ADA511: Data science and data-driven engineering

Steffen Mæland
PierGianLuca Porta Mana

2023-07-06

Table of contents

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)

**WARNING: THIS IS A WORKING DRAFT. TEXT WILL CHANGE A LOT. MANY PASSAGES ARE JUST TEMPORARY, INCOHERENT, AND DISJOINTED.

To be written.

- Difference between car mechanic and automotive engineer
- "Engineering based on data" is just how engineering and science in general have been in the past 400 years or so. Nothing new there.
- The amount of available data has changed. This may lead to a reduction – or in some cases an increase – in uncertainty, and therefore to different solutions.
- Luckily the fundamental theory to deal with large amount of data is exactly the same to deal with small amounts. So the foundations haven't changed.

This course makes you acquainted with the foundations.

Part I

An invitation

1 Accept or discard?

Let's start with a question that could arise in a particular 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 with every newly produced component coming out of the line.
>
> The inspection device first makes some tests on the new 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.
>
> When a new electronic component is sold, the manufacturer has a net gain of 1\$. If the component fails within a year of use, however, the manufacturer incur 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.
>
> For a specific 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 or discard the new component?

First, try to give and motivate an answer.

This is not the real question of this exercise, however. In fact 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.

Very first exercise!

The purpose here is for you to do some introspection about your own reasoning. Then examine and discuss these 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 let's say that we `accept` the component. What happens?

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 that 10% probability that the component fails within a year. Different people do this with different imagination tricks. We can imagine, for instance, that this 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.

In each of the 10 imaginary repetitions in which 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\$. 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 say that we `discard` the component instead. What happens? In this case we don't need to invoke imaginary repetitions, but even if we do, it's clear that the manufacturer doesn't gain or lose anything – that is, the "gain" is 0\$ – in each and all of the repetitions.

The conclusion is that if in a situation like this we accept the component, then we'll lose 0.2\$ on average; whereas if we discard it, then on average we won't lose anything or gain anything.

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 of the failure cost: is it still best to discard? For instance, try with

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

   Feel free to get wild and do plots.

2. Identify the failure probability at which accepting the component doesn't lead to any loss or any gain, 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$`. This means no new component will ever fail. To decide in such a case we do not need imaginary repetitions; but **confirm** that we arrive at the same logical conclusion whether we reason through imaginary repetitions or not.

4. Consider this completely different problem:

   > A patient is examined by a brand-new medical diagnostics AI system.
   >
   > The AI first performs some clinical tests on the patient. The tests give an uncertain forecast of 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 the 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 the 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 the patient has 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, if we simply know that the probability of failure is less than 50%?

  Obviously not. We found that if the failure probability is 10% then it's best to discard; but if it's 5% then it's best to accept. In both cases the component was less likely to fail than not, but the decisions were different. Moreover, we found that the probability affected the loss if one made the non-optimal decision. Therefore:

  Knowledge of exact probabilities is absolutely necessary for making the best decision

- Is it enough if we simply know that failure leads to a cost? that is, that its gain is less than the gain for non-failure?

  Obviously not. The situation is similar to that with the probability. In the exercise we found that if the failure cost is 11$ then it's best to discard; but if it's 5$ then

it's best to accept. It's also best to accept if the failure cost is 11$ but the non-failure gain is 2$. Therefore:

Knowledge of the exact 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 found a clinical situation that's exactly analogous: there's uncertainty, there are gains and losses (of time rather than money), and the best decision depends on both.

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

Every data-driven engineering project is unique, with its unique difficulties and problems. But there are also problems common to all engineering projects.

In the scenarios 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 choice). Making a particular decision will lead to some consequences, some leading to a desired goal, others leading to something undesirable. The decision is difficult because its consequences are not known with certainty, given 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. But our explored scenarios also suggest that this problem-pattern has a sort of systematic solution method.

In this course we're going to focus on decision problems and their systematic solution method. We'll learn a framework and some abstract notions that allow us to frame and analyse this kind of problem, and we'll learn a universal set of

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

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 and tightly related on many different planes:

- We saw that *probability* values are essential in a decision problem. How do we find them? As you can imagine, *data* play an important part in their calculation. In our intro example, the failure probability must come from observations or experiments on similar electronic components.
- We saw that also the values of *gains and losses* are essential. *Data* play an important part in their calculation as well.
- *Data science* is based on the laws of *Decision Theory*. Here's an analogy: a rocket engineer relies on fundamental physical laws (balance of momentum, energy, and so on) for making a rocket work. Failure to account for those laws leads at best to sub-optimal solutions, at worst to disasters. As we shall see, the same is true for a data scientist and the rules of decision theory.
- *Machine-learning* algorithms, in particular, are realizations or approximations of the rules of *Decision Theory*. This is clear, for instance, considering that the main task of a machine-learning classifier is to decide among possible output labels or classes.
- The rules of *Decision Theory* are also the foundations upon which *artificial-intelligence* agents, which must make optimal inferences and decisions, are built.

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

For the extra curious

Decision theory in expert systems and artificial intelligence

@@ TODO add examples: algorithm giving outputs is a decision agent. @@ Include one with https://hjerterisiko.helsedirektoratet.no

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. These aspects can barely be planned or decided; but their fruits, once available, should be handled and used optimally – thus leading to a decision problem.

Artificial intelligence is proving to be a valuable aid in these more creative aspects too. This kind of use of AI is outside the scope of the present notes. Some aspects of this creativity-assisting use, however, 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, then 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. This also means that a priori we cannot blame an engineer for making an unsuccessful decision in a situation of uncertainty.

Imagine two persons, Henry and Tina, who must bet on "heads" or "tails" under the following conditions (but who otherwise don't get any special thrill from betting):

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

Henry chooses the first bet, on "heads". Tina chooses the second bet, on "tails". 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 in the second bet, without any possibility of at least winning more. Tina's decision was unlucky, and yet *rational*: the possibility and amount of winning was the same in the two bets, and she chose the bet with the least amount of loss. 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.

This example shows two points. First, "success" is generally not a good criterion to judge a decision under uncertainty; success can be the pure outcome of luck, not of smarts. Second, even if there is no method to determine which decision is successful, there is 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.

Let us emphasize, however, that we are not giving up on "success", or trading it for "optimality". Indeed we'll find that **Decision Theory automatically leads to the *successful* decision** in problems where uncertainty is not present or is irrelevant. It's a win-win. It's important to keep this point in mind:

> Aiming to find the solutions that are *successful* can make us *fail* to find those that are optimal when the successful ones cannot be determined.
> Aiming to find the solutions that are *optimal* makes us automatically find those that are *successful* when those can be determined.

We shall later witness this fact with our own eyes, and will take it up again 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?

If we find an optimal decision in regards to some outcome, it may still happen that the decision can be realized in several ways that are equivalent in regard to the outcome, but inequivalent in regard to time or resources. In the assembly-line scenario, for example, the decision `discard` could be carried out by burning, recycling, and so on. 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 in a specific kind of uncertain situations. Think for instance of an aeronautic engineer designing an autopilot system, or a software company designing an image classifier.

Decision Theory turns out to meet these 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 automated or AI devices

Decision Theory has a long history, going back to Leibniz in the 1600s and partly even to Aristotle in the −300s, and appearing in its present form around 1920–1960. What's remarkable about it is that it is not only *a* framework, but *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, any "alternative" framework may use different technical terminology and

rewrite mathematical operations in a different way, but it boils down to the same notions and operations of Decision Theory. So if you wanted to invent and use another framework, then either (a) it would lead to some irrational or illogical consequences, or (b) it would lead to results identical to Decision Theory's. 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. Because human decision-makers typically make irrational, sub-optimal, or biased decisions. That's exactly what we want to avoid and improve!

For the extra curious

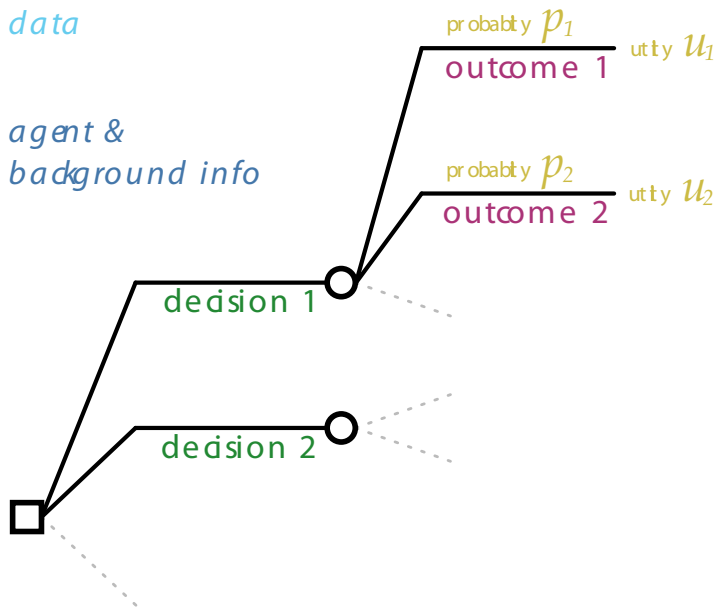
- *Judgment under uncertainty*
- *Heuristics and Biases*
- *Thinking, Fast and Slow*

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 *uncertainty node*, usually represented by a circle ●, from which the possible outcomes depart as lines. Each outcome leads to a particular utility value. The uncertainty of each outcome is quantified by a probability.

A basic decision problem is analysed in terms of these elements:

- **Agent**, and **background** or **prior information**. The agent is the person or device that has to make the decision. An agent always possess (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 and utilities of the outcomes, together with other available data and information. Since different agents typically have different background information, we shall somehow conflate agents and prior information.

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

- **Decisions**, also called **courses of actions**, 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.
- **Probabilities** for each of the outcomes. Their values typically depend on the background information, the decision, and the additional data.
- **Utilities**: the gains or losses associated with each of the possible outcomes. Their values also depend on the background information, the decision, and the additional data.
- **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.

Note that it is 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.

> Reading
>
> § 1.1.4 in *Artificial Intelligence*

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

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

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 Inference, utility, maximization

The solution of a basic decision-making problem can be roughly divided into three main stages: inference, utility assessment, and expected-utility maximization.

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.

For the extra curious

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

Expected-utility maximization is the final stage where the probabilities and gains or costs of the possible outcomes are combined, in order to determine the optimal decision.

Part II

Inference

4 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 in view of the test results were already given. In real decision problems, however, 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 goal 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", more precisely? What important aspects about inferences 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?

4.1 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 a patient, so as 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. From the current set of camera frames, the computer of a self-driving car needs to assess whether a particular patch of colours in the frames is a person, so as to slow down the car and stop.

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 reading of 100 mV. The electronic component turns out to be defective. An engineer wants to assess whether the voltage-test reading could have been 100 mV, if the component had not been defective.

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

K. From measurements of the Sun's energy output and of concentrations of various substances in the Earth's atmosphere over the past 500 000 years, and 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 perfectly, that is, without any uncertainty, based the data given explicitly or 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, their object or the data on which they are based have never happened before and will never happen again?

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

12. Are any of the inferences 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 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: 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**.

4.2 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 of 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, for example, 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 to prove some *theorem*, we must always start from some *axioms*. There are "inferences" – *tautologies* – that can be drawn without requiring others; but they are all trivial, such as "this component failed early, or it didn't". They are of little use in a real problem, although they have a deep theoretical importance.

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 anyone. In this case the final inference has a somewhat contingent character, and we accept it (as well as the

> For the extra curious
>
> *Mathematics: The Loss of Certainty*.

Figure 4.1: Sci-fi films like *The Matrix* ultimately draw on the fact that we must take some inferences for granted without further proof.

solution of any underlying decision problem) as the best available for the moment. This is partly the origin of the term "**model**".

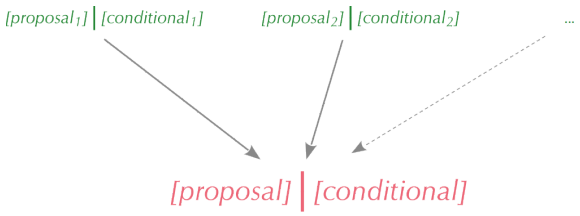
4.3 Basic elements of an inference

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

Every inference has an "object" – what is to be assessed – as well as data, information, or hypotheses on which it is based. We call **proposal**[1] the object of the inference, and **conditional**[2] what the inference is based upon. We separate them with a vertical bar[3] " | ", which can be pronounced *given* or *conditional on*:

$$[\textit{proposal}] \mid [\textit{conditional}]$$

We have seen that to calculate the probability for an inference, we must start from the probabilities of other inferences. A basic inference process therefore can be schematized like this:



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

[1] Johnson's (1924) terminology. Keynes (1921) uses "conclusion". Modern textbooks do not seem to use any specialized term.

[2] Modern terminology. Other terms used: "evidence", "premise", "supposal".

[3] Originally a solidus, introduced by Keynes (1921).

5 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 "items" 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 "items"?

5.1 The central components of knowledge representation

When speaking of "data", what comes to mind to many people is basically numbers or collections of numbers. Maybe numbers, then, could be used to represent all the variety of items exemplified above. This option, however, 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. So "data" is not just numbers: a number is not "data" unless there's an additional verbal, 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" (we imagine to know who's the patient and what was 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 essentially non-numeric data. Sure, in some situations we can represent it as "1", while "0" would represent "not recovered"; but the opposite convention could also be used, or the numbers "0.3" and "174". These numbers have intrinsically nothing to do with the clinician's "recovery" data.

But the examples above 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.

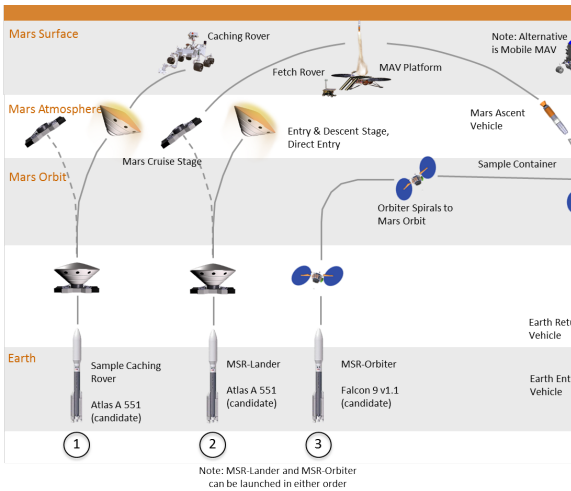
We shall therefore use **sentences**, also called **propositions** or **statements**,[1] to represent and communicate all the kinds of "items" that can be the proposal or conditional of an inference. In some cases we can of course summarize a sentence by a number, as a shorthand, when the full meaning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

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.

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 com-*

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



(From the *SMART* paper)

plex mission
- around p. 22 in *No More Band-Aids: Integrating FM into the Onboard Execution Architecture*
- § 2.1 in *Deliberation for autonomous robots: A survey*
- part IV in *Model-based programming of intelligent embedded systems and robotic space explorers*

5.2 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 we can determine whether it is `true` or `false`, at least in principle and in such a way that all interested receivers of the message would agree.

For instance, in most engineering contexts the phrase "This valve will operate for at least two months" is a sentence; whereas the phrase "Apples are much tastier than pears" is not, because it's a matter of personal taste – there's no objective criterion to determine its truth or falsity (however, the phrase "Rita finds apples tastier than pears" could be a sentence; its truth is found by asking Rita). In a data-science context, the phrase "The neural-network algorithm has better performance than the random-forest one" is *not* a sentence unless we have objectively specified what "*better*" means, for example by using a particular comparison metric.

Some expressions in fact, even involving technical terms, may appear to be sentences at first, but a deeper analysis may reveal 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 deter-

> For the extra curious
>
> Propositions

mine whether such an expression is true or false. Its truth turns out to be a matter of convention (also in Newtonian mechanics). The Theory of Relativity was born from this observation.

One sentence can be expressed by many 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.

> For the extra curious
>
> *On the electrodynamics of moving bodies*.

A sentence can contain numbers, pictures, and graphs.

Working with sentences, and keeping in mind that inference is about sentences, is important in several respects:

First, it 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 and engineering a solution, it is important to understand what information or outcomes the end user exactly wants. The "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, it 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 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 only be true or

false. Many inferences, especially when they concern numerical measurements, are actually collections of inferences. For example, an inference about the result of rolling a die actually consists of six separate inferences with the 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 without thinking about this detail. In real applications it can be useful, on some occasions, to pause and reduce an inference to its basic set of `true`/`false` inferences; this analysis may reveal contradictions in our inference. 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 § 4.1 in a formal way, as one or more inferences
>
> $$[\textit{proposal}] \mid [\textit{conditional}]$$
>
> where proposal and conditional are well-defined sentences.
> In ambiguous cases, use your judgement and motivate your choices.

5.3 Notation

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 such as the one above,

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

or even more compactly introducing physical symbols:

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

where v is a physical variable denoting the speed; or even writing simply

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

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

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

which means "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:

'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. These differences 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 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?

5.4 Connecting sentences

5.4.1 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 complex 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 instance, 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.

5.4.2 Connectives

How do we construct complex 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 encountered them already 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: $\neg$ example:

$\neg s$ = 'The component is broken after the shock test'

And: $\wedge$ example:

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

Or: $\vee$ example:

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

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

! Important subtleties of the connectives:

- There is *no strict correspondence* between the words "not", "and", "or" in natural language and the three connectives. For instance the `and` connective could 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 both the sentences it connects can be 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 complex 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."

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

Reading

Just 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.

5.5 "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

$$\text{'The component is defective'} \mid \text{'The voltage reading is 200 mV'}$$

is `true`.

This kind of information is very important because it often is the starting point from which to arrive at the final inferences we're interested in. We shall discuss it more in 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.

6 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 kind of inferences *truth inferences*. Mathematical inferences are a typical instance of this kind. You probably have some acquaintance with rules for drawing truth inferences, so we start from these.

6.1 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 heating test or the shock 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.

6.2 Analysis and representation of the problem

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

6.2.1 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)

6.2.2 Proposal

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

6.2.3 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.

6.2.4 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 heating test or the shock test. We write this as

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

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

$$s \mid f \wedge l$$

can*not* be `false`.

6.2.5 Target inference

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

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

6.3 Truth-inference rules

6.3.1 Deduction systems; a specific choice

Formal logic gives us a set of rules for correctly drawing sure inferences, *when such 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). 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.

$$\dfrac{\dfrac{}{l \wedge f \vdash \neg h \vee \neg s} \qquad \dfrac{}{s \vdash s}}{\neg s \wedge s \vdash}$$
$$l \wedge f \wedge s \vdash \neg h$$

Figure 6.1: The bottom formula is the target inference. Each line denotes the application of an inference rule, from one or more inferences above the line, to one below the line. The two formulae with no line above are our starting inference, and a tautology.

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

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

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 complex sentences X, Y, Z:

Rule for "not":

$$\mathrm{T}(\neg X \mid Z) + \mathrm{T}(X \mid Z) = 1 \tag{6.1}$$

Rule for "and":

$$\mathrm{T}(X \wedge Y \mid Z) = \mathrm{T}(X \mid Y \wedge Z) \cdot \mathrm{T}(Y \mid Z) = \mathrm{T}(Y \mid X \wedge Z) \cdot \mathrm{T}(X \mid Z) \tag{6.2}$$

Rule for "or":

$$\mathrm{T}(X \vee Y \mid Z) = \mathrm{T}(X \mid Z) + \mathrm{T}(Y \mid Z) - \mathrm{T}(X \wedge Y \mid Z) \tag{6.3}$$

Rule of self-consistency:

$$\mathrm{T}(X \mid X \wedge Z) = 1 \tag{6.4}$$

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.

6.3.2 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

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

One possibility is to work backwards from the target inference:

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

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 ap-

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

6.3.3 [*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 (6.1)–(6.4) 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 (6.1)–(6.4) 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.

6.4 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 § 5.1, are programmed according to these rules.

> **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 take a look at
>
> - *Logic in Computer Science*
> - *Mathematical Logic for Computer Science*
> - Natural Deduction Systems in Logic

7 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.

7.1 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 for "probability" are *degree of belief*, *plausibility*, *credibility*.

> **Avoid *likelihood* as a synonym for *probability***
>
> In technical discourse, "likelihood" means something different and is *not* a synonym of "probability", as we'll explain later.

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.

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

- **Probabilities are assigned to *sentences*.** We already discussed this point in § 5.3, 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'

  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 deciding how to set that output, the probability is 100%). 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 "$O = 100\,\mathrm{W}$" provided that the full sentence behind this kind of shorthand is understood.

- **Probabilities are agent- and context-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 that groundless and silly. We would be baffled if Alice assigned 50% to 'heads', because she saw the outcome was actually heads; we would hypothesize that 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 sub-optimal 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.

> 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[1] – or even cause deaths.

7.2 An unsure inference

Consider now the following variation of the trivial inference problem of § 6.1.

> This electric component had an early failure. If an electric component fails early, then at production it either didn't pass the heating test or didn't pass the shock test. The probability that it didn't pass

[1] This fact can be mathematically proven.

both tests is 10%. There's no reason to believe that the component passed the heating test, more than it passed the shock test.

The inspector wants to assess, also in this case, 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 drawn with certainty in this case. One way to argue this is to present two different scenarios that fit the data given 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?

7.3 Probability notation

For this inference problem we can't find a `true` or `false` final value. The truth-inference rules (6.1)–(6.4) 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 that the belief in the truth of the sentence is as strong as that in its falsity.

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

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

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

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

7.4 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 (6.1)–(6.4), extend also to probability-inference*. The only difference is that they now hold for all values in the range $[0, 1]$, rather than for 0 and 1 only.

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 the 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 complex sentences X, Y, Z:

It is amazing that **ALL** inference is nothing else but a repeated application of these four rules – billions of times or more in some cases. 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, at bottom, just a study of how to apply these rules in 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

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

Self-consistency rule

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

📖 Reading

- *Probability, Frequency and Reasonable Expectation*
- Ch. 2 of *Bayesian Logical Data Analysis for the Physical Sciences*
- §§ 1.0–1.2 of *Data Analysis*
- Feel free to skim through §§ 2.0–2.4 of *Probability Theory*

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

7.5 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, find what are its proposal and conditional, and which starting inferences are given in the problem.

7.5.1 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.

7.5.2 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

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

7.5.3 Starting inferences

We are told that if an electric component fails early, then at production it didn't pass either the heating test or the shock test. Let's write this as

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

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

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

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:

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

Note this interesting situation: 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.

7.5.4 Final inference

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

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

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?

7.6 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*[2] 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

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

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

Exercise

Try to prove the two constraints above.

7.7 Derived rules

The fundamental rules above are in principle all we need to use to draw inferences from other inferences. But from them it is possible to derive some "shortcut" rules.

7.7.1 Boolean algebra

First, it is possible to show that all rules you may know from Boolean algebra *are a consequence of the fundamental rules*. So we can always make the following convenient replacements anywhere in a probability expression:

Boolean algebra

$$\neg\neg X = X \qquad X \wedge X = X \vee X = X$$

$$X \wedge Y = Y \wedge X \qquad X \vee Y = Y \vee X$$

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

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

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

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

Suppose we have a set of n sentences $\{Y_1, Y_2, \ldots, Y_n\}$ having these two properties:

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

  $$\mathrm{P}(Y_1 \wedge Y_2 | Z) = 0\,, \quad \mathrm{P}(Y_1 \wedge Y_3 | Z) = 0\,, \quad \ldots\,, \quad \mathrm{P}(Y_{n-1} \wedge Y_n | Z) = 0$$

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

$$\mathrm{P}(Y_1 \vee Y_2 \vee \cdots \vee Y_n \mid Z) = 1$$

Then the probability of a sentence X, conditional on Z, is equal to a combination of probabilities conditional on $Y_1, Y_2, \ldots$:

Derived rule: extension of the conversation

$$\mathrm{P}(X \mid Z) = \mathrm{P}(X \mid Y_1 \wedge Z) \cdot \mathrm{P}(Y_1 \mid Z) + \mathrm{P}(X \mid Y_2 \wedge Z) \cdot \mathrm{P}(Y_2 \mid Z) + \cdots + \mathrm{P}(X \mid Y_n \wedge Z) \cdot \mathrm{P}(Y_n \mid Z)$$

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 probabilities of that sentence conditional on several auxiliary sentences – often representing hypotheses that exclude one another, and of which we know at least one is true. The name **extension of the conversation** for this derived rule comes from the fact that we are able to call the additional sentences into play.

This situation occurs very often in concrete applications, especially in problems where the probabilities of several competing hypotheses have to be assessed.

The next derived rule is used extremely often, so we discuss it separately.

7.8 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 where two sentences Y, X appear in the proposal and the conditional, with one where they are exchanged:

Derived 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.

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

Bayes's theorem is extremely useful when we want to assess the probability of a sentence, typically a hypothesis, given some conditional, typically data; and we can easily 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 does not always need to be a "hypothesis", and X "data".

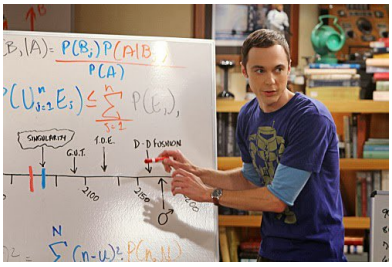


Figure 7.1: Bayes's theorem guest-starring in *The Big Bang Theory*

7.8.1 Combining with the extension of the conversation

Bayes's theorem is often 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:

> **Derived 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.

7.8.2 Many facets

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 according to which we "update our beliefs". The meaning of this proclamation is the following. Let's say that at some point Z represents all your knowledge. Your 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, let's say that you get to know – maybe thanks to an observation you made – that the sentence X is true. Your whole knowledge at that point is represented no longer by Z, but by $X \wedge Z$. Your degree of belief about Y is then given by the value of $\mathrm{P}(Y \mid X \wedge Z)$. Bayes's theorem allows you to find your degree of belief about Y conditional on your new state of knowledge, from the one conditional on your 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. Is this result reasonable?

Reading

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

  https://www.youtube.com/watch?v=HZGCoVF3YvM

7.9 consequences of not following the rules

@@ §12.2.3 of AI

- *Exercise: Monty-Hall problem & variations*
- *Exercise: clinical test & diagnosis*

7.10 Remarks on terminology and notation

7.10.1 Likelihood

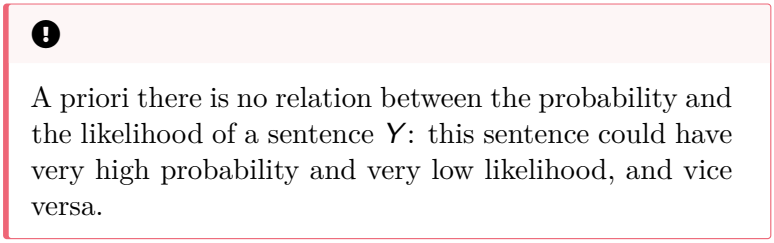
In everyday language, "likely" is often a synonym of "probable", and "likelihood" of "probability". But in technical questions about probability, inference, and decision-making, "likelihood" has a very different meaning. Keep in mind this important difference of definition:

$P(Y \mid X)$ is:

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

Let's express this also in a different way:

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



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".

7.10.2 Omitting background information

In the analyses of the inference examples of § 6.1 and § 7.2 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 background information.

In many concrete inference problems the background information usually stays there 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 when you see in books probabilities "$P(x)$" 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 proceed, for example using the rule of extension of the conversation. If the background information was not explicitly written, this may lead to confusion and mistakes. For instance 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 different agents's degrees of belief.

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

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

7.10.3 "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, but it has some shortcomings:

- As discussed in § 7.1, 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.
- Very often the object (proposal) of a probability is not a "variable": it is actually a *constant* value that is simply unknown.
- What does "random" (or "chance") mean? Good luck finding an understandable and non-circular definition in texts that use that word; strangely enough, they 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 7.2: James Clerk Maxwell is one of the main founders of statistical mechanics and kinetic theory (and electromagnetism). Yet he never used the word "random" in his technical writings. Maxwell is known for being very clear and meticulous with explanations 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.

8 Probability distributions

(Make sure you're familiar with § 12 before you begin.)

8.1 Distribution of probabilities 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 regarding the true value. 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 271 K", and so on up to 275 K. We can abbreviate these sentences, denoting the temperature with T, as

$$T = 271\,\mathrm{K}\,, \quad T = 272\,\mathrm{K}\,, \quad T = 273\,\mathrm{K}\,, \quad T = 274\,\mathrm{K}\,, \quad T = 275\,\mathrm{K}\,.$$

We recognize these as *mutually exclusive* and *exhaustive* sentences.

Our belief about the quantity is then expressed by a collection of probabilities, conditional on the agent's state of knowledge I. For instance:

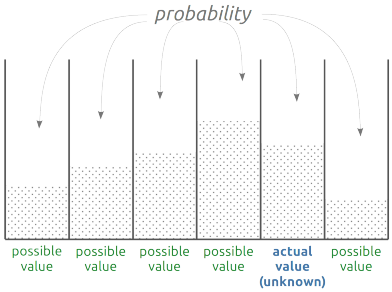
$$\begin{aligned}
\mathrm{P}(T = 271\,\mathrm{K} \mid I) &= 0.04 \\
\mathrm{P}(T = 272\,\mathrm{K} \mid I) &= 0.10 \\
\mathrm{P}(T = 273\,\mathrm{K} \mid I) &= 0.18 \\
\mathrm{P}(T = 274\,\mathrm{K} \mid I) &= 0.28 \\
\mathrm{P}(T = 275\,\mathrm{K} \mid I) &= 0.40
\end{aligned}$$

that sum up to one:

$$\begin{aligned}
&\mathrm{P}(T = 271\,\mathrm{K} \mid I) + \cdots + \mathrm{P}(T = 275\,\mathrm{K} \mid I) \\
&\quad = 0.04 + 0.10 + 0.18 + 0.28 + 0.40 \\
&\quad = 1
\end{aligned}$$

This collection of probabilities is called a **probability distribution**.

> **What's "distributed"?**
>
> It's the *probability* that's distributed among the possible values, not the quantity, as illustrated in the side picture. The quantity cannot be "distributed": it has one, definite value, which is however unknown to us.



> **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 § 7, 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.

8.2 Discrete probability distributions

8.2.1 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 the function

$$\mathrm{P}(T \mid I) = \frac{(T/\mathrm{K} - 269)^2}{90} \quad \text{(rounded to two decimals)}$$

A graphical representation is often helpful to detect features, peculiarities, and even inconsistencies in one or more probability distributions.

8.2.2 Histograms and area-based representations

A probability distribution for a nominal, ordinal, and 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. Typically the rectangles are contiguous. 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 the unit of area. 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.

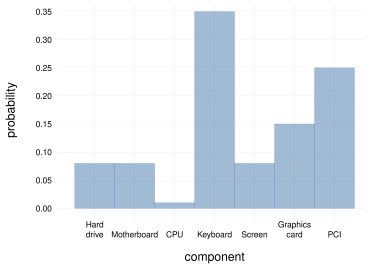


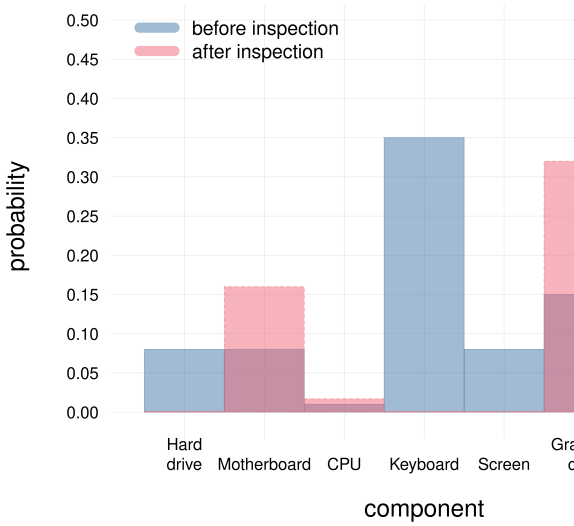
Figure 8.1: Histogram for the probability distribution over possible component failures

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.

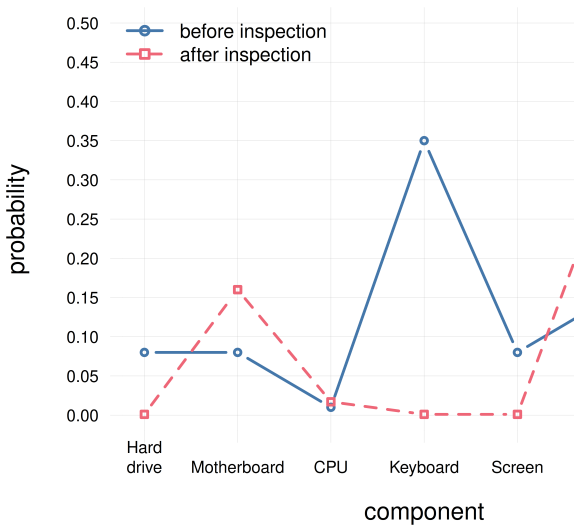
8.2.3 Line-based representations

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 distribution varies across values (for ordinal or interval quantities); or when we want to compare several probability distributions over the same values.

In these cases we can use standard line plots, or variations thereof. Compare the examples on the margin figure: the line plot displays more cleanly the differences between the "before-inspection" and "after-inspection" probability distributions.



Representation of the same pair of probability distributions with a histogram plot and a line plot



8.3 Probability distributions over infinite discrete values

@@ TODO

8.4 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 still using linguistic expressions that make at most 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 such a 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. For example,

for double-precision values stored in a computer, the width is $\epsilon \approx 2 \cdot 10^{-16}$:

```
## R code
> 1.234567890123456 == 1.234567890123455
[1] FALSE

> 1.2345678901234567 == 1.2345678901234566
[1] TRUE
```

and a value `1.3` represents a number between `1.2999999999999998223643160599749535322189331054687 5` and `1.3000000000000002664535259100375697016716003417968 75`, this range coming from the internal binary representation of `1.3`. But often the width ϵ is much larger than the computer's precision, and comes from the precision with which the value is experimentally measured.

Probabilities are therefore assigned to such small ranges, not to single values. Since these ranges are very small, they are also very numerous. The total probability assigned to all of them must still amount to 1; therefore each small range receives an extremely small amount of probability. A standard Gaussian distribution for a real quantity, for instance, assigns a probability of approximately $8 \cdot 10^{-17}$, or 0.00000000000000008, 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 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 simpler 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 $\epsilon = 2 \cdot 10^{-16}$ each 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

$$\frac{P+P}{2\,\epsilon} = \frac{1.59577 \cdot 10^{-16}}{4 \cdot 10^{-16}} = 0.398942$$

In these notes we'll denote probability densities with a *lowercase* p, with the following notation:

$$\mathrm{p}(X = x \mid I) := \frac{\mathrm{P}(\text{“}X \text{ has value between } x - \epsilon/2 \text{ and } x + \epsilon/2\text{'}” \mid I)}{\epsilon}$$

uppercase *lowercase*

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

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 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 range width, that is in principle arbitrary. This fact shows that they cannot be probabilities.
It is important not to mix up probability 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 “$\mathrm{d}x$” 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.

8.5 Representation of probability densities

8.5.1 Line-based representations

The histogram and the line representations become indistinguishable for a probability density.

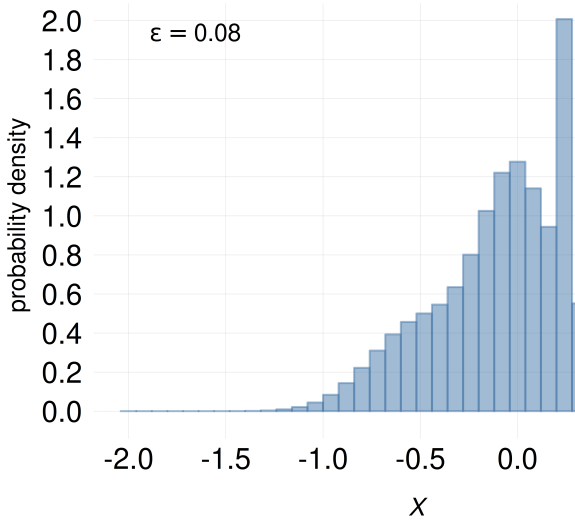
If we represent the probability P assigned to a small range of width ϵ as the area of a rectangle, and the width of the rectangle is equal to ϵ, then the height P/ϵ of the rectangle is numerically equal to the probability density. The difference from histograms for discrete quantities lies in the values 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 plots on the side.

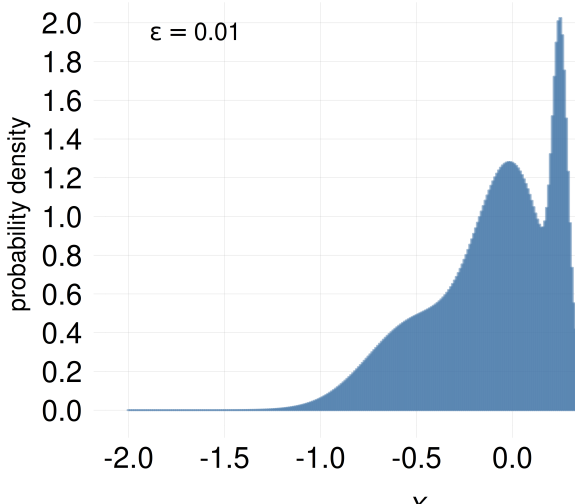
The rectangles, however, are so thin (usually thinner than a pixel on a screen) that they appear just as vertical lines, and together they look just like a curve delimiting a coloured area. If we don't colour the area underneath we just have a line-based representation of the probability density.

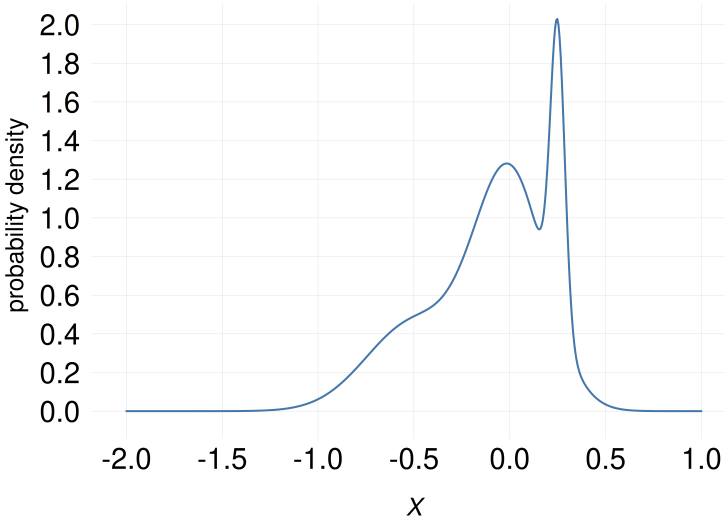
8.5.2 Scatter plots

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 with 10 lines has twice the probability of a range with 5 lines. The probability 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?

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.)

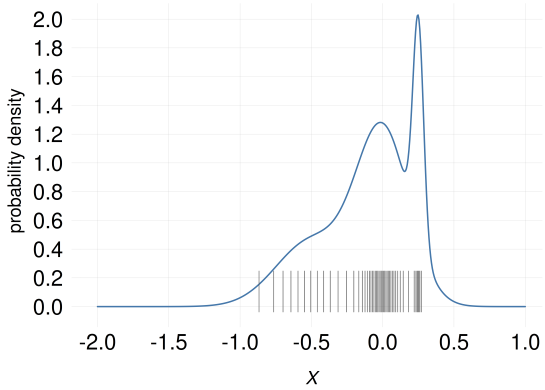


Figure 8.2: (A)

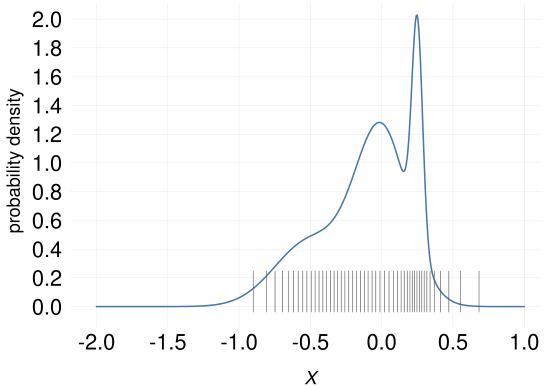
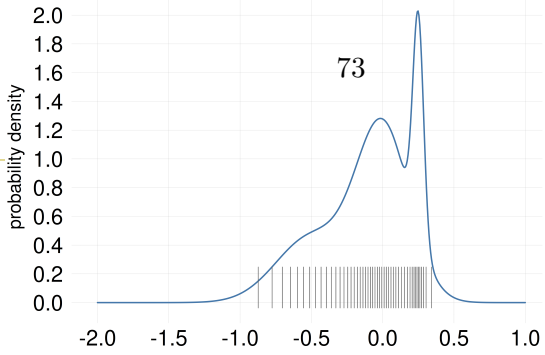
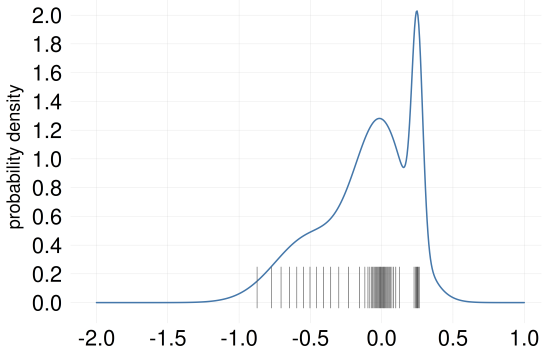


Figure 8.3: (B)





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 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.

8.6 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 call this a *combined probability*. Let's explain this procedure with an example.

Back to our initial assembly-line scenario from § 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 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’

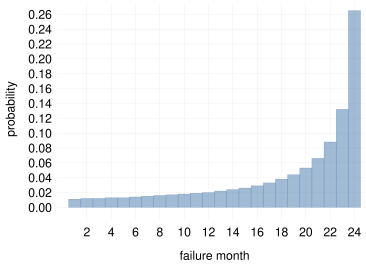
‘The component will fail during it 24th month of use or after’

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

$$\mathrm{P}(1 \mid I), \quad \mathrm{P}(2 \mid I), \quad \ldots, \quad \mathrm{P}(24 \mid I)$$



Their values are stored in this `csv` file 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’

$\equiv$ ‘The component will fail during it 1st month of use’ $\vee$

‘The component will fail during it 2nd month of use’ $\vee \cdots$

$\cdots \vee$ ‘The component will fail during it 12th month of use’

$\equiv (t=1) \vee (t=2) \vee \cdots \vee (t=12)$

The probability needed by the agent is therefore

$$\mathrm{P}(1 \vee 2 \vee \cdots \vee 12 \mid I)$$

which can be calculated using the `or`-rule, considering that the sentences involved are mutually exclusive:

$$\begin{aligned}
&\mathrm{P}(\text{'The component will fail within a year of use'} \mid I) \\
&\quad = \mathrm{P}(1 \vee 2 \vee \cdots \vee 12 \mid I) \\
&\quad = \mathrm{P}(1 \mid I) + \mathrm{P}(2 \mid I) + \cdots + \mathrm{P}(12 \mid I) \\
&\quad = \sum_{t=1}^{12} \mathrm{P}(t \mid I)
\end{aligned}$$

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 `csv` file 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 or use or after a year of use".

9 Multivariate probability distributions

So far we have considered probability distributions for quantities of a basic (nominal, ordinal, binary, 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 multivariate quantities.

(Make sure you're familiar with § 13 before you begin.)

9.1 Joint probability distributions

A multivariate quantity, as discussed in § 13.1, is just a collection or set of quantities of basic types. Saying that a multivariate 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 multivariate 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 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:

$$\mathrm{P}\big[(A = 25\,\mathrm{y}) \wedge (S = \mathrm{F}) \mid I\big]\,, \qquad \mathrm{P}\big[(A = 31\,\mathrm{y}) \wedge (S = \mathrm{M}) \mid I\big]\,, \qquad \ldots$$

where I is the agent's state of knowledge, and the probabilities sum up to one. We call this a **multivariate** or **joint probability distribution**. Usually these probabilities are written in much abbreviated form, and the $\wedge$ is represented by a comma; for instance

$$\mathrm{P}(25, \mathrm{F} \mid I) \ , \quad \mathrm{P}(31, \mathrm{M} \mid I) \ , \quad \ldots$$

9.2 Joint probability densities

@@ TODO

9.3 Representation of joint probability distributions

There is a wide variety of ways of representing multivariate 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 from the full distribution. Whenever you see a plot of a multivariate 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.

9.3.1 Tables

When a quantity is *bi*variate and its two component quantities are both discrete and finite, the joint probabilities can be reported as a table.

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` `non-urgent` care. These can be seen as two quantities A (nominal) and U (binary). When these two quantities are

taken together; their joint probability distribution is as follows, conditional on the hospital's data I_H:

Table 9.1: Joint probability distribution for transportation at arrival and urgency

$P(A, U \mid I_H)$		**arrival** A		
		ambulance	helicopter	other
urgency U	urgent	0.11	0.04	0.03
	non-urgent	0.17	0.01	0.64

We see for instance that the most probable possibility is that the next patient will arrive by transportation means other than ambulance and helicopter, and won't require urgent care.

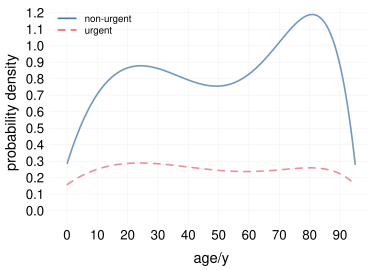
It is also possible to replace the numerical probability values with graphical representations; for example as shades of a colour, or squares with different areas.

@@ TODO ref to marginals section

9.3.2 Multi-line plots

Joint probability distributions over one discrete and one continuous quantity (or ordinal discrete with numerous values) can be represented as a collection of line plots: each line plot represent the probability density for the continuous quantity and a specific value of the discrete quantity.

Consider for instance the probability that the next patient who arrives at a particular hospital has a given age (continuous quantity) and may require urgent care or not (binary quantity). The joint probability density can be visualized as in the plot in the margin. From the plot we can see that the probability of urgent patients is generally lower than non-urgent ones. A possibly disadvantage of this kind of plots is evident: the details, such as peaks, of the density for some values of the discrete quantity may be barely visible.



9.3.3 Surface plots

9.3.4 Scatter plots

@@ TODO work also for 3D

9.4 Marginal probabilities

In some situations we have the joint probability distribution for a multivariate quantity, but we are interested in the probability for a particular value of one of the component quantities only, or in the probability distribution for that component – irrespective of what the values for the other components might be.

Consider for instance the joint probability of table 9.1. We may be interested in the probability that the next patient will need urgent care, *independently of how the patient got 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 § 7.4.

The sentence we want is

'The next patient will require urgent care'

which is equivalent to

'The next patient will require urgent care, and will arrive by ambulance, helicopter, or in by other means'

This last sentence can be written by means of `and` and `or` connectives:

'The next patient will require urgent care' ∧

('The next patient will arrive by ambulance' ∨
'The next patient will arrive by helicopter' ∨
'The next patient will arrive by other means')

which we can compactly write as

urgent ∧ (ambulance ∨ helicopter ∨ other)

= (urgent ∧ ambulance) ∨ (urgent ∧ helicopter) ∨ (urgent ∧ other)

where the second, equivalent form comes from the derived rules of Boolean algebra of § 7.7.1.

The last sentence is an `or` of mutually exclusive sentences. Its probability is therefore given by the `or` rule:

$$\begin{aligned}&\mathrm{P}\big[(\mathsf{urgent} \wedge \mathsf{ambulance}) \vee (\mathsf{urgent} \wedge \mathsf{helicopter}) \vee (\mathsf{urgent} \wedge \mathsf{other}) \mid I_{\mathrm{H}}\big] \\ &\qquad = \mathrm{P}(\mathsf{urgent} \wedge \mathsf{ambulance} \mid I_{\mathrm{H}}) + \mathrm{P}(\mathsf{urgent} \wedge \mathsf{ambulance} \mid I_{\mathrm{H}}) + \mathrm{P}(\mathsf{urgent} \wedge \mathsf{other} \mid I_{\mathrm{H}})\end{aligned}$$

The probability for a value of the urgency quantity, independently of the value of the arrival-means quantity, can be found by summing all joint probabilities with all possible arrival-means values. Using the $\sum$-notation we can write

$$\mathrm{P}(\mathsf{urgent} \mid I_{\mathrm{H}}) = \sum_{A=\mathsf{ambulance}}^{\mathsf{other}} \mathrm{P}(\mathsf{urgent}, A \mid I_{\mathrm{H}})$$

This is called a **marginal probability**.

> **Exercise**
>
> Using the values from table 9.1, calculate:
>
> - the marginal probability that the next patient will need urgent care
> - the marginal probability that the next patient will arrive by helicopter

@@ TODO point out connection with § 8.6

@@ TODO write general formula for multivariate

10 Conditional probability distributions

A joint probability distribution quantifies an agent's uncertainty about all the quantities that compose a multivariate quantity. This means that in general the agent has no sure certainty about any of them.

@@ TODO because of change in state of knowledge or hypothetical reasoning.

11 The most general inference problem

Part III

Information and data

12 Basic data types

12.1 Quantities

12.1.1 Quantities, values, domains

Most decisions and inferences in engineering and data science involve things or properties of real things. We represent them by mathematical objects – most often, collections of numbers – with particular mathematical properties and operations. The mathematical properties reflect the kind of activities that we can do with these things. For instance, colours are represented by particular tuples of numbers, and these tuples can be multiplied by some numeric weights and added, to obtain another tuple. This mathematical operation represents the fact that colours can be obtained by mixing other colours in different proportions. Physics and engineering are founded on this approach.

It's difficult to find a general term to denote any instance of such "things" and their mathematical representation, but it's convenient to find one for presenting the general theory without getting bogged down in individual cases. We'll borrow the term **quantity** from physics and engineering. A specific instance of a quantity is called its **value**. The set of possible values is called the **domain** of the quantity.

For instance:

- *quantity*: the image taken by a particular camera at a particular time, represented by a specific collection of numbers (say $128 \times 128 \times 3$ integer numbers between 0 and 255)
- *values*: one value is this:  (corresponding to a grid of $128 \times 128 \times 3$ *specific* numbers), another is this:  – and there are many other possible values

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

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

Other examples of quantities and domains:

1. The distance between two objects in the Solar System at a specific 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 some 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 $[-100\,\mathrm{N}, +100\,\mathrm{N}]$.
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 10000 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 $[0\,\mathrm{V}, 10\,\mathrm{V}]$ to $[0\,\mathrm{A}, 1\,\mathrm{A}]$.
7. A 1-minute audio track recorded by a device with a sampling frequency of 48 kHz. The domain could be all possible 2 880 000 values 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 $(-180°, +180°] \times (-90°, +90°] \times (-180°, +180°]$.

12.1.2 Notation

We shall denote quantities by italic letters, such as X for example. 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, for instance  . This kind of sentences are often abbreviated like "$X = x$".

> ❗ but keep in mind our discussion from § 5.3: we must make clear what that "=" means; it could mean "observed", "set", "reported", and so on.

12.2 Basic types of quantities

As the examples above show, quantities and data come in all sorts and with different 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 it does not make sense to say, for instance, that some value is "twice" or "1.5 times" another, or "larger" or "later" than another one. Nor does it make sense to "add" two quantities. In particular, *there is no notion of average for a nominal quantity.*

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
`Dalmatian` − `Labrador` = `Labrador` − `Dachshund`
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 it does not make sense to say that some value is "twice" or "1.5 times" another, and we cannot "add" two values. But it does make sense to say, for any two values, which one has higher rank, for example "stronger", or "later", "larger", and similar. Also in this case *there is no notion of average 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 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 does not make sense. For this reason the use of numbers can be deceptive at times. A less deceptive possibility is to represent ordered values by alphabet letters, for example.

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 we can say, at the very least, whether the interval or "distance" between a pair of values is the same, or larger, or smaller than the interval between another pair. For this reason we can also say whether a value is larger than another. We can also take weighted sums of values, called *convex combinations* (but simple *addition* of values may still be meaningless, though).

Owing to these mathematical properties, **it does make sense to speak of the average 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 one.

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 kind. But if we see values "1", "2", "3", it's unclear whether the quantity is interval, ordinal, nominal, or maybe of yet some other kind.

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; but in a specific situation we could simply label its values as `slow`, `medium`, `fast`, thus making it 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. In practice 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. In many situations the available precision is so high that we can consider the quantity as continuous for all practical purposes and use the mathematics of continuous sets – derivation, integration, and so on – to our advantage.

@@ TODO comment on repetition

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 *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 (right-)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 can a website (or the Earth?) last, and how many people can live in such an extent of time.
- What about the age of a person? What 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 – at least in theory – an infinite number of possible values (the domain of a continuous quantity always has an infinite number of values). 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, owing to the way it's measured. In this case the quantity could be considered discrete rather than continuous. Rounding can impact the way we do inferences about such a quantity.

The famous *Iris* dataset, for instance, 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 have intermediate values, of course. The age of a person is another frequent example of an in-principle continuous quantity which is rounded, say to the year or the month.

In some situations it's important to be aware of rounding, because it can lead to quantities with different unrounded values to have 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.

Iris setosa				*Iris versicolor*				
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9
5·1	3·7	1·5	0·4	6·1	2·8	4·0	1·3	5·6
4·6	3·6	1·0	0·2	6·3	2·5	4·9	1·5	7·7
5·1	3·3	1·7	0·5	6·1	2·8	4·7	1·2	6·3
4·8	3·4	1·9	0·2	6·4	2·9	4·3	1·3	6·7
5·0	3·0	1·6	0·2	6·6	3·0	4·4	1·4	7·2
5·0	3·4	1·6	0·4	6·8	2·8	4·8	1·4	6·2
5·2	3·5	1·5	0·2	6·7	3·0	5·0	1·7	6·1
5·2	3·4	1·4	0·2	6·0	2·9	4·5	1·5	6·4
4·7	3·2	1·6	0·2	5·7	2·6	3·5	1·0	7·2
4·8	3·1	1·6	0·2	5·5	2·4	3·8	1·1	7·4
5·4	3·4	1·5	0·4	5·5	2·4	3·7	1·0	7·9
5·2	4·1	1·5	0·1	5·8	2·7	3·9	1·2	6·4
5·5	4·2	1·4	0·2	6·0	2·7	5·1	1·6	6·3
4·9	3·1	1·5	0·2	5·4	3·0	4·5	1·5	6·1
5·0	3·2	1·2	0·2	6·0	3·4	4·5	1·6	7·7
5·5	3·5	1·3	0·2	6·7	3·1	4·7	1·5	6·3
4·9	3·6	1·4	0·1	6·3	2·3	4·4	1·3	6·4
4·4	3·0	1·3	0·2	5·6	3·0	4·1	1·3	6·0
5·1	3·4	1·5	0·2	5·5	2·5	4·0	1·3	6·9
5·0	3·5	1·3	0·3	5·5	2·6	4·4	1·2	6·7
4·5	2·3	1·3	0·3	6·1	3·0	4·6	1·4	6·9
4·4	3·2	1·3	0·2	5·8	2·6	4·0	1·2	5·8
5·0	3·5	1·6	0·6	5·0	2·3	3·3	1·0	6·8
5·1	3·8	1·9	0·4	5·6	2·7	4·2	1·3	6·7
4·8	3·0	1·4	0·3	5·7	3·0	4·2	1·2	6·7
5·1	3·8	1·6	0·2	5·7	2·9	4·2	1·3	6·3
4·6	3·2	1·4	0·2	6·2	2·9	4·3	1·3	6·5
5·3	3·7	1·5	0·2	5·1	2·5	3·0	1·1	6·2
5·0	3·3	1·4	0·2	5·7	2·8	4·1	1·3	5·9

The *Iris* dataset from its original paper

13 Multivariate 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. A useful, though approximate, distinction can be made on how these sets of basic quantities are made.

13.1 Multivariate quantities

Some sets of basic types are just that: sets, in the sense that they do not have new properties or allow for new kinds of operations. We shall call these **multivariate quantities** when necessary, although there's no standard terminology.

The values of a multivariate quantity are just tuples of values of its basic component quantities. Their 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 multivariate quantity "(age, sex, nationality)" which can be denoted by (A, S, N). One value of this multivariate quantity is, for example, $(25\,\mathrm{y}, \mathrm{F}, \text{Norwegian})$. The domain could be

$$[0, +\infty) \times \{\mathrm{F}, \mathrm{M}\} \times \{\text{Afghan}, \text{Albanian}, \ldots, \text{Zimbabwean}\}$$

[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 multivariate 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 multivariate quantity is called, for instance, "continuous" if all its basic components are continuous; but other conventions are also used.

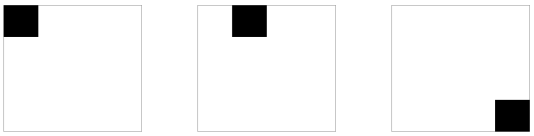
> Exercises
>
> Consider again the examples of § 12.1.1. Do you find any examples of multivariate quantities?

13.2 Complex quantities

Some complex quantities can be represented as sets of quantities of basic types. These sets, however, are "more than the sum of their parts": they possess new physical and mathematical properties and operations that do not apply or do not make sense for the single components.

Familiar examples are vectorial quantities from physics and engineering, such as location, velocity, force, torque. Another example are images, when represented as grids of basic quantities.

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:



represented by the numeric matrices $\begin{pmatrix} 1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0 \end{pmatrix}$, $\begin{pmatrix} 0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0 \end{pmatrix}$, $\begin{pmatrix} 0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1 \end{pmatrix}$.
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 not a property of each of them individually.

More generally, complex quantities have additional, peculiar properties, represented by mathematical structures, which distinguish them from simpler multivariate 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. The importance of machine-learning methods lies to a great extent in the fact that 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 correctly but for the wrong reasons.

14 Statistics

14.1 The difference between Statistics and Probability Theory

Statistics is the study of collective properties of collections of data. It does not imply that there is any uncertainty.

Probability theory is the quantification and propagation of uncertainty. It does not imply that we have collections of data.

Part IV

Decision theory

15 Making decisions

15.1 Decisions, possible situations, and consequences

15.2 Gains and losses: utilities

15.2.1 Factors that enter utility quantification

Utilities can rarely be assigned a priori.

15.3 Making decisions under uncertainty: maximization of expected utility