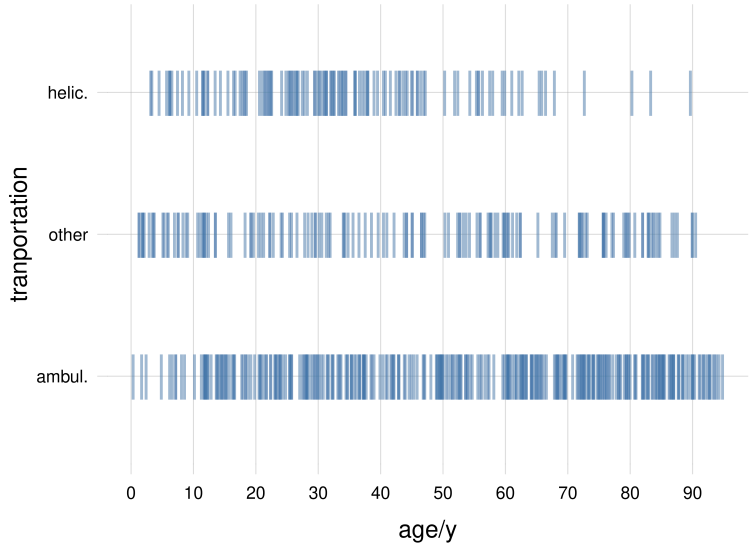
The area under the solid blue curve is equal to 0.55, the area under the dashed red curve is 0.25, and the area under the dotted green curve is 0.20 . The total area under the three curves (counting also the overlapping regions) is equal to 1, as it should.

A possible disadvantage of this kind of plots is that some details, such as peaks, of the densities for some values of the discrete quantity, may be barely discernible.

15.6.2 Scatter plots

As discussed before, in a scatter plot we represent the probability density by a cloud of "representative" objects, such as points, obtained from it. The density of these objects is approximately proportional to the density of probability.

Here is an example of scatter plot for the probability density of table 15.1:



In the plot above, the probability density is reflected by the density of vertical lines. Using points instead of vertical lines, the density would have been difficult to discern, since the points would all lie on three lines.

We can use points if we give some variation, usually called jitter, to their vertical coordinate; but we must keep in mind that such vertical variation has no meaning. The idea is similar to the one of fig. 15.1. In our current example of table 15.1 we obtain a plot like this:

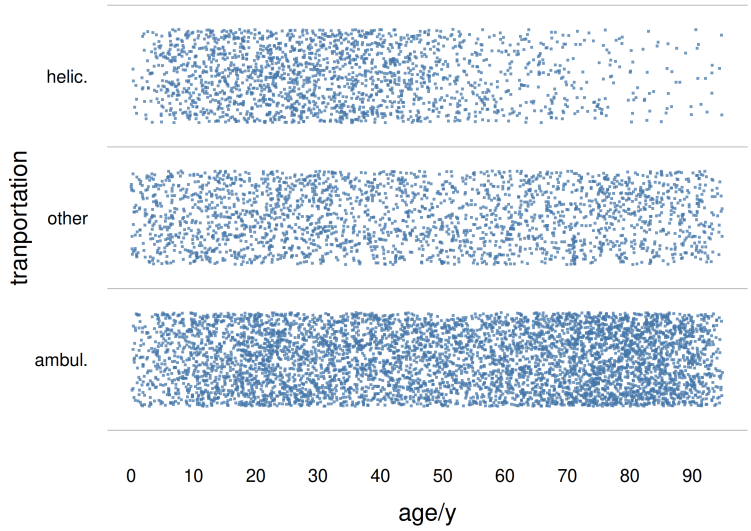


Figure 15.4: Point-scatter plot for the age-transportation joint probability

Exercise

Compare the line plot of fig. 15.3 and the point-scatter plot of fig. 15.4, which represent the same joint probability density. Do some introspection, and analyse the contrasting impressions that the two kinds of representations may give you. For instance, does the line plot give you a wrong intuition about the sharpness of the peaks in the density?
Compare with what you did in the exercise of § 14.4.

15.7 Representation of more general probability distributions and densities

Probability distributions for complex types of quantity can be quite tricky to visualize and represent in an informative way. They typically require a case-by-case approach.

Often the idea behind the scatter plot works also in these complex cases: the probability distribution or density is represented by a "representative" sample of objects. The objects can even depict the quantity itself.

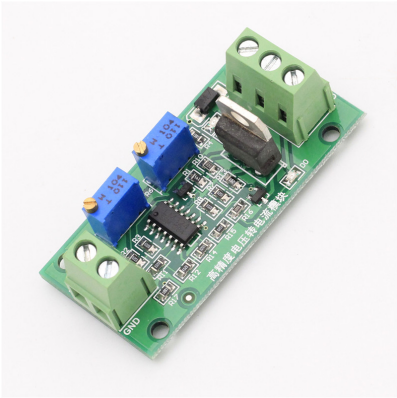
For instance, imagine the complex quantity L defined as "the linear relationship between input voltage and output current of a specific electronic component". The possible values of this quantity are *straight lines*, that is, functions of the form "$y = m\,x + q$", where x is the input voltage and y the output current. These possible values – straight lines – can differ in their angular coefficient m or in their intercept q. One possible value could be the straight line

$$y = (2\,\mathrm{A/V})\,x - 3\,\mathrm{A}$$

another possible value could be the straight line

$$y = (-1\,\mathrm{A/V})\,x + 5\,\mathrm{A}$$

and so on. The quantity L so defined is a continuous quantity, but it isn't a quantity of a basic type.



A voltage-current converter

An agent may be uncertain about the actual value of L, that is, about what is the straight line that correctly expresses the voltage-current relationship of this particular electronic component. The agent therefore assignes a probability density over all possible values: over all possible straight lines. How to visually represent such a "probability density over lines"?

One way is to use a *scatter plot*. The probability distribution is represented by a collection of *straight lines*, whose density is approximately proportional to the probability density. Here is an example using 360 representative straight lines:

From this plot we can read some important semi-quantitative information about the agent's degrees of belief. For instance:

Figure 15.5: Scatter plot for a probability density over the voltage-current relationship

- It's most probable that the voltage-current relationship has a positive angular coefficient m with value around $0.5\,\mathrm{A/V}$, and an intercept q around $3\,\mathrm{A}$.
- It is improbable, but not impossible, that the voltage-current relationship has a negative angular coefficient (that is, the output current decreases as the input voltage is increased).
- It's practically impossible that the voltage-current relationship is almost vertical (say, changes in current larger than $\sim 5\,\mathrm{A}$ with changes in voltage smaller than $\sim 0.2\,\mathrm{V}$).

Exercises

- Explore datasets in a database such as the UC Irvine Machine Learning Repository, for example
  - The adult-income dataset
  - The heart-disease dataset

  Assume that the data given are *representative "points"* of a probability distribution or density (of which we don't know the analytic formula). Plot the probability distributions and probability densities as scatter plots using some of these representative points.

- Look around for analytic formulae of some probability distributions and densities of simple and joint quantities, and plot them using different representations.

Study reading

- §5.3.2 of *Risk Assessment and Decision Analysis with Bayesian Networks*
- §12.2.2 of *Artificial Intelligence*

16 Marginal probability distributions

16.1 Marginal probability: neglecting some quantities

In some situations an agent has a joint distribution of degrees of belief for the possible values of a joint quantity, but it needs to consider its belief in the value of *one* component quantity alone, *irrespective of what the values for the other components quantities might be*.

Consider for instance the joint probability for the next-patient arrival scenario of table 15.1 from § 15.2, with joint quantity (U, T). We may be interested in the probability that the next patient will need `urgent` care, *independently of how the patient is transported to the hospital*. This probability can be found, as usual, by analysing the problem in terms of *sentences* and using the basic rules of inference from § 8.4.

The sentence of interest is "The next patient will require urgent care", or in symbols

$$U{=}\texttt{urgent}$$

This sentence is equivalent to "The next patient will require urgent care, and will arrive by ambulance, helicopter, or other means", or in symbols

$$U{=}\texttt{urgent} \wedge (T{=}\texttt{ambulance} \vee T{=}\texttt{helicopter} \vee T{=}\texttt{other})$$

Using the derived rules of Boolean algebra of § 9.2 we can rewrite this sentence in yet another way:

$$(U{=}\texttt{urgent} \wedge T{=}\texttt{ambulance}) \vee (U{=}\texttt{urgent} \wedge T{=}\texttt{helicopter}) \vee (U{=}\texttt{urgent} \wedge T{=}\texttt{other})$$

This last sentence is an `or` of *mutually exclusive* sentences. Its probability is therefore given by the `or` rule, with the `and` terms being zero (we shall now use the comma "," for `and`):

$$\begin{aligned}
&\mathrm{P}(U{=}\texttt{urgent} \mid I_\mathrm{H}) \\
&\quad = \mathrm{P}\big[(U{=}\texttt{urgent}\,, T{=}\texttt{ambulance}) \vee (U{=}\texttt{urgent}\,, T{=}\texttt{helicopter}) \vee (U{=}\texttt{urgent}\,, T{=}\texttt{other}) \mid I_\mathrm{H}\big] \\
&\quad = \mathrm{P}(U{=}\texttt{urgent}\,, T{=}\texttt{ambulance} \mid I_\mathrm{H}) + \\
&\qquad\quad \mathrm{P}(U{=}\texttt{urgent}\,, T{=}\texttt{helicopter} \mid I_\mathrm{H}) + \\
&\qquad\quad \mathrm{P}(U{=}\texttt{urgent}\,, T{=}\texttt{other} \mid I_\mathrm{H})
\end{aligned}$$

We have found that the probability for a value of the urgency quantity U, independently of the value of the transportation quantity T, can be calculated by summing all joint probabilities with all possible T values. Using the $\sum$-notation we can write this compactly:

$$\mathrm{P}(U{=}\texttt{urgent} \mid I_\mathrm{H}) = \sum_t \mathrm{P}(U{=}\texttt{urgent}\,, T{=}t \mid I_\mathrm{H})$$

where it's understood that the sum index t runs over the values $\{\texttt{ambulance;}, \texttt{helicopter;}, \texttt{other;}\}$.

This is called a **marginal probability**.

> Exercise
>
> Using the values from table 15.1, calculate:
>
> - the marginal probability that the next patient will need urgent care
> - the marginal probability that the next patient will arrive by helicopter

Considering now a more generic case of a joint quantity with component quantities X and Y, the probability for a specific value of X, conditional on some information I and irrespective of what the value of Y might be, is given by

$$\mathrm{P}(X{=}x \mid I) = \sum_{y} \mathrm{P}(Y{=}y\,, X{=}x \mid I)$$

You may notice the similarity with the expression for a *combined probability* from § 14.5. Indeed a marginal probability is just a special case of a combined probability: we are combining all probabilities that exhaust the possibilities for the sentence $Y{=}y$.

> Exercise: test your understanding
>
> Using again the values from table 15.1, calculate the probability that *the next patient will need urgent care and will be transported either by ambulance or by helicopter.*

16.2 Marginal density distributions

In the example of the previous section, suppose now that the quantities X and Y are continuous. Then the joint probability is expressed by a density:

$$\mathrm{p}(Y{=}y\,, X{=}x \mid I)$$

with the usual meaning. The marginal probability density for X is still given by a sum, but this sum occurs over intervals of values of Y, intervals with very small widths. As a consequence the sum will have a very large number of terms. To remind ourselves of this fact, which can be very important in some situations, we use a different notation in terms of integrals:

$$\mathrm{p}(X{=}x \mid I) = \int_{\Upsilon} \mathrm{p}(Y{=}y\,, X{=}x \mid I)\,\mathrm{d}y$$

where Υ represents the domain of the quantity Y.

This is called a **marginal probability density**.

The appearance of integrals is sometimes extremely useful, because it allows us to use the theory of integration to calculate marginal probabilities quickly and precisely, instead of having to compute sums of a large numbers of small terms – a procedure that can be computationally expensive and lead to numerical errors owing to underflow or similar computation problems.

16.3 Marginal probabilities and scatter plots

In the previous chapters we have often discussed scatter plots (§ 14.4.2, § 15.6.2) for representing probability distributions of various kinds: discrete, continuous, joint, mixed, and so on.

One more advantage of representing a joint distribution with a scatter plot is that it can be quickly modified to represent any marginal distribution, again with a scatter plot. Whereas the use of a surface plot would require analytical calculations or approximations thereof.

Consider for instance the joint probability density from § 15.4, represented by the formula

$$\mathrm{p}(X{=}x\,, Y{=}y \mid I) = \tfrac{3}{8\,\pi}\,\mathrm{e}^{-\frac{1}{2}(x-1)^2-(y-1)^2} + \tfrac{3}{64\,\pi}\,\mathrm{e}^{-\frac{1}{32}(x-2)^2-\frac{1}{2}(y-4)^2} + \tfrac{1}{40\,\pi}\,\mathrm{e}^{-\frac{1}{8}(x-5)^2-\frac{1}{5}(y-2)^2}$$

and suppose we would like to visualize the marginal probability density for X:

$\mathrm{p}(X{=}x \mid I)$.

In order to represent this marginal probability density with a line plot, we would first need to calculate the integral of the formula above over Y:

$$\mathrm{p}(X{=}x \mid I) = \int_{-\infty}^{\infty} \left[\tfrac{3}{8\,\pi}\,\mathrm{e}^{-\frac{1}{2}(x-1)^2-(y-1)^2} + \tfrac{3}{64\,\pi}\,\mathrm{e}^{-\frac{1}{32}(x-2)^2-\frac{1}{2}(y-4)^2} + \tfrac{1}{40\,\pi}\,\mathrm{e}^{-\frac{1}{8}(x-5)^2-\frac{1}{5}(y-2)^2} \right] \mathrm{d}y$$

Now instead suppose that we have stored the points used to represent the joint probability density $\mathrm{p}(X{=}x\,, Y{=}y \mid I)$ as a scatter plot, as in fig. 15.2. Each of these points is a pair of coordinates (x, y), representing an X-value and a Y-value. It turns out that *these same points can be used to make a scatter-plot of the marginal density for* X, simply by considering their x-coordinates only, that is, by discarding their y-coordinates. Often we use a subsample (unsystematically chosen) of them, so that the resulting one-dimensional scatter plot doesn't become too congested and difficult to read.

As an example, here is a scatter plot for the marginal probability density $\mathrm{p}(X{=}x \mid I)$ above, obtained by selecting a subset of 400 points from the scatter plot (fig. 15.2) for the joint distribution. The points are replaced by vertical lines for better visibility:

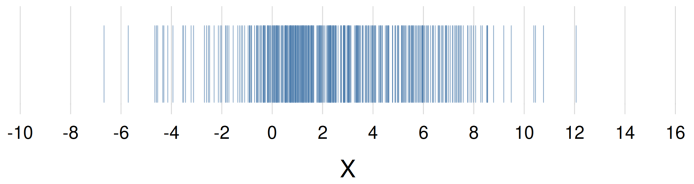


Figure 16.1: Scatter-plot representation of the marginal probability density $\mathrm{p}(X{=}x \mid I)$

Exercise

The points for the scatter plot of fig. 15.2 (§ 15.4) are saved in the file `scatterXY_samples.csv`. Use them to represent the marginal probability density $\mathrm{p}(Y{=}y \mid I)$, for the other quantity Y, as a scatter plot.

16.4 Uses and pitfalls of marginal probability distributions

An agent's distribution of degrees of belief for a multi-dimensional joint quantity is not easily – or at all – visualizable. This shortcoming is especially bad because, as discussed in

§ 10.1, our intuition often fails us horribly in multi-dimensional problems.

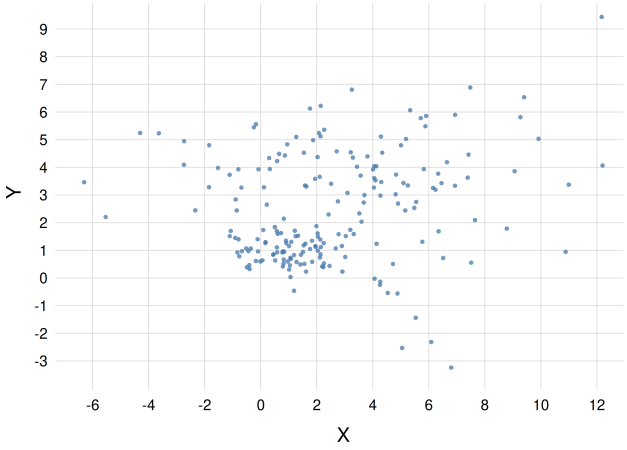
Marginal probability distributions for one or two of the component quantities are useful because they offer us a little glimpse of the multi-dimensional "monster" distribution. In concrete engineering and data-science problem, when we need to discuss a multi-dimensional distribution it is good practice to visually report at least its one-dimensional marginal distributions.

In the machine-learning literature, this low-dimensional glimpse is often used to qualitatively assess whether two multi-dimensional distributions are similar. Their one-dimensional marginals are visually compared and, if they overlap, one hopes (but some works in the literature even erroneously *conclude*) that the multi-dimensional distributions are somewhat similar as well.

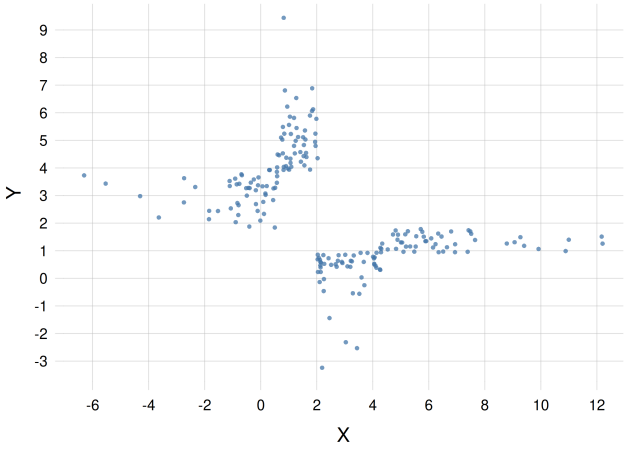
Keep in mind that this may very well not be the case. Marginal distributions can also be quite deceiving:

Exercise

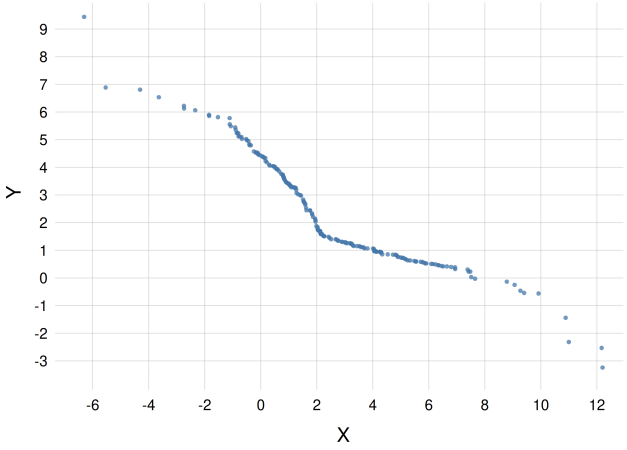
Here are three different joint probability densities for the joint quantity (X, Y), each density represented by a scatter plot with 200 points. the files containing the coordinates of the scatter-plot points are also given:
A. File `scatterXY_A.csv`:



B. File `scatterXY_B.csv`:



C. File `scatterXY_C.csv`:



- Reproduce the three scatter plots above using the points from the three files, just to confirm that they are correct.
- For each density, plot the marginal density for the quantity X as a scatter plot. Use the method described in § 16.3; do not subsample the points. What can you say about the three marginal densities you obtain?

- Do the same, but for the marginal densities for Y. What can you say about the three marginal densities you obtain?
- If two joint probability distributions have the same marginals, can we conclude that they are identical, or at least similar?

Study reading

- §§5.3.2–5.3.3 of *Risk Assessment and Decision Analysis with Bayesian Networks*
- §12.3 of *Artificial Intelligence*
- §§5.1–5.5 of *Probability*

17 Conditional probability and learning

17.1 The meaning of the term "conditional probability"

When we introduced the notion of degree of belief – a.k.a. probability – in chapter 8, we emphasized that *every probability is conditional on some state of knowledge or information*. So the term "conditional probability" sounds like a pleonasm, just like saying "round circle".

This term must be understood in a way analogous to "marginal probability": it applies in situations where we have two or more sentences of interest. We speak of a "conditional probability" when we want to emphasize that additional sentences appear in the conditional (right side of " | ") of that probability. For instance, in a scenario with these two probabilities:

$$\mathrm{P}(A \mid B, I) \qquad \mathrm{P}(A \mid I)$$

we call the first **conditional probability** of A (**given** B) to emphasize or point out that its conditional includes the additional sentence B, whereas the conditional of the second probability doesn't include this sentence.

17.2 The relation between *learning* and conditional probability

Why do we need to emphasize that a particular degree of belief is conditional on an additional sentence? Because the ad-

ditional sentence usually represents *new information that the agent has learned*.

Remember that the conditional of a probability usually contains all factual information known to the agent[1]. Therefore if an agent acquires new data or a new piece of information expressed by a sentence D, it should draw inferences and make decisions using probabilities that include D in their conditional. In other words, the agent before was drawing inferences and making decisions using some probabilities

$$\mathrm{P}(\dots \mid K)$$

where K is the agent's knowledge until then. Now that the agent has acquired information or data D, it will draw inferences and make decisions using probabilities

$$\mathrm{P}(\dots \mid D, K)$$

Vice versa, if we see that an agent is calculating new probabilities conditional on an additional sentence D, then it means[2] that the agent has acquired that information or data D.

Therefore **conditional probabilities represent an agent's learning** and **should be used when an agent has learned something**.

This learning can be of many different kinds. Let's examine two particular kinds by means of some examples.

[1] Exceptions are, for instance, when the agent does *counterfactual* or *hypothetical* reasoning, as we discussed in § 5.1.

[2] But keep again in mind exceptions like counterfactual reasoning; see the previous side note.

17.3 Learning about a quantity from a *different* quantity

Consider once more the next-patient arrival scenario of § 15.2, with joint quantity (U, T) and an agent's joint probability distribution as in table 15.1, reproduced here:

Table 17.1: Joint probability distribution for transportation and urgency

$\mathrm{P}(U{=}u, T{=}t \mid I_{\mathrm{H}})$		**transportation at arrival** T		
		ambulance	helicopter	other
urgency U	urgent	0.11	0.04	0.03
	non-urgent	0.17	0.01	0.64

Suppose that the agent must forecast whether the next patient will require `urgent` or `non-urgent` care, so it needs to calculate the probability distribution for U (that is, the probabilities for U=`urgent` and U=`non-urgent`).

In the first exercise of § 16.1 you found that the marginal probability that the next patient will need urgent care is

$$\mathrm{P}(U{=}\texttt{urgent} \mid I_{\mathrm{H}}) = 18\%$$

this is the agent's degree of belief if it has nothing more and nothing less than the knowledge encoded in the sentence I_{H}.

But now let's imagine that the agent *receives a new piece of information*: it is told that the next patient is being transported by helicopter. In other words, **the agent has learned that the sentence** T=`helicopter`; **is true**. The agent's complete knowledge is therefore now encoded in the `and`ed sentence

$$T{=}\texttt{helicopter} \;\wedge\; I_{\mathrm{H}}$$

and this composite sentence should appear in the conditional. The agent's belief that the next patient requires urgent care, given the new information, is therefore

$$\mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{helicopter}\,, I_{\mathrm{H}})$$

Calculation of this probability can be done by just one application of the `and`-rule, leading to a formula connected with Bayes's theorem (§ 9.4):

$$\mathrm{P}(U{=}\texttt{urgent}\,, T{=}\texttt{helicopter} \mid I_{\mathrm{H}}) = \mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{helicopter}\,, I_{\mathrm{H}}) \cdot \mathrm{P}(T{=}\texttt{helicopter} \mid I_{\mathrm{H}})$$

$$\implies \quad \mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{helicopter}\,, I_{\mathrm{H}}) = \frac{\mathrm{P}(U{=}\texttt{urgent;}\,, T{=}\texttt{helicopter;} \mid I_{\mathrm{H}})}{\mathrm{P}(T{=}\texttt{helicopter;} \mid I_{\mathrm{H}})}$$

Let's see how to calculate this. The agent already has the joint probability for $U{=}\texttt{urgent} \wedge T{=}\texttt{helicopter}$ that appears in the numerator of the fraction above. The probability in the denominator is just a marginal probability for T, and we know how to calculate that too from § 16.1. So we find

$$\mathrm{P}(U{=}\texttt{urgent}|T{=}\texttt{helicopter},I_{\mathrm{H}}) = \frac{\mathrm{P}(U{=}\texttt{urgent;}\,, T{=}\texttt{helicopter;} \mid I_{\mathrm{H}})}{\sum_u \mathrm{P}(U{=}u\,, T{=}\texttt{helicopter;} \mid I_{\mathrm{H}})}$$

where it's understood that the sum index u runs over the values $\{\texttt{urgent;}, \texttt{non-urgent;}\}$.

This is called a **conditional probability**; in this case, the conditional probability of $U{=}\texttt{urgent}$ **given** $T{=}\texttt{helicopter}$.

The collection of probabilities for all possible values of the quantity U, given a *specific* value of the quantity T, say `helicopter`:

$$\mathrm{P}(U{=}\texttt{urgent}|T{=}\texttt{helicopter},I_{\mathrm{H}})\,, \qquad \mathrm{P}(U{=}\texttt{non-urgent}|T{=}\texttt{helicopter},I_{\mathrm{H}})$$

is called the **conditional probability distribution** for U **given** $T{=}\texttt{helicopter}$. It is indeed a probability distribution because the two probabilities sum up to 1.

⚠

Note that the collection of probabilities for, say, U=`urgent`, but for *different* values of the conditional quantity T, that is:

$$\mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{ambulance}\,, I_{\mathrm{H}})\ ,$$

$$\mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{helicopter}\,, I_{\mathrm{H}})\ ,$$

$$\mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{other}\,, I_{\mathrm{H}})$$

is **not** a probability distribution. Calculate the three probabilities above and check that in fact they do *not* sum up to 1.

Exercise

- Using the values from table 15.1 and the formula for marginal probabilities, calculate:
  - The conditional probability that the next patient needs urgent care, given that the patient is being transported by helicopter.
  - The conditional probability that the next patient is being transported by helicopter, given that the patient needs urgent care.
- Now discuss and find an intuitive explanation for these comparisons:
  - The two probabilities you obtained above. Are they equal? why or why not?
  - The *marginal* probability that the next patient will be transported by helicopter, with the *conditional* probability that the patient will be transported by helicopter *given* that it's urgent. Are they equal? if not, which is higher, and why?

17.4 Learning about a quantity from instances of *similar* quantities

In the previous section we examined how learning about one quantity can change an agent's degree of belief about a *different* quantity, for example knowledge about "transportation" affects beliefs about "urgency", or vice versa. The agent's learning and ensuing belief change are reflected in the value of the corresponding conditional probability.

This kind of change can also occur with "similar" quantities, that is, quantities that represent the same kind of phenomenon and have the same domain. The maths and calculations are identical to the ones we explored above, but the interpretation and application can be somewhat different.

As an example, imagine a scenario similar to the next-patient arrival above, but now consider the *next three patients* to arrive and their urgency. Define the following three quantities:

U_1 : urgency of the next patient
U_2 : urgency of the second future patient from now
U_3 : urgency of the third future patient from now

Each of these quantities has the same domain: {`urgent;`, `non-urgent;`}.

The joint quantity (U_1, U_2, U_3) has a domain with $2^3 = 8$ possible values:

- U_1=`urgent` , U_2=`urgent` , U_3=`urgent`
- U_1=`urgent` , U_2=`urgent` , U_3=`non-urgent`
- ...
- U_1=`non-urgent` , U_2=`non-urgent` , U_3=`urgent`
- U_1=`non-urgent` , U_2=`non-urgent` , U_3=`non-urgent`

Suppose that an agent, with background information I, has a particular joint belief distribution for the joint quantity (U_1, U_2, U_3). For example consider the joint distribution implicitly given as follows:

- If `urgent` appears in the probability 0 times out of 3: probability = 53.6%
- If `urgent` appears 1 times out of 3: probability = 11.4%
- If `urgent` appears 2 times out of 3: probability = 3.6%
- If `urgent` appears 3 times out of 3: probability = 1.4%

Here are some examples of how the probability values are determined by the description above:

$$\mathrm{P}(U_1{=}\texttt{urgent}\,, U_2{=}\texttt{non-urgent}\,, U_3{=}\texttt{urgent} \mid I) = 0.036 \qquad (\texttt{urgent}; \text{ appears twice})$$

$$\mathrm{P}(U_1{=}\texttt{non-urgent}\,, U_2{=}\texttt{urgent}\,, U_3{=}\texttt{non-urgent} \mid I) = 0.114 \qquad (\texttt{urgent}; \text{ appears once})$$

$$\mathrm{P}(U_1{=}\texttt{urgent}\,, U_2{=}\texttt{urgent}\,, U_3{=}\texttt{non-urgent} \mid I) = 0.036 \qquad (\texttt{urgent}; \text{ appears twice})$$

$$\mathrm{P}(U_1{=}\texttt{non-urgent}\,, U_2{=}\texttt{non-urgent}\,, U_3{=}\texttt{non-urgent} \mid I) = 0.536 \qquad (\texttt{urgent}; \text{ doesn't appear})$$

Exercise

- Check that the joint probability distribution as defined above indeed sums up to 1.
- Calculate the marginal probability for $U_1{=}\texttt{urgent}$, that is, $\mathrm{P}(U_1{=}\texttt{urgent} \mid I)$.
- Calculate the marginal probability that the second and third patients are non-urgent cases, that is

$$\mathrm{P}(U_2{=}\texttt{non-urgent}\,, U_3{=}\texttt{non-urgent} \mid I)\ .$$

From this joint probability distribution the agent can calculate, among other things, its degree of belief that the *third* patient will require urgent care, regardless of the urgency of the preceding two patients. It's the marginal probability

$$\begin{aligned}\mathrm{P}(U_3{=}\texttt{urgent} \mid I) &= \sum_{u_1}\sum_{u_2} \mathrm{P}(U_1{=}u_1\,, U_2{=}u_2\,, U_3{=}\texttt{urgent} \mid I) \\ &= 0.114 + 0.036 + 0.036 + 0.014 \\ &= 20.0\%\end{aligned}$$

where each index u_1 and u_2 runs over the values $\{\texttt{urgent;}, \texttt{non-urgent;}\}$. This double sum therefore involves four terms. The first term in the sum corresponds to "U_1=urgent,U_2=urgent,U_3=urgent" and therefore has probability 0.014 . The second term corresponds to "U_1=urgent , U_2=non-urgent , U_3=urgent" and therefore has probability 0.036. And so on.

Therefore the agent, with its current knowledge, has a 20% degree of belief that the third patient will require urgent care.

Now fast-forward in time, after *two* patients have arrived and have been taken good care of; or maybe they haven't arrived yet, but their urgency conditions have been ascertained and communicated to the agent. Suppose that *both patients were or are non-urgent cases*. The agent now knows this fact. The agent needs to forecast whether the third patient will require urgent care.

The relevant degree of belief is obviously not $\mathrm{P}(U_3=\texttt{urgent}|I)$, calculated above, because this belief represents an agent knowing only I. Now, instead, the agent has additional information about the first two patients, encoded in this **and**ed sentence:

$$U_1=\texttt{non-urgent}\,, U_2=\texttt{non-urgent}$$

The relevant degree of belief is therefore the *conditional* probability

$$\mathrm{P}(U_3=\texttt{urgent}\,|\,U_1=\texttt{non-urgent}\,, U_2=\texttt{non-urgent}\,, I)$$

Which we can calculate with the same procedure as in the previous section:

$$
\begin{aligned}
&\mathrm{P}(U_3{=}\texttt{urgent} \mid U_1{=}\texttt{non-urgent}\,, U_2{=}\texttt{non-urgent}\,, I) \\
&\quad = \frac{\mathrm{P}(U_1{=}\texttt{non-urgent};\,, U_2{=}\texttt{non-urgent};\,, U_3{=}\texttt{urgent};\mid I)}{\mathrm{P}(U_1{=}\texttt{non-urgent};\,, U_2{=}\texttt{non-urgent};\mid I)} \\
&\quad = \frac{0.114}{0.65} \\
&\quad \approx 17.5\%
\end{aligned}
$$

This conditional probability 17.5% for $U_3{=}\texttt{non-urgent}$ is *lower* than 20.0% calculated previously, which was based only on knowledge I. **Learning about the two first patients has thus affected the agent's degree of belief about the third**.

Let's also check how the agent's belief changes in the case where the first two patients are both *urgent* instead. The calculation is completely analogous:

$$
\begin{aligned}
&\mathrm{P}(U_3{=}\texttt{urgent} \mid U_1{=}\texttt{urgent}\,, U_2{=}\texttt{urgent}\,, I) \\
&\quad = \frac{\mathrm{P}(U_1{=}\texttt{urgent};\,, U_2{=}\texttt{urgent};\,, U_3{=}\texttt{urgent};\mid I)}{\mathrm{P}(U_1{=}\texttt{urgent};\,, U_2{=}\texttt{urgent};\mid I)} \\
&\quad = \frac{0.030}{0.107} \\
&\quad \approx 28.0\%
\end{aligned}
$$

In this case the conditional probability 28.0% for $U_3{=}\texttt{urgent}$ is *higher* than the 20.0%, which was based only on knowledge I.

One possible intuitive explanation of these probability changes, *in the present scenario*, is that observation of two non-urgent cases makes the agent slightly more confident that "this is a day with few urgent cases". Whereas observation of two urgent cases makes the agent more confident that "this is a day with many urgent cases".

The diversity of inference scenarios

In general we cannot say that the probability of a particular value (such as `urgent` in the scenario above) will decrease or increase as similar or dissimilar values are observed. Nor can we say how much the increase or decrease will be.

In a different situation the probability of `urgent` could actually **increase** as more and more `non-urgent` cases are observed. Imagine, for instance, a scenario where the agent initially knows that there are 10 urgent and 90 non-urgent cases ahead (maybe these 100 patients have already been gathered in a room). Having observed 90 non-urgent cases, the agent will give a much higher, in fact 100%, probability that the next case will be an urgent one. Can you see intuitively why this conditional degree of belief must be 100%?

The differences among scenarios are reflected in differences in joint probabilities, from which the conditional probabilities are calculated. One particular joint probability can correspond to a scenario where observation of a value *increases* the degree of belief in subsequent instances of that value. Another particular joint probability can instead correspond to a scenario where observation of a value *decreases* the degree of belief in subsequent instances of that value.

All these situations are, in any case, correctly handled with the four fundamental rules of inference and the formula for conditional probability derived from them!

Exercises

a. Using the same joint distribution above, calculate

$$\mathrm{P}(U_1{=}\texttt{urgent} \mid U_2{=}\texttt{non-urgent}\,, U_3{=}\texttt{non-urgent}\,, I)$$

that is, the probability that the *first* patient will require urgent care *given that the agent knows the second and third patients will not require urgent care.*

- Why is the value you obtained different from $\mathrm{P}(U_1{=}\texttt{urgent} \mid I)$?
- Describe a scenario in which the conditional probability above makes sense, and patients 2 and 3 still arrive after patient 1. That is, a scenario where the agent learns that patients 2 and 3 are non-urgent, but still doesn't know the condition of patient 1.

b. Do an analysis completely analogous to the one above, but with different background information J corresponding to the following joint probability distribution for (U_1, U_2, U_3):

- If `urgent` appears 0 times out of 3: probability = 0%
- If `urgent` appears 1 times out of 3: probability = 24.5%
- If `urgent` appears 2 times out of 3: probability = 7.8%
- If `urgent` appears 3 times out of 3: probability = 3.1%

1. Calculate

$$\mathrm{P}(U_3{=}\texttt{urgent} \mid J)$$

and

$$\mathrm{P}(U_3{=}\texttt{urgent} | U_1{=}\texttt{non-urgent}, U_2{=}\texttt{non-urgent}, J)$$

and compare them.

2. Find a scenario for which this particular change in degree of belief makes sense.

17.5 Learning in the general case

Take the time to review the two sections above, focusing on the application and meaning of the two scenarios and calculations, and noting the similarities and differences:

- 〓 The calculations were completely analogous. In particular, the conditional probability was obtained as the quotient of a joint probability and a marginal one.
- ≠ In the first (urgency & transportation) scenario, information about one aspect of the situation changed the agent's belief about another aspect. The two aspects were different (transportation and urgency). Whereas in the second (three-patient) scenario, information about analogous occurrences of an aspect of the situation changed the agent's belief about a further occurrence.

A third scenario is also possible, which combines the two above. Consider the case with three patients, where each patient can require `urgent` care or not, and can be transported by `ambulance`, `helicopter`, or `other` means. To describe this situation, introduce three pairs of quantities, which together form the joint quantity

$$(U_1, T_1,\ U_2, T_2,\ U_3, T_3)$$

whose symbols should be obvious. This joint quantity has $(2 \cdot 3)^3 = 216$ possible values, corresponding to all urgency & transportation combinations for the three patients.

Given the joint probability distribution for this joint quantity, it is possible to calculate all kinds of conditional probabilities, and therefore consider all the possible ways the agent may learn new information. For instance, suppose the agent learns this:

- the first two patients have not required urgent care
- the first patient was transported by ambulance
- the second patient was transported by other means
- the third patient is arriving by ambulance

and with this learned knowledge, the agent needs to infer whether the third patient will require urgent care. The required conditional probability is

$$\mathrm{P}(U_3{=}\texttt{urgent} \mid T_3{=}\texttt{ambulance}\,, U_1{=}\texttt{non-urgent}\,, T_1{=}\texttt{ambulance}\,, U_2{=}\texttt{non-urgent}\,, T_2{=}\texttt{other}\,, I)$$

$$= \frac{\mathrm{P}(U_3{=}\texttt{urgent;}\,, T_3{=}\texttt{ambulance;}\,, U_1{=}\texttt{non-urgent;}\,, T_1{=}\texttt{ambulance;}\,, U_2{=}\texttt{non-urgent;}\,, T_2{=}\texttt{other;} \mid I)}{\mathrm{P}(T_3{=}\texttt{ambulance;}\,, U_1{=}\texttt{non-urgent;}\,, T_1{=}\texttt{ambulance;}\,, U_2{=}\texttt{non-urgent;}\,, T_2{=}\texttt{other;}\,, I)}$$

and is calculated in a way completely analogous to the ones already seen.

All three kinds of inference scenarios that we have discussed occur in data science and engineering. In machine learning, the second scenario is connected to "unsupervised learning"; the third, mixed scenario to "supervised learning". As you just saw, the probability calculus "sees" all of these scenarios as analogous: information about something changes the agent's belief about something else. And the handling of all three cases is perfectly covered by the four fundamental rules of inference.

So let's write down the general formula for all these cases of learning.

Let's consider a more generic case of a joint quantity with component quantities X and Y. Their joint probability distribution is given. Each of these two quantities could be a complicated joint quantity by itself.

The conditional probability for $Y{=}y$, given that the agent has learned that X has some specific value x^*, is then

$$\mathrm{P}(Y{=}y \mid X{=}x^*\,, I) = \frac{\mathrm{P}(Y{=}y\,, X{=}x^* \mid I)}{\sum_{\upsilon} \mathrm{P}(Y{=}\upsilon\,, X{=}x^* \mid I)} \tag{17.1}$$

where the index υ runs over all possible values in the domain of Y.

17.6 Conditional probabilities as initial information

Up to now we have calculated conditional probabilities, using the derived formula (17.1), starting from the joint probability distribution, which we considered to be given. In some situations, however, an agent may initially possess not a joint probability distribution but **conditional probabilities** together with **marginal probabilities**.

As an example let's consider a variation of our next-patient scenario one more time. The agent has background information I_S that provides the following set of probabilities:

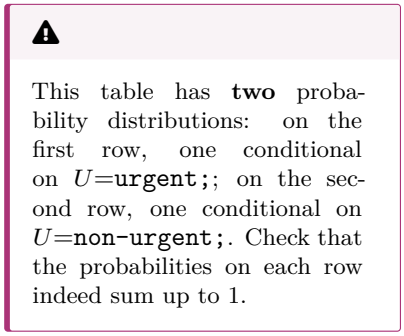
- Two conditional probability distributions $\mathrm{P}(T{=}\ldots \,|\, U{=}\ldots, I_S)$ for transportation T given urgency U, as reported in the following table:

Table 17.2: Probability distributions for transportation given urgency

$\mathrm{P}(T{=}t \,\|\, U{=}u, I_S)$		**transportation at arrival** T |		
		ambulance	helicopter	other
given	urgent	0.61	0.22	0.17
urgency $\|\, U$	non-urgent	0.21	0.01	0.78

- Marginal probability distribution $\mathrm{P}(U{=}\ldots \,|\, I_S)$ for urgency U:

$$\mathrm{P}(U{=}\texttt{urgent} \mid I_S) = 0.18\ , \quad \mathrm{P}(U{=}\texttt{non-urgent} \mid I_S) = 0.82 \tag{17.2}$$



This table has **two** probability distributions: on the first row, one conditional on $U{=}$`urgent;`; on the second row, one conditional on $U{=}$`non-urgent;`. Check that the probabilities on each row indeed sum up to 1.

With this background information, the agent can also compute all joint probabilities simply using the `and`-rule. For instance, the joint probability for U=`urgent` , T=`helicopter` is

$$
\begin{aligned}
&P(U\text{=}\texttt{urgent}\,, T\text{=}\texttt{helicopter} \mid I_{\mathrm{S}}) \\
&\quad = P(T\text{=}\texttt{helicopter} \mid U\text{=}\texttt{urgent}\,, I_{\mathrm{S}}) \cdot P(U\text{=}\texttt{urgent} \mid I_{\mathrm{S}}) \\
&\quad = 0.22 \cdot 0.18 = 3.96\%
\end{aligned}
$$

And from the joint probabilities, the marginal ones for transportation T can also be calculated. For instance

$$
\begin{aligned}
&P(T\text{=}\texttt{helicopter} \mid I_{\mathrm{S}}) \\
&\quad = \sum_u P(T\text{=}\texttt{helicopter}\,, U\text{=}u \mid I_{\mathrm{S}}) \\
&\quad = \sum_u P(T\text{=}\texttt{helicopter} \mid U\text{=}u\,, I_{\mathrm{S}}) \cdot P(U\text{=}u \mid I_{\mathrm{S}}) \\
&\quad = 0.22 \cdot 0.18 + 0.01 \cdot 0.82 \\
&\quad = 4.78\%
\end{aligned}
$$

Now suppose that the agent learns that the next patient is being transported by `helicopter`, and needs to forecast whether `urgent` care will be needed. This inference is the conditional probability $\mathrm{P}(U\text{=}\texttt{urgent} \mid T\text{=}\texttt{helicopter}\,, I_{\mathrm{S}})$, which can also be rewritten in terms of the conditional probabilities given initially:

$$
\begin{aligned}
&\mathrm{P}(U{=}\texttt{urgent} \mid T{=}\texttt{helicopter}\,, I_{\mathrm{H}}) \\
&\quad = \frac{\mathrm{P}(U{=}\texttt{urgent};\,, T{=}\texttt{helicopter};\mid I_{\mathrm{H}})}{\mathrm{P}(T{=}\texttt{helicopter};\mid I_{\mathrm{H}})} \\
&\quad = \frac{P(T{=}\texttt{helicopter};\mid U{=}\texttt{urgent};\,, I_{\mathrm{S}}) \cdot P(U{=}\texttt{urgent};\mid I_{\mathrm{S}})}{\sum_u P(T{=}\texttt{helicopter};\mid U{=}u\,, I_{\mathrm{S}}) \cdot P(U{=}u \mid I_{\mathrm{S}})} \\
&\quad = \frac{0.0396}{0.0478} \\
&\quad = 82.8\%
\end{aligned}
$$

This calculation was slightly more involved than the one in § 17.3, because in the present case the joint probabilities were not directly available. Our calculation involved the steps $T \mid U \;\longrightarrow\; T \wedge U \;\longrightarrow\; U \mid T$.

In this same scenario, note that if the agent were instead interested, say, in forecasting the transportation means knowing that the next patient requires urgent care, then the relevant degree of belief $\mathrm{P}(T{=}\ldots \mid U{=}\texttt{urgent}\,, I_{\mathrm{S}})$ would be immediately available and no calculations would be needed.

Let's find the general formula for this case, where the agent's background information is represented by conditional probabilities instead of joint probabilities.

Consider a joint quantity with component quantities X and Y. The conditional probabilities $\mathrm{P}(X{=}\ldots \mid Y{=}\ldots\,, I)$ and $\mathrm{P}(Y{=}\ldots \mid I)$ are encoded in the agent from the start.

The conditional probability for $Y{=}y$, given that the agent has learned that $X{=}x^*$, is then

$$
\mathrm{P}(Y{=}y \mid X{=}x^*\,, I) = \frac{\mathrm{P}(X{=}x^* \mid Y{=}y\,, I) \cdot \mathrm{P}(Y{=}y \mid I)}{\sum_v \mathrm{P}(X{=}x^* \mid Y{=}v\,, I) \cdot \mathrm{P}(Y{=}v \mid I)} \tag{17.3}
$$

In the above formula we recognize **Bayes's theorem** from § 9.4.

This formula is often exaggeratedly emphasized in the literature; some texts even present it as an "axiom" to be used in situations such as the present one. But we see that this formula is simply a by-product of the four fundamental rules of inference in a specific situation. An AI agent who knows the four fundamental inference rules, and doesn't know what "Bayes's theorem" is, will nevertheless arrive at this very formula.

17.7 Conditional densities

The discussion so far about conditional probabilities extends to conditional probability *densities*, in the usual way explained in §§15.3 and 16.2.

If X and Y are continuous quantities, the notation

$$\mathrm{p}(Y{=}y \mid X{=}x\,, I) = q$$

means that, given background information I and given the sentence "X has value between $x-\delta/2$ and $x+\delta/2$", the sentence "Y has value between $y-\epsilon/2$ and $y+\epsilon/2$" has probability $q\cdot\epsilon$, as long as δ and ϵ are small enough. Note that the small interval δ for X is *not* multiplied by the density q.

The relation between a conditional density and a joint density or a different conditional density is given by

$$\begin{aligned}
&\mathrm{p}(Y{=}y \mid X{=}x\,, I) \\
&\quad = \frac{\mathrm{p}(Y{=}y\,, X{=}x \mid I)}{\displaystyle\int_{\Upsilon} \mathrm{p}(Y{=}\upsilon\,, X{=}x \mid I)\,\mathrm{d}\upsilon} \\
&\quad = \frac{\mathrm{p}(X{=}x \mid Y{=}y\,, I)\cdot\mathrm{p}(Y{=}y \mid I)}{\displaystyle\int_{\Upsilon} \mathrm{p}(X{=}x \mid Y{=}\upsilon\,, I)\cdot\mathrm{p}(Y{=}\upsilon \mid I)\,\mathrm{d}\upsilon}
\end{aligned}$$

where Υ is the domain of Y.

17.8 Graphical representation of conditional probability distributions and densities

Conditional probability distributions and densities can be plotted in all the ways discussed in chapters 15 and 16. If we have two quantities A and B, often we want to compare the different conditional probability distributions for A, conditional on different values of B:

- P(A=... | B=`one-value` , I),
- P(A=... | B=`another-value` , I),
- ...

and so on. This can be achieved by representing them by overlapping line plots, or side-by-side scatter plots, or similar ways.

In § 16.3 we saw that if we have the scatter plot for a joint probability *density*, then from its points we can often obtain a scatter plot for its marginal densities. Unfortunately no similar advantage exists for the conditional densities that can be obtained from a joint density. In theory, a conditional density for Y, given that a quantity X has value in some small interval δ around x, could be obtained by only considering scatter-plot points having X coordinate in a small interval between $x-\delta/2$ and $x+\delta/2$. But the number of such points is usually too small and the resulting scatter plot could be very misleading.

Study reading

- §5.4 of *Risk Assessment and Decision Analysis with Bayesian Networks*
- §§12.2.1, 12.3, and 12.5 of *Artificial Intelligence*
- §§4.1–4.3 in *Medical Decision Making*
- §§5.1–5.5 of *Probability* – yes, once more!

Learning and conditional probability: a summary

The previous chapter 17 discussed many concepts that are important for what follows, and for artificial intelligence and machine learning in general. So let's stop for a moment to emphasize and point out some things to keep in mind.

- 👉 What does it means that an agent has "*learned*"? It means that the agent has acquired new information, knowledge, or data. But this acquisition is not just some passive memory storage. As a result of this acquisition, the agent *modifies its degrees of belief* about any inferences it needs to draw, and consequently may make different decisions ([ch.@sec-basic-decisions]).

  This change is an important aspect of learning. Think of when a person receives useful information; but the person, in actions or word, doesn't seem to make use of it. We typically say "that person hasn't learned anything".

- 👉 The relation between the acquired knowledge and the change in beliefs is perfectly represented and quantified by *conditional probabilities*. These probabilities take account of the acquired information in their conditionals (the right side of the bar " | "). And the probability calculus automatically determines how to calculate the modified belief based on this new conditional.

In other words, the probability calculus already has everything we need to deal with and calculate "learning". This is therefore *the* optimal, self-consistent way to deal with learning. We may use approximate versions of it in some situations, for instance when the computations would be too expensive. But we must keep in mind that such approximations are also deviations from optimality and self-consistency.

- ☛ The formula for conditional probability – that is, for the belief change corresponding to learning – involves and requires a *joint distribution* over several possibilities ([ch.@sec-prob-joint]). Therefore such distribution *must be somehow built into the agent* from the beginning, for the agent to be able to learn.

- ☛ The beliefs and behaviour arising from learning can be very different, depending on the context. For example, in some situations frequently observing a phenomenon may *increase* an agent's belief in observing that phenomenon again; but in other situations such frequent observation may *decrease* an agent's belief instead. Both kinds of behaviour can make sense in their specific circumstances.

  These differences in behaviour are also encoded in the *joint distribution* built into the agent.

- ☛ The formula for belief change arising from learning is amazingly flexible and universal: it holds whether the agent is learning about different kinds of quantities or about past instances of similar quantities.

  From a machine-learning point of view, this must therefore be the formula underlying the use of "features" by a classifier, as well as its "training".

This formula is moreover extremely simple in principle: it only involves addition and division! Computational difficulties arise from the huge amount of terms that may need to be added in specific data-science problems, not because of complicated mathematics. (A data engineer should keep this in mind, in case new hardware technologies may make it possible to deal with larger number of terms.)

18 Information, relevance, independence, association

18.1 Independence of sentences

In an ordinary situation represented by background information I, if you have to infer whether a coin will land heads, then knowing that it is raining outside has no impact on your inference. The information about rain is **irrelevant** for your inference. In other words, your degree of belief about the coin remains the same if you include the information about rain in the conditional.

In probability notation, representing "The coin lands heads" with H, and "It rains outside" with R, this irrelevance is expressed by this equality:

$$\mathrm{P}(H \mid I) = \mathrm{P}(H \mid R\,, I)$$

More generally two sentences A, B are said to be **mutually irrelevant** or **informationally independent given knowledge** I if any one of these three conditions holds:

- $\mathrm{P}(A \mid B\,, I) = \mathrm{P}(A \mid I)$
- $\mathrm{P}(B \mid A\,, I) = \mathrm{P}(B \mid I)$
- $\mathrm{P}(A\,, B \mid I) = \mathrm{P}(A \mid I) \cdot \mathrm{P}(B \mid I)$

"independ**E**nt" is written with an **E**, not with an **A**.

These three conditions turn out to be *equivalent* to one another. In the first condition, $\mathrm{P}(A \mid B, I)$ is undefined if $\mathrm{P}(B \mid I) = 0$, but in this case independence still holds; analogously in the second condition.

⚠ Irrelevance is not absolute and is not a physical notion

- Irrelevance or independence is not an absolute notion, but **relative to some background knowledge**. Two sentences may be independent given some background information, and **not** independent given another.
- Independence as defined above is an **informational** or **logical**, not physical, notion. It isn't stating anything about physical dependence between phenomena related to the sentences A and B. It's simply stating that information about one does not affect an agent's beliefs about the other.

18.2 Independence of quantities

The notion of irrelevance of two sentences can be generalized to quantities. Take two quantities X and Y. They are said to be **mutually irrelevant** or **informationally independent given knowledge I** if any one of these three equivalent conditions holds *for all possible values x of X and y of Y*:

- $\mathrm{P}(X{=}x \mid Y{=}y\,, I) = \mathrm{P}(X{=}x \mid I) \quad \text{all } x, y$
- $\mathrm{P}(Y{=}y \mid X{=}x\,, I) = \mathrm{P}(Y{=}y \mid I) \quad \text{all } x, y$
- $\mathrm{P}(X{=}x\,, Y{=}y \mid I) = \mathrm{P}(X{=}x \mid I) \cdot \mathrm{P}(Y{=}y \mid I) \quad \text{all } x, y$

Note the difference between independence of *two sentences* and independence of *two quantities*. The latter independence involves not just two, but many sentences: as many as the combinations of values of X and Y.

In fact it may happen that for some particular values x^* of X and y^* Y the probabilities become independent:

$$\mathrm{P}(X{=}x^* \mid Y{=}y^*, I) = \mathrm{P}(X{=}x^* \mid I)$$

while at the same time this equality does *not* occur for other values. In this case the quantities X and Y are *not* independent given information I. The general idea is that two quantities are independent if knowledge about one of them cannot change an agent's beliefs about the other, *no matter what their values might be*.

Irrelevance is not absolute and not physical

- Also in this case, irrelevance or independence is not an absolute notion, but **relative to some background knowledge**. Two quantities may be independent given some background information, and **not** independent given another.
- Also in this case, independence is an **informational** or **logical**, not physical, notion. It isn't stating anything about physical dependence between phenomena related to the quantities X and Y. It's simply stating that information about one quantity does not affect an agent's beliefs about the other quantity.

Exercise

Consider our familiar next-patient inference problem with quantities urgency U and transportation T. Assume a different background information J that leads to the following joint probability distribution:

$\mathrm{P}(U{=}u, T{=}t \mid J)$		**transportation at arrival** T		
		ambulance	helicopter	other
urgency U	urgent	0.15	0.08	0.02
	non-urgent	0.45	0.04	0.26

- Calculate the marginal probability distribution $\mathrm{P}(U \mid J)$ and the conditional probability distribution $\mathrm{P}(U \mid T{=}\texttt{ambulance}, J)$, and compare them. Is the value $T{=}\texttt{ambulance}$ relevant for inferences about U?
- Calculate the conditional probability distribution $\mathrm{P}(U \mid T{=}\texttt{helicopter}, J)$, and compare it with the marginal $\mathrm{P}(U \mid J)$. Is the value $T{=}\texttt{helicopter}$ relevant for inferences about U?
- Are the quantities U and T independent, given the background knowledge J?

18.3 Information and uncertainty

The definition of irrelevance given above appears to be very "black or white": either two sentences or quantities are independent, or they aren't. But in reality there is no such dichotomy. We can envisage some scenario I where for instance the probabilities $\mathrm{P}(Y{=}y \mid X{=}x, I)$ and $\mathrm{P}(Y{=}y \mid I)$ are extremely close in value, although not exactly equal:

$$\mathrm{P}(Y{=}y \mid X{=}x, I) = \mathrm{P}(Y{=}y \mid I) + \delta(x, y)$$

with $\delta(x, y)$ very small. This would mean that knowledge about X modifies an agent's belief just a little. And depending on the situation such modification could be unimportant. In this situation the two quantities would be "independent" for all practical purposes. Therefore there really are *degrees of relevance*, rather than a dichotomy "relevant vs irrelevant".

This suggests that we try to quantify such degrees. This quantification would also give a measure of how "important" a quantity can be for inferences about another quantity.

This is the domain of **Information Theory**, which would require a course by itself to be properly explored. In this chapter we shall just get an overview of the main ideas and notions of this theory.

For the extra curious

- *Information Theory, Inference, and Learning Algorithms*
- *Elements of Information Theory*

-
-

18.4 Exploring “importance”: some scenarios

18.4.1 First two required properties: a lottery scenario

Clue A: ✔ ✔ ✔ ✔ ? ?

Clue B: ✔ ✔ ✔ ? ✔ ?

Clue C: ? ? ? ✔ ✔ ✔

Scenario 1: choose one clue

Scenario 2: discard one clue

Scenario 3: discard one more clue

Importance is context-dependent

It doesn't make sense to ask which aspect or feature is "most important" if we don't specify the context of its use. Important if *used alone*? Important if *used with others*? and *which* others?
Depending on the context, an importance ranking could be completely reversed. **A quantitative measure of "importance" must therefore take the context into account**.

Importance is non-additive

A quantitative measure of importance cannot be *additive*, that is, it cannot quantify the importance of two or more features as the sum of their individual importance.

18.4.2 Third required property: A two-quantity scenario

Table 18.2: Example conditional distribution for two discrete quantities

1.

2.

3.

The importance of a quantity depends on its probability distribution

The importance of a quantity is not only determined by the relation between its possible values and what we need to infer, but also by the probability with which its values can occur.
A quantitative measure of "importance" of a quantity must therefore take the probability distribution for that quantity into account.

18.5 Entropies and mutual information

-
-
-

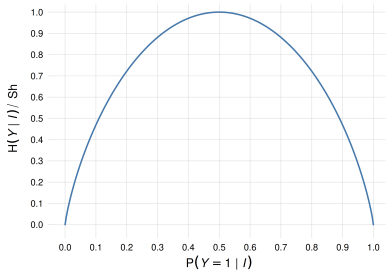
interna-
tional standards

18.5.1 Shannon entropy

[1]

[1]With the logarithm is in base 10, the unit is the *hartley* (Hart); with the natural logarithm, the unit is the *natural unit of information* (nat). 1 Sh ≈ 0.301 Hart ≈ 0.693 nat.

-
-
-
-
-



18.5.2 Conditional entropy

2

[2]or “equivocation” according to ISO standard.

$$\mathrm{H}(Y \mid X\,, I) := -\sum_{x}\sum_{y} \mathrm{P}(X{=}x \mid I) \cdot \mathrm{P}(Y{=}y \mid X{=}x\,, I)\ \log_2 \mathrm{P}(Y{=}y \mid X{=}x\,, I)\ \mathrm{Sh}$$

Context-dependence

Non-additivity

Probability-awareness

-
- 18.2
-

18.2

1.

2.

3.

18.5.3 Mutual information

[3]

[4]

$$\mathrm{H}(Y : X \mid I) := \sum_x \sum_y \mathrm{P}(Y{=}y\,, X{=}x \mid I)\ \log_2 \frac{\mathrm{P}(Y{=}y\,, X{=}x \mid I)}{\mathrm{P}(Y{=}y \mid I) \cdot \mathrm{P}(X{=}x \mid I)}\ \mathrm{Sh}$$

- 18.2

[3] There's no contradiction with the second remarkable property previously discussed: in this case the maximal value that the conditional entropy can take is zero.

[4] or "mean transinformation content" according to ISO standard.

-
-
-

18.2

1.

2.

3.

18.5.4 Uses

For the extra curious

- *Information Theory, Inference, and Learning Algorithms*
- *Probability and Information Theory, with Applications to Radar*

⚠ Mutual information is superior to the correlation coefficient

The Pearson correlation coefficient is actually a very poor measure of correlation or association. It is more a measure of "linearity" than correlation. It can be very dangerous to rely on in data-science problems, where we can expect non-linearity and peculiar associations in large-dimensional data. The Pearson correlation coefficient is widely used not because it's good, but because of (1) computational easiness, (2) intellectual inertia.

📖 Study reading

- Chapter 8 of *Information Theory, Inference, and Learning Algorithms*
- §12.4 of *Artificial Intelligence*

18.6 Utility Theory to quantify relevance and importance

2.4

19 Third connection with machine learning

11

$D_{\text{outcome?}}[0ex]D_N \wedge \cdots \wedge D_{2\text{ training data?}} \wedge D_1 \wedge [0ex]I_{\text{? architecture?}})$

12 13

-
-

-

$Z_{N+1}{=}z_{N+1}$ [0ex] $Z_N{=}z_N$, $\cdots$, $Z_2{=}z_2$, $Z_1{=}z_1$, [0ex] l_2

Nutdome? training data? architecture?

17.4

Figure 19.1: label = “Saitama”

Part V

Data II

20 Populations and variates

20.1 Collections of similar quantities: motivation

17.4

- **Stock exchange**
  closing price

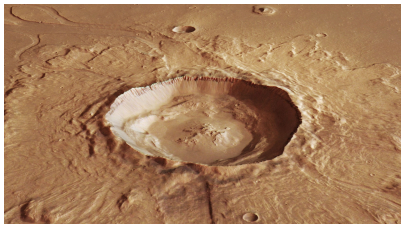
+; −;

- Mars **prospecting**

haematite

Y; N;

- **Glass forensics**



refrac-
tive index
Cal-
cium
Sili-
con

20.2 Units, variates, statistical populations

-
-
-

-
-
-
-

[1]

-
-
-
-
-

12.1.1

[1]The term *features* is frequently used in machine learning

-
-

Exercises

- Which of the following descriptions does properly define a statistical population? explain why it does or does not.
    1. People.
    2. Electronic components produced in a specific assembly line, since the line became operational until its discontinuation, and measured for their electric resistance, with possible values in $[0\ , \infty\ ]$, and for their result on a shock test, with possible values $\{\texttt{pass;}, \texttt{fail;}\}$.
    3. People born in Norway between 1st January

1990 and 31st December 2010.

4. The words contained in all websites of the internet.
5. Rocks, of volume between $1\,\mathrm{cm}^3$ and $1\,\mathrm{m}^3$, found in the Schiaparelli crater (as defined by contours on a map), and tested to contain haematite, with possible values {`Y;`, `N;`}.

- Browse some datasets at the UC Irvine Machine Learning repository. Each dataset is a statistical population. The variate in most of these populations is a joint variate (to be discussed below), that is, a collection of several variates.

  Examine and discuss the specification of some of those datasets:

  – Is it well-specified what constitutes a "unit"? Are the criteria for including or excluding datapoints, their origin, and so on, well explained?
  – Are the variates well-defined? Is it explained what they mean, how they were measured, what is their domain, and so on?

⚠ Subtleties in the notion of statistical population

- A statistical population is only a conceptual device for simplifying and facing some decision or inference problem. There is no objectively-defined population "out there".

  Any entity, object, person, and so on has some characteristics that makes it completely unique (say, its space-time coordinates). Otherwise we wouldn't be able to distinguish it from other entities. From this point of view any entity is just a one-member population in itself. If we consider two or more entities as being "similar" and belonging to the same pop-

ulation, it's because we have decided to disregard some characteristics of these entities, and only focus on some other characteristics. This decision is arbitrary, a matter of convention, and depends on the specific inference and decision problem.

To test whether an entity belongs to a given population, we have to check whether that entity satisfies the agreed-upon definition of that population.

- Any physical entity, object, person, etc. can be a "unit" in very different and even statistical populations. For instance, a $100\,\mathrm{cm}^3$ rock found in the Schiaparelli crater on Mars could be a unit in these populations:

  A. Rocks, of volume between $1\,\mathrm{cm}^3$ and $1\,\mathrm{m}^3$, found in the Schiaparelli crater and tested for haematite

  B. Rocks, of volume between $10\,\mathrm{cm}^3$ and $200\,\mathrm{cm}^3$, found in the Schiaparelli crater and tested for haematite

  C. Rocks, of volume between $10\,\mathrm{cm}^3$ and $200\,\mathrm{m}^3$, found in any crater on any planet of the solar system, and tested for haematite

  D. Rocks, of volume between $1\,\mathrm{cm}^3$ and $1\,\mathrm{m}^3$, found in the Schiaparelli crater and measured for the magnitude of their magnetic field.

  Note the following differences. Populations A, B, C above have the same variate but differ in their definition of "unit". Populations A and D have the same definition of unit but different variates. Population B is a subset of population A: they have the same variate, and any unit in B is also a unit in A; but not every unit in A is also a unit in B. Populations A and C have some overlap: they have the same variate, and some units of A are also units of C, and vice versa.

20.3 Populations with joint variates

13.1

glass forensics

-
-
  - *R*efractive *I*ndex
  - *Ca*lcium
  - *Si*licon
  -

Table 20.1: Glass fragments

4 4 4 4

- Remember the difference between *variate* and *quantity*, discussed previously. Consider the population of glass fragments introduced above, and suppose I say "Ca=8.1". Can you check if what I said is true? No, because you don't know which unit I'm referring to.

  The variate *for a specific unit* is a quantity instead. We can indicate this by appending the unit label to the variate symbol, as we did with "Ca_4" above. If I tell you "Ca_4=8", you can check that what i said is false; therefore Ca_4 is a *quantity*.

- The units' IDs don't need to be consecutive numbers; in fact they don't even need to be numbers: any label that completely distinguishes all units will do.

Exercises

- Download the dataset[2] `income_data_nominal_nomissing.csv` (4 MB):
  - How many variates does this population have?
  - What types of variate (binary, nominal, etc.) do they seem to be?

  - What are their domains?

- Explore datasets from the UC Irvine Machine Learning Repository, answering the three questions above.

[2]This is an adapted version of the UCI "adult-income" dataset

21 Statistics

21.1 What's the difference between Probability Theory and Statistics?

Probability theory

8.1

Statistics

-

-

James Clerk Maxwell

-
-

For the extra curious

Maxwell explains the statistical method and its use in the molecular description of matter:

- *Introductory Lecture on Experimental Physics*
- *Molecules*

21.2 Frequencies and frequency distributions

12.2

13

In the following we shall call relative frequencies simply "frequencies", and explicitly use the word "absolute" when we speak about absolute frequencies.

12.1.1

Exercise

Consider the statistical population defined as follows:

- *units:* the bookings at a specific hotel during a specific time period
- *variate:* the market segment of the booking
- *variate domain:* the set of five values {`Aviation;`, `Complementary;`, `Corporate;`, `Offline;`, `Online;`}

The population data is stored in the file `hotel_bookings-market.csv`. Each row of the file corresponds to a unit, and lists the unit id (this is not a variate in the present population) and the market segment.
Use any method you like (a script in your favourite programming language, counting by hand, or whatever) to answer these questions:

- What is the size of the population?
- What are the absolute frequencies of the five values?
- What are their relative frequencies?
- Which units have the value `Corporate`?

21.2.1 Differences between frequencies and probabilities

1. •

-

2. -

-

3. -
   -

4. -

-

21.3 Joint frequencies

-
-
-
-

Table 21.1: Income

Exercise

Try to write a function that takes as input a dataset with a small number of variates and outputs the joint frequency distribution for all combinations of variate values. The best output format is a multidimensional array having one dimension per variate, and for each dimension a length equal to the number of possible values of that variate. The value of the array in each cell is the corresponding frequency.

For instance, consider the case of the income dataset above but *without the age variate*. The output of the function would then be an array with $5 \times 2 \times 2$ dimensions

21.4 Marginal frequencies

X Y

y Y

y

Y

$Y=y$

x

16.1

$Y=y$ $Y=y$ $X=x$

x

21.1

Exercises

- Download again the dataset `income_data_nominal_nomissing.csv`:
  - Calculate the marginal frequencies of some of its variates.
  - Does any variate have a value appearing with *marginal absolute frequency* equal to 1?

21.5 Summary statistics

21.5.1 Mode

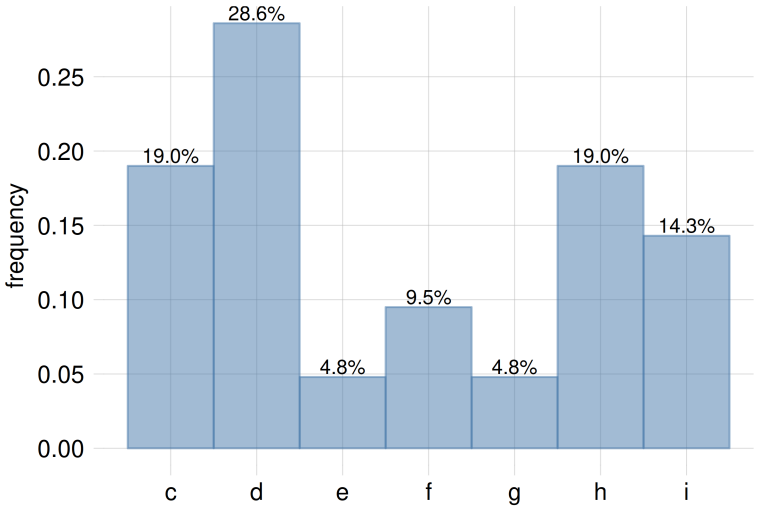
Be careful in relying too much on the "mode" for a continuous quantity. Continuous quantities can be transformed in a one-to-one way into other, equivalent ones; and such a transformation also give the equivalent frequency or probability *density* for the new quantity. **There is no general relationship** between the modes of the densities for the two equivalent quantities. In fact, the density for one quantity can have one mode, whereas the density for the equivalent quantity can have no mode, or many modes. This is true for all kinds of distributions represented by densities, for example a continuous distribution of energy.

21.5.2 Median and quartiles

12.2

For the extra curious

Some paradoxes, errors, and resolutions concerning the spectral optimization of human vision


28.6%
19.0%
19.0%
14.3%
9.5%
4.8%
4.8%
0.25
0.20
0.15
0.10
0.05
0.00
frequency
c
d
e
f
g
h
i

21.5.3 Mean and standard deviation

21.5.4 Uses and pitfalls

2

Study reading

- §2.6 of *Risk Assessment and Decision Analysis with Bayesian Networks*
- § "**The median estimate**" of *Meaningful expression of uncertainty in measurement*
- *The Median Isn't the Message*

21.6 Outliers vs out-of-population units

Study reading

§2.1 of *Risk Assessment and Decision Analysis with Bayesian Networks*

For the extra curious

Ch. 21 of *Probability Theory*

22 Subpopulations and conditional frequencies

22.1 Subpopulations

-
-
  -
  -
  -

Table 22.1: Simplified glass-fragment population data