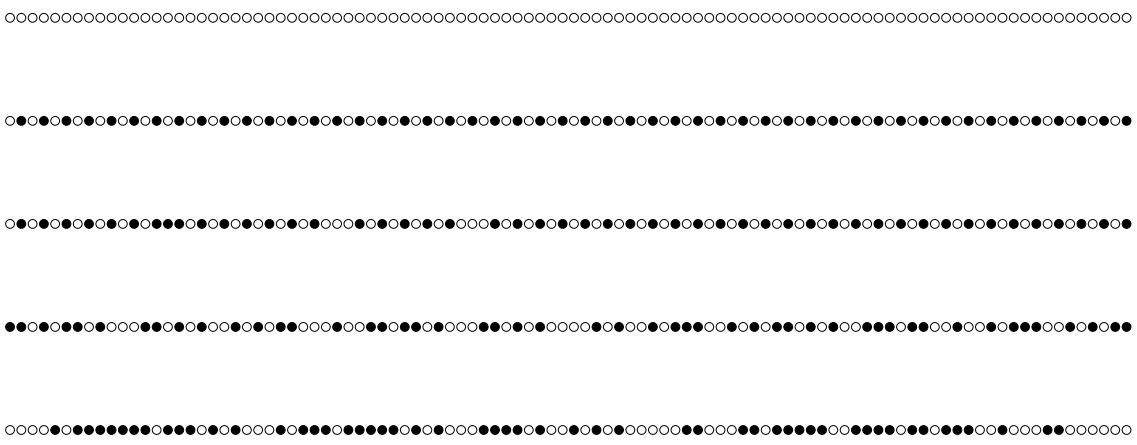
Lecture Notes: Probability A (ST111)

Warwick University, 2012

Julia Brettschneider

Version β.3 (6.3.2012)



What others have said about probability

Many returning Warwick mathematics graduates:
Why didn't someone tell us that probability and statistics are *the key mathematical subjects* in applied quantitative work?

A randomly selected first year student:
Given I passed the exam, what is the probability I studied for it?

Another first year student:
Is there any way I can buy that hat off you?

Another student:
You keep telling me to read books.

A rumour:
The exam will only be taken by a sample of 40 students selected at random from this class.

A coin:
This module tossed me around in my sleep!

Pierre Simon Laplace:
It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

Niels Bohr:
Einstein, stop telling God what to do with his dice.

Stephen Jay Gould:
Misunderstanding of probability may be the greatest of all impediments to scientific literacy.

ebay:

Shadowfist CCG Probability Manipulator SE Rare
0 Bids £0.99 + Postage £0.75 12h 21m

Oscar Wilde:
Always be a little improbable.

1 Preface

Lecture notes

It is very difficult to simultaneously listen to a lecturer, read what is being written on the blackboard and copy the essential of both sources. First year modules are a good opportunity to experiment with note taking strategies, because you can get hold of most of the material (and more) from textbooks and/or lecture notes. In later year module this may not be the case. In research seminars or meetings (in or outside academia) there will be no notes and note taking skills will come in handy.

My recommendation for taking notes is to scribble down as much as you can of both the material on the blackboard and of what is said without worrying about whether it looks neat and whether you understand it in real-time. Later, you rewrite your notes as a way of deepening your understanding of the material.

These typed lectures notes are an example of what this might look like. They are very close to what happened in the actual lectures, but they are not exactly the same. In particularly, they contain additional examples, details or some proofs that were only sketched in the lecture and additional background on applications and foundations.

The lectures and the notes are based on a number of other sources. First, there are previous versions of this module taught by my colleagues. I thank Saul Jacka for sharing his lecture notes with me – some parts are taken straight from his notes – and I thank Roger Tribe for providing a sketch about content and motivation for his lectures. Second, some of the fascinating probability textbooks; as listed below. Third, my notes from teaching at University of California at Berkeley, Technical University Berlin and Humboldt University Berlin. Finally, I have been inspired by Hans Föllmer's lectures who taught me probability in the first place. Special thanks goes to Charlie and Lola[2].

ω's nickname omi is chosen in memory of my grandmother who passed away during the first time I taught this module. She introduced me to dice and cards games. She gave me a set theory kit that had made its way into toy stores in the wake of the 1960s *New Math*[3] movement introducing $\cap$ and $\cup$ between sets of colourful triangles, squares and circles to help 6 year olds doing their math homework.

Websites

There are different websites for this module:
`http://www2.warwick.ac.uk/fac/sci/statistics/modules/st1/st111`
`http://www2.warwick.ac.uk/fac/sci/statistics/modules/st1/st111/resources`
The first one is very general and is intended to help students with their module choice or with anticipating the role of this module within the course program. The second website, the *resources website,* is the one that is more relevant for you now. There, I will post lecture notes, exercise sheets, information about the module organisation, old exams, computer code and interesting links about probability. In previous years, some of the lecturers have used the *MathStuff* website by the Warwick Mathematics Department to post exercises,

[2] `http://www.bbc.co.uk/cbeebies/charlieandlola/stories/`
[3] to get a taste of what this is see `www.bbc.co.uk/cbeebies/charlie-and-lola/`

lecture notes, old exams and other information. This could still be useful to you; just go to ST111 > Module Pages > Archived Material; or at this address:
`http://mathstuff.warwick.ac.uk/ST111/archive`
Finally, there is my website:
`http://www2.warwick.ac.uk/fac/sci/statistics/staff/academic/brettschneider/`

Textbooks

Please check out the list on the resources website. Which one is best? Some people say the one by Ross tends to more or less suit the majority of students. It certainly has a lot of examples and it covers essentially all the material from this module, as well as a lot more. The textbook by Pitman teaches you to love probability and to do so with the minimum of notation and formality. Examples are chose with care and discussed in a way that connects intuition with mathematically formal language.

By the way, have you experienced the *law of the second source*? The second source where you read about a piece of mathematics new to you tends to be the one that explains it better – regardless of the order of the sources. Interestingly, there is no such law for the third source. In fact, there seems to be a $k \geq 3$ such that while reading the kth source you end up being more confused than while reading the first one, again regardless of the order of the sources. Without getting to that point you will never know how well you had already understood the material before getting hold of source k. For this and many other reasons,

$$\sum_{i=1}^{\infty} \text{“read another book”}$$

Cover pictures sources

The cover picture is made using the statistical programming language R which is publicly available, in source code form, at `www.r-project.org`
In other words, you can use it, too. A real strength of R is the huge amount of packages contributed by the community of R Users from all over the world. In 2011, the R community had their annual conference at Warwick `www.warwick.ac.uk/statsdept/user-2011/`

Assessment and exams

The one most frequently asked question is: How do I study for the exams?
The answer is, in a nutshell,

$$\sum_{i=1}^{\infty} \text{“do another exercise”}$$

Apologies about ∞, this is an idealised model. Reality is more complex. You also need to work through the material until you have embraced all key concepts (i.e., understand definitions and techniques) and you need to get enough sleep before the day of the exam.

Synopsis of the lecture

This is a rough plan which may be modified as we move through the term. General remarks about the place of probability theory within mathematics and about its connections with the so-called "real world" tend to pop up in the lecture when they to fit in; in the notes they are summarised in Sections 2.2 to 2.6. The schedule for Part A is detailed below; for a schedule of Part B see below the notes to Part A.

Week I

1) Probability as a science of patterns, random sequences, birthday problem
2) Three door puzzle, fallacies in dealing with probabilities, two-headed coin example, connection between probability theory and statistics

Week II

3) Terminology for random events, some classical random experiments
4) Classical probability for equally likely events, summary set theory and event language
5) Combinatorics: counting and sampling

Week III

6) Combinatorics
7) Axioms of probability
8) Deductions from the axioms, examples for probability spaces, cumulative distribution function

Week IV

9) Conditional probability
10) Total probability theorem, Bayes theorem
11) General multiplication rule, independence

Week V

12) Sequences of independent events: binomial, geometric, negative binomial
13) Poisson approximation to the binomial
14) Poisson distribution, Law of large numbers

Have a great journey into probability spaces!

2 Motivation

2.1 Observing random phenomena

We start by looking at a few pictures from the "real world" that exhibit random phenomena. The slide show used during this lecture can be downloaded from the ST111 resource website. It includes ice crystals, animal patterns, genetics, traffic, networks, statistical mechanics, stock markets, gambling, knitting and more.

The common definition of mathematics as the *science of patterns* goes back to G.F. Hardy, if not longer. Number theorist look for patterns in the integers, topologists in shapes, analysts in motion and change and so on.

Look at the bottom picture on the cover sheet of these lecture notes. Are the black and white dot sequences regular or not? In which sense (not)?

- Row 1: Yes, it is constant.
- Row 2: Yes, it is periodic.
- Row 3: No, but almost. It is the sequence from row (2) with a few errors.
- Row 4: No, there seems to be no rule that generates this sequence.
- Row 5: No, there seems to be no rule that generates this sequence.

Question 2.1. *Statistical regularity in dot sequences. The sequences in Row 4 and Row 5 clearly are not regular in the classical sense: There is no deterministic rule that generates them. But are they statistically regular in some sense? And if so, do they have the same kind of statistical regularity in the sense that they show similar statistical characteristics? One of them has actually been generated by flipping a fair coin 100 times and printing a white dot for head and a black dot for tail. Can you say which one? And why do you think so?*

ω**:** But excuse me, I do not like black and white. My very most favourite colour is totally tomato red and my other favourite colour is bitter chocolate brown with lime green sparkles in it and YOU should really try this link: `http://biscuitsandjam.com/stripe_maker.php`

Ω**:** Speaking of black and white, have you heard that half of the students taking this module loves us and the other half hates us? I have decided that the most sensible step for us is to turn grey and to shrink bit, so so we do not overstay our welcome.

About Question 2.1: Most likely, Row 5 is the coin toss. Does the first one look more random to you? Well, it does switch more often between black and white, and we seem to take that as a sign for proper randomness. But could this be too much of a good thing? Consider that staying with the same colour is as likely as switching to another colour we would expect that a sequence of 100 coin tosses has about 50 colour changes. However, there are 66 colour changes in Row 4 and 46 Row 5. While we can not say with certainty, we have argued that the difference in the number of observed colour changes makes Row 5 much more likely to be generated by the coin tossing by than Row 4. This is an example for statistical thinking (see also Section 2.5).

2.2 Studying random phenomena

It seems desirable to get more insight into the phenomenon of randomness, but what kind of discoveries can we expect from studying something as inconceivable as randomness?

> The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.
>
> — James Clerk Maxwell

> How dare we speak of the laws of chance? Is not chance the antithesis of all law?
>
> — Joseph Bertrand, *Calcul des probabilités*

Why would anyone want to study randomness? There are a number of reasons including

- describing and understanding patterns in random phenomena,
- unrevealing scientific principles,
- predicting and forecasting events,
- learning and developing degrees of belief about events,
- quantifying and comparing risks,
- taking decisions under uncertain circumstances.

Probability theory provides a formal theory to describe and analyse random phenomena.

2.3 Probability is pure mathematics (with some probability $p > 0$)

Probability can be practiced as pure math. The birth of of probability theory as a mathematical discipline is often associated with the axiomatisation the field underwent during the twenties and early thirties of the 20th century.

It borders other fields in mathematics. Most obvious are the overlaps with measure theory, analysis, functional analysis, PDEs, geometry, combinatorics, mathematical statistics, computing, ergodic theory, number theory and mathematical physics. The probability group at Warwick is among the leading ones in the UK. Their major home is P@W; see `http://www2.warwick.ac.uk/fac/sci/statistics/paw/`

2.4 Probability is applied mathematics, with some probability $q > 0$

> The instrument that mediates between theory and practice, between thought and observation, is mathematics; it builds the bridge and makes it stronger and stronger. Thus it happens that our entire present day culture, to the degree that it reflects intellectual achievement and the harnessing of nature, is founded on mathematics.
>
> — David Hilbert, radio speech (1930)

Probability provides a language and a methodology for the human mind's inquiries into random phenomena in a variety of fields.

Physical world: e.g. astronomy (in particular, theories about measurement error), mechanics, statistical physics, quantum mechanics.

Living world: e.g. integrative biology, evolution, genetics, genomics, medicine, neuroscience.

Social world: e.g. economics (in particular, financial markets), sociology, demography.

Engineering: e.g. computer science (in particular, machine learning), electrical engineering, risk assessment, reliability, operations research, information theory, communication theory, control theory, traffic engineering.

Humanities: e.g. epistemology, determinism, philosophy of language, philosophy of religion, philosophy of logic, political philosophy, belief systems.

Games of chance: e.g. roulette, dice, card games.

Arts: e.g. stochastic music, visual art (inspired by chaos theory and fractals, patterns simulated by stochastic algorithms).

Probability theory is based on a set of axioms that can be summarised in just a few lines. What makes it applicable to so many different kinds of applications is a huge repertoire of probabilistic models. For example: Classical discrete probability models in games of chance, genetic inheritance, quality control, measurement error, electrical circuits and lattice modes for particle interaction. Discrete time stochastic processes (often involving a certain depth of dependency on the past) in queuing, coding and stochastic composition algorithm. Continuous stochastic processes for particle motion and stock prices. Probabilistic networks in sociology, mobile phone and genomics.

2.5 Probability as the mathematical foundation of statistics.

"Probability and statistics used to be married, then they separated, then they got divorced and now they hardly see each other," says the British (actually, Welsh) probabilist David Williams in his textbook about probability and statistics[1]. We add that at Warwick they have moved on to be good friends offering a wide range of research activities. They cover any possible blend of the two disciplines from pure probability to applied statistics and extending to interdisciplinary collaborations with other Warwick research groups such at Complexity, Systems Biology, Medical School, Analytical Sciences, Economics and Finance.

Probability and statistics are connected by studying similar objects from different perspectives:

Probability starts out with a model for a random experiment that describes potential outcomes and assumes certain basic parameters. Probabilities for events of interest are *deduced* from these model assumptions.

For **statistics,** the starting point are the *observed outcomes* (aka *oberservations* or *data*) of an experiment that has already been performed. The task is to *infer* characteristics of an optimal model for the unknown (or partially known) mechanism of the experiment. The goal is to find the model that is, among a certain class of models, most likely to have

[1] David Williams (2001), Weighing the odds, Cambridge University Press.

created the observations. A statistical judgement can not be made with certainty. Instead, usually come with

We will proceed with a toy example invoking a simple probabilistic model and a first glance at statistical thinking.

2.5.1 Two-headed coin

Question 2.2. *Suppose you have a bag with one two-headed coin (i.e., a coin with heads on both sides rather than a head on one side and a tail on the other side) and $n-1$ normal coins (in particularly, we assume they are fair). You pick a coin at random from this bag. Without inspecting the coin, you throw it three times. You observe three heads. Do you think it is the two-headed coin? How sure are you about that?*

Answer to Question 2.2: We can not be sure. It seems more likely, though, that it was the two-headed coin. Unless n was large, as this would make it very unlikely to pick the two-headed coin. Let us do some explicit computations.

We will compute the probabilities for the two possible ways to obtain the outcome three heads. Then we will compare them by looking at their ratio R.

If the two-headed coin was picked we certainly get three heads. So the probability for this possibility is $1/n$. If the normal coin was picked, the probability for obtaining three heads is $1/8$, because it is one out of a total of eight possible outcomes of tossing a coin three times (you can check by listing all of them explicitly). So the probability for this possibility is $(n-1)/n \cdot 1/8$. So we get

$$R_n = \frac{1/n}{(n-1)/n \cdot 1/8} = 8/(n-1) \tag{1}$$

So we obtain $R_2 = 8, R_3 = 4, R_4 = 2\frac{2}{3}, R_3 = 2$ and the series keeps decreasing until it hits $R_9 = 1$, which means that both possible ways for obtaining three heads are equally like. For larger n, R_n is even smaller than 1, which means that explaining the outcome by the two-headed coin is less likely to be true than explaining it by a normal coin.

Say $n = 3$. Would you be ready to bet £100 that the three heads come from using the two-headed coin? It's a bet in your favour, but it's still somewhat risky. Uncertainty could be reduced by throwing the coin more often. In the language of statistics, you are increasing the size of your sample.

Now let us do the math for the general case of throwing the coin k times. The bit that changes is the likelihood of getting an outcome with all heads when you use a normal coin. You may know from school that this equals $1/2^k$. (Otherwise, wait just until next week's lectures to see how these things are computed.) This leads to

$$R_n = \frac{1/n}{(n-1)/n \cdot 1/2^k} = 2^k/(n-1) \tag{2}$$

We can fix n and study the ratio R_n as a function of k. For n of the form $n = 2^j + 1$ for some integer j the expressions simplify to a convenient expression: $R_n(2^j+1) = 2k - j$, for example, $R_2(k) = 2^k, R_3(k) = 2^{k-1}, R_5(k) = 2^{k-2}$. We can see that for larger n, we need a

higher number of k to obtain a higher probability for the explanation that the two-headed coin was used than that the normal coin was used. Roughly speaking, we have to offset the low probability that the two-headed coin was selected in the first place.

This question is an example for statistical thinking: given the outcome, what is the most likely explanation for it. Probability theory was used to compute the different probabilities involved.

Ω: It is a bit boring with the two-headed coin. I'd have just inspected the coin properly and hence avoided all this uncertainty!

ω: But sometimes you can not find out everything by just looking at it. I saw someone making loaded dice on youtube!

Ω: I guess the question above was just a toy example to focus on the most essential. A more subtle version would involve a biased coins rather than a two-headed coin. The comparison of likelihoods given different explanation is an example for the kinds of approach used in the paradigm of hypothesis testing.

ω: You know, Omega, the lecturer told me that they had taught hypothesis testing in her school and, as a result, never looked at statistics again until 15 years later 10,000 miles away from home.

Ω: Maybe, if someone had told her earlier that there is nothing logical about these mysterious 95%, but it's just some sort of standard that leading figures make up based on their priorities, like policies regulating university tuition fees.

2.6 How are we going to study probability in this module?

In his classical book *An Introduction to Probability Theory and its Applications,* William Feller stresses that in probability, as in other mathematical disciplines, we must distinguish three aspects of the theory:

(a) the formal logical content,

(b) the intuitive background,

(c) the applications.

He continues: "The character, and the charm, of the whole structure cannot be appreciated without considering all three aspects in their proper relation."

Considering both the history of probability and the fact that this is a first year module we start with models for equally like probabilities. This approach is minimalistic in the formal sense but already provides enough mathematical foundation to properly dive into some interesting examples typical for probability as a field.

While some students are extremely good at solving probability problems intuitively, others would prefer to see a more formal approach first, which is why we will then introduce the modern axiomatic approach to probability including the notions of σ-algebras, probability measures and random variables.

Besides these objects, the mathematical theory of probability includes crucial concepts such as conditioning and independence, expectations and variances. In the easier examples, the introduction of such formalism may feel like breaking a butterfly on a wheel. It serves the purpose to get used to the formalism, which becomes very useful for more complex questions some of which even have counterintuitive answers.

Along with the theory we will get to know classical probability distributions such as

Bernoulli, uniform, binomial and geometric. All of this will be motivated by "real world" questions and illustrated by further applications and by examples constructed from probabilistic experiments (such as coin tossing or gambling examples).

2.7 Some examples to surprise and enthuse you

ω: Some people use enthuse you as euphemism for *confuse* you. But it is your birthday, so that's why I will stick around.

2.7.1 Birthday problem

Question 2.3. *Suppose there are r students in this class. What is the probability that at least two students in the class have the same birthday?*

ω: But I need to *know* everybody's birthday first to find out whether any two are on the same day!

Ω: Don't be silly. In the question they don't care about exactly *which* students these are. They just want to know *what are the chances* of two identical birthdays in a *random* class of r students.

ω: You are the one being silly. A *random* class is how our school teacher calls us for being naughty.

Ω: Not that kind of random. I mean a class *picked at random* from *all the classes* of r students in the universe.

ω: *Universe, probability space, outer space* – you are just making up funny words. How is the universe going to help you out? You do not even know who is out there. Maybe on Mars everybody is born in May, and on Saturn everybody is born on a Saturday, and in the Milky Way everybody is born on a Lucky Day!

Ω: OK. That is a very good point, we do need to be clear on what kind of space we are solving that problem. I am inclined to assume that every student is equally like to be born an any of the days of the year, and that it has nothing to do with when others are born.

ω: But excuse me, there are leap years, there are more birthdays in November and twins have the same birthday!

Ω: You're absolutely right, but I just want to start somewhere while insuring that my assumptions still capture the essence of the problem. Once we have solved it this way we can see how important the assumptions are in the calculation and then tackle it with a different set of assumptions.

ω: OK, so let's go for it. There are 23 kids in my class. They are all very special, hopefully there no high risk that any two of us have the same birthday. The probability that Alphie is born on my birthday is only 1 in 365, for Betty it is also 1 in 365 and so on, which makes 24/365 all together. Now, Alphie makes the same calculation for her, and so does Betty and and everybody else. Putting it all together makes 25 times 24 divided by 365.

Ω: Which is about 1.64. How can a probability be larger than 100%?

ω: OK, I see, I counted some things twice, like Alphie on the same day as me is the same as me on the same day as Alphie. Let me try again....

Ω: I know you like questions. But this is actually *my* question and I will approach it by looking at a simpler version it. Say there are only three students in a class and we are looking at any two of them having their birthday in the same month denoted by $a, b, c, \ldots, l$. Now I will figure out all possibilities for the birthday months of the three students and count how many of them have at least two identical months in it. I am using two PROBLEM SOLVING TECHNIQUES: *Look at a toy problem, consider all options.* Now, $aaa, aab, aac, \ldots, aal$ obviously all belong in this group, makes 12. Now, $aba, abb, abc, \ldots, abl$. Of those, only two have two identical months. Same for all the once starting with ac or ad and so on up to al. So far, we have 12 plus 2 times 11. Now, there are $baa, bab, \ldots, bal$ and $bba, bbb, \ldots, bbl$...

ω: And bla, bla, bla and... Omega, I am hungry, and there's got to be a better way for this!

Answer to Question 2.3: We make three assumptions.

(i) *The year has $n = 365$ days.*

(ii) *Every student in this class is equally likely to be born on any day of the year.*

(iii) *The birthdays are independent of each other.*

Are these assumptions are they realistic enough? (i) is correct for 3 in 4 years, otherwise it is almost correct. (ii) is a simplification. There is a surge of birthdays in autumn in the UK, but it's not that big and we will neglect it for now. To be sure the models suits, we would also need to check out the data from overseas to account for the students not born here. Finally, we need to think about how the class was selected from the general population and whether this might in any way be affected by birthdays. For example, if the students in a year was split into classes directly or indirectly defined by birthdays. (iii) is usually realistic, but there may be exceptions such as a meeting of people born as twins.

To answer the question we first order the students in some way. With the above assumptions our model is that the birthdays are *equally likely* to form any ordered r-tuple with numbers chosen from $\{1, 2, \ldots, n\}$.

We want to compute the probability $p_{n,r}$ that there are at least two students with the same birthday among the r students in the class. Obviously, if $r > n$ this has to happen, so $p_{n,r} = 1$. Otherwise, we can compute it by dividing the number of r-tuples containing repeats by the total number n^r of r-tuples from a set of n elements.

Omega attempted to list all possible r-tuples, identify which of them have repeats in them and count those. Whereas this is a correct way of solving the problem, it is more efficient to just list and count those r-tuples that do *not* have any repeats in them. This will allow us to compute the probability $q_{n,r}$ of the *opposite*, that is, not any two students have the same birthday?
Problem solving technique: *Check if the opposite event is easier to handle.*

In how many ways can birthdays be assigned without ever repeating one? Let the first student have whatever birthday. Avoiding that day for the second student leaves $n - 1$ options. The third student has to avoid both previously taken birthdays which leaves $n-2$ choices. The forth has $n - 3$ choices and so on up to the $n - r + 1$ choices for the rth student. So we have

$$q_{n,r} = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1)/n^r,$$

which yields

$$p_{n,r} = 1 - n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1)/n^r.$$

Let us look at some numerical results:

$$\begin{array}{ll} p_{365,10} \approx 11.7\% & p_{365,40} \approx 89.1\% \\ p_{365,20} \approx 41.1\% & p_{365,50} \approx 97.1\% \\ p_{365,30} \approx 70.6\% & p_{365,60} \approx 99.4\% \end{array}$$

In particular, we can see that the smallest number of students r such that the probability of having at least two of them with the same birthday is 23:

$$p_{365,22} \approx 47.6\% \qquad p_{365,23} \approx 50.7\%$$

2.7.2 Door problem

An old kind of problem that became famous worldwide via the Monty Hall game show in 1990. You are being shown three closed doors. Behind one of the doors is a car; behind each of the other doors is a goat. You choose one of the doors. The show's host, who knows which door conceals the car, opens one of the two remaining doors which he knows will definitively reveal a goat. Then he asks you whether or not you want to switch your choice to the remaining closed door. What are you going to to? Here are some attempts to answer this question.

(i) Either the prize is behind the door you bet on originally or behind the other still closed door makes it a fifty-fifty chance for each, so it doesn't matter whether you change or not.

(ii) All doors were equally like originally. That is not going to change because of whatever that show's host did. So it doesn't matter whether you change or not.

(iii) Imagine the same problem but with 100 doors instead of 3. Behind one of the doors is a car, behind all the other 99 doors is a goat. The show's host knows where the prize is and opens all doors but the one you picked and one other one. Now it seems intuitive, that you would want to switch.
PROBLEM SOLVING TECHNIQUE: *Exaggerating the original question helps to reveal the essential characteristics of a problem. Often, the qualitative aspects of the answer to the exaggerated question can be carried over to answer the original question.*

(iv) We compare the two different strategies *stick* (with your original choice) and *switch*. There are two possibilities:
Case 1: Original choice is door with car. Now you will get a goat.
Case 2: Original choice is a doors with a goat. Now you will get a car.
The probability for Case 1 is 1/3, the probability for Case 2 is 2/3. If your strategy is *stick* then your chance of winning the car are 1/3. If your strategy is *switch* then your chance of winning the car are 2/3.
PROBLEM SOLVING TECHNIQUE: *Distinguishing cases allows to find solutions of the problem under more constraint conditions. In probability, it is usually done to get rid of (some of) the randomness.*

(v) A qualitative approach. The shop's host knows where the car is and opens, on purpose, a door that reveals a goat. If you make use of that information it should increase your chances of getting the car. At least it should not decrease it.

(vi) A variation of the problem. The show's host does not know where the car is, he just *happens* to reveal a goat when he opens the door. Does this change your answer to the problem?

2.7.3 Two short questions

Question 2.4. *Which one of the following birth orders in a family with six children is more likely, or are they the same? (G means girls, B means boy.)*

GBGBBG *BGBBBB*

Question 2.5. *A hexagonal die with 2 red and 4 green faces is thrown 20 times. You have to bet on the occurrence of one of the following patterns. Which one would you choose?*

RGRRR *GRGRRR* *GRRRRR*

Answers to these two questions are given below. Before you read them, try to come up with your own answers.

About Question 2.4: Both are equally likely. Experiments with people who have no training in probability have shown that they tend to believe that the first order is more likely than the second order. The second sequence is perceived to be too regular. The answers suggest that the respondents compare the frequency of families with 3 girls and 3 girls with the frequency of families with 5 boys and 1 girl. Yet both *exact orders* of birth, $GBGBBG$ and $BGBBBB$, are equally like, because they both represent one of 64 equally like possibilities. (See Kahneman and Tversky, 1972b, p.432 see Chapter 5, Neglecting exact birth order.)

About Question 2.5: Your best bet is (i). Most people bet on (ii), because G is more likely than R and (ii) has two Gs in it. Yet (i) is more likely, simply because $RGRRR$ is nested in $GRGRRR$. This is an example of committing the *conjunction fallacy* without realising that there is a conjunction. (See Kahneman and Tversky, 1983, p.303, Failing to detect a hidden binary sequence.)
The conjunction fallacy has been studied by many researchers using the example about the (hypothetical) person *Linda*. The description reads:

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*

Then people are asked to rank statements about Linda by their probability, in particularly the following two:

> *Linda is a bank teller.*
> *Linda is a bank teller and is active in the feminist movement.*

People assigned higher probabilities to the *second* statement. This is *not* in tune with the mathematical rules about probability and has stimulated a discussion about how real people deal with probabilities. (See Kahneman and Tversky, 1982b, p.92.)

ω: People's minds don't follow the usual probability rules, incy wincy tiny little particles do not follow them either; are they of any use?

Ω: Of course they do work a lot of the times, for particles and even for human. It's just that we enjoy these curious examples. By the way, for those fractions of elementary particles, well, that's why *quantum probability* was invented.

3 Models for equally probable events

> "On Tuesday 9th October 1972, in the first serious lecture given to a group of 45 second-year mathematicians, entitled Possibilities & Probabilities, the founding professor tossed a 2p coin high in the air. The coin descended to the vinyl floor of lecture theatre L5, spun as a perfect sphere, and, in full view, slowly came to rest on its edge! Stunned silence turned into massive applause. No further publicity was necessary truly the Statistics Department had arrived in style!"

> — Jeff Harrison, A Brief History of the Early Years of the Statistics Department [at Warwick].[2]

> "All models are wrong, but some are useful"
> — George Box

In this chapter we develop a classical mathematical framework to model *random experiments,* i.e. actions with unknowns results. More generally, it serves to model systems that incorporate uncertainty. The basic ingredients are:

- An *outcome* is a result of the experiment. It is not known until the experiment has been performed.
- The set of all possible outcomes is called *outcome space* or *sample space.* This set is known beforehand.
- An *event* is a collection of outcomes. In other words, a subset of the outcome space. Often, we are interested in obtaining certain kinds of outcomes (e.g. an odd number) rather than a specific one.

Typically, we use Ω for the outcome space, ω for its elements and $A, B, C, \ldots$ for events.

3.1 Classical random experiments

The probabilists' favourite example is the coin flip. One coin flip in the easiest non-trivial example for a probabilistic model. And it can be made more interesting by flipping two coins. Or many coins, making this a discrete stochastic process.

Example 3.1. Coins. Flip a coin and observe what it is facing. Then $\Omega = \{h, t\}$ with h for the coin facing heads and t for the coin facing tails. The standard model for flipping two coins is $\Omega = \{hh, ht, th, tt\}$. (See homework sheet 1 for a discussion of alternative models.)

Ω: Look omega, `www.btwaters.com/probab/flip/coinmainD.html` flips coins for you.

ω: Look! Here you can really see how all of these coins land `www.random.org/coins/`.

Ω: Thank you, omega. I'll start with flipping a 1 Pound coin and then go on to a *Constatius II - Silver Siliqua.*

ω: What is an East German mark?

Ω: I don't know, but it says obsolete.

ω: And what about these European coins? Mum said they don't work anymore either.

Ω: They still do have the 1 Euro coins here... Actually, I heard they were going to replace them by dice reflecting uncertainty of the value.

Example 3.2. Dice. The most common ones are made from *Platonic Solids,* i.e. regular convex polyhedrons: the tetrahedron (4 faces), the cube or regular hexahedron (6 faces), the octahedron (8 faces), the dodecahedron (12 faces) and the icosahedron (20 faces). The outcome space for rolling an n-faced die and observing the number it faces is $\Omega = \{1, \ldots, n\}$.

[2] `www.maths.warwick.ac.uk/general/institute/histories-small.pdf`

For a cubic die,$\Omega = \{1, 2, 3, 4, 5, 6\}$. Here are some events:

$$\begin{aligned} A &= \text{“number is even”} = \{1, 3, 5\} \\ B &= \text{“number is smaller than 5”} = \{1, 2, 3, 4\} \\ C &= \text{“number is divisible by 3”} = \{3, 6\} \end{aligned}$$

They can be combined using set operations, e.g.

$$\begin{aligned} A \cap B &= \text{“number is even AND smaller than 6”} = \{2, 4\} \\ A \cup B &= \text{“number is even OR smaller than 6”} = \{1, 2, 3, 4, 6\} \\ A \cap B \cap C &= \text{“number is even AND smaller than 6 AND divisible by 3”} = \emptyset \\ A \cup B \cup C &= \text{“number is even OR smaller than 6 OR divisible by 3”} = \Omega \end{aligned}$$

What all these models have in common is that they have finitely many outcomes. Here is a model that covers all such situations.

Example 3.3. Finite number of tickets. A box contains a finite number n of tickets enumerated $1, 2, \ldots, n$. Draw a ticket at random. The standard choice for an outcome space is $\Omega = \{1, 2, \ldots, n\}$. Events can be subsets of Ω such as:

$$\begin{aligned} \text{“ticket shows 3”} &= \{3\} \\ \text{“ticket shows a number between 5 and 10”} &= \{5, 6, 7, 8, 9, 10\} \\ \text{“ticket shows an even number”} &= \{2k \,|\, k \in \mathbb{N}, k \leq n/2\} \end{aligned}$$

In Examples 3.1 to 3.3 we can define probabilities corresponding to *equally like outcomes* as follows. Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ and set

$$P(\omega_1) = 1/n,\ P(\omega_2) = 1/n, \ldots\ ,\ P(\omega_n) = 1/n. \tag{3}$$

3.2 Classical probability

Classical probability models are based on symmetric characteristics of the random mechanism generating the outcomes. Often, these are physical characteristics such as a die with faces of the same shape or a box with balls of the same size. Calculations can usually be performed directly or indirectly using the equiprobable model (3).

The axioms for classical probability we will define below are only a slight extension of that model with the main difference being the mathematical framework. Rather than defining probabilities for outcomes, we use events. In other words, we introduce probability as a function on subsets of the outcome space. First we put up a structure that allows us to combine events without falling out of that structure.

Definition 3.4. Algebra of sets.
Let Ω be a set of points. A system $\mathcal{A}$ of subsets of Ω is called algebra if

(A1) $\Omega \in \mathcal{A}$

(A2) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$

(A3) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$

To get an idea what such an algebra is, think of it as a system of sets which, whenever it includes a number of sets, it also includes all combinations of these sets that are obtained by a finite number of the admissible operations (union and complement). In particularly, this applies that the algebra is also closed under intersections, because it can be represented as $A \cap B = (A^c \cup B^c)^c$.

That tells us what *other* sets are in the algebra, but what sets do we actually start with? One very common way to create an algebra of sets is to start out with a partition of Ω.

Definition 3.5. Partition.
A system $B_1, \ldots, B_n$ of subsets of Ω is called partition *if it has the following two properties:*

(i) $B_1, \ldots, B_n$ *is* exhaustive *i.e.* $\bigcup_{i=1}^n B_i = \Omega$

(ii) $B_1, \ldots, B_n$ *is* mutually exclusive*, i.e.* $B_i \cap B_j = \emptyset$ *for all* $i \neq j$

Example 3.6. Algebra generated by a finite partition. Let $B_1, \ldots, B_n$ be a partition of Ω. The algebra $\mathcal{A}(B_1, \ldots, B_n)$ generated by this partition consists of the empty set and all possible unions of any number of elements of the partition. $\mathcal{A}$ consists of 2^n subsets of Ω. In particular, for finite Ω, the partition defined by all possible subsets of containing exactly one element results in $\mathcal{A}$ being the power set $\mathcal{P}(\Omega)$.

In this situation, any set in the algebra can be represented explicitly by elements of the partition. For any $A \in \mathcal{A}(B_1, \ldots, B_n)$ there is an index set $I \subset \{1, \ldots, n\}$ such that $A = \bigcup_{i \in I} B_i$. Note that this includes the empty set by using $I = \emptyset$.

The finest possible partition is the one consisting of all elements of Ω and the algebra it generates is $\mathcal{P}(\Omega)$. On the other extreme, the collection $(\emptyset, \Omega)$ is also an algebra.

Definition 3.7. Classical measurable space.
Let Ω be a non-empty set of points and $\mathcal{A} = \mathcal{A}(B_1, \ldots, B_n)$ the algebra generated by a partition $B_1, \ldots, B_n$ of Ω. The pair $(\Omega, \mathcal{A})$ is called a classical measurable space*. $B_1, \ldots, B_n$ are called* basic events*.*

Definition 3.8. Axioms for classical probability.
$(\Omega, \mathcal{A}, P)$ *is called* classical probability space *if the following two axioms are fulfilled:*

(C1) $(\Omega, \mathcal{A})$ *is a classical measurable space.*

(C2) $P : \mathcal{A} \longrightarrow [0, 1]$ *is a set function such that, for any event $A \in \mathcal{A}$ which compromises exactly k distinct basic events B_i, it is $P(A) = k/n$.*

All the classical examples from the last section, i.e. 3.1 to 3.3 with (3), are special cases of this definition:
$\Omega = \{\omega_1, \ldots, \omega_n\}, \mathcal{A} = \mathcal{A}(\{w_1\}, \ldots, \{w_n\}) = \mathcal{P}(\Omega), P(\{\omega_i) = 1/n$ $(i = 1, \ldots, n)$. The next example shows how we can use this framework when different observations are not equally likely.

Example 3.9. Balls in a bag. Draw a ball at random from a bag with k black balls and $n - k$ white balls. More generally, there are n_k $(k = 1, \ldots, K)$ balls of colour k in the bag.

We observe only the colour, but to build an equiprobable model we go back to outcomes.

$$
\begin{aligned}
&b_{k_i} = i\text{th ball of colour } k \ (i_k = 1, \ldots, n_k, k = 1, \ldots, K)\\
&\Omega = \{b_{k_i} \,|\, k = 1, \ldots, K,\ i = 1 \ldots, n_k\}\\
&\mathcal{A} = \mathcal{A}(\{b_{k_i}\}, k = 1, \ldots, K,\ i = 1 \ldots, n_k)\\
&P = (n_1 + \ldots + n_K)^{-1}
\end{aligned}
$$

The following model has infinitely many outcomes. Distinguishing only a finite number of them allows us to model it in the available framework.

Example 3.10. Spinner. A spinner looks like a clock with one hand. The hand is spun until it comes to rest. The random angle between a reference line and the hand is the observation, so a natural model. Fix $N \in \mathbb{N}$. Then the following model for the spinner defines a classical probability space.

$$
\begin{aligned}
&\Omega = [0, 2\pi),\ B_i = [(i-1)/N 2\pi, i/N 2\pi)\ (i = 1, \ldots, N),\ \mathcal{A} = \mathcal{A}(B_1 \ldots, B_N)\\
&\text{for } A \in \mathcal{A} \ \text{ with } A = \bigcup_{i \in I} B_i = |I|/N
\end{aligned}
$$

In examples with a finite outcome space the power set is the default choice for $\mathcal{A}$. The fact that you *could* define a probability measures on $\mathcal{P}(\Omega)$ does not mean you always *should.* Sometimes it is more appropriate to look at a smaller collection of subsets.

Definition 3.11. Submodel.
A measurable space $(\Omega, \mathcal{A}^*)$ *is called submodel of a measurable space* $(\Omega, \mathcal{A})$ *if* $\mathcal{A}^* \subset \mathcal{A}$.

One of the reasons to define P on a submodel $(\Omega, \mathcal{A}^*)$ of $(\Omega, \mathcal{P}(\Omega))$ is to reflect limited knowledge about probabilities. Say we know P for all $A \in \mathcal{A}^*$, but we do not have enough information to uniquely extend P to $\mathcal{P}(\Omega)$.

Example 3.12. Cards. The deck most often seen in English-speaking cultures, and common in other countries where the deck has been introduced, is the Anglo-American poker deck. This deck contains 52 unique cards in the four French suits, Spade (♠), Heart (♡), Diamond (♢) and Club (♣) and thirteen ranks running from two (deuce) to ten, Jack, Queen, King, and Ace. For simplicity, introduce the coding Ace=1, Jack=11, Queen=12, King=13 and choose Draw a card from the deck and observe the suit and the rank.

$$\Omega = \{Sk \,|\, S = ♠, ♡, ♢, ♣; k = 1, \ldots, 13\}.$$

The full model uses $\mathcal{P}(\Omega)$. The submodel obtained from observing only the suit is the algebra generated by the partition $(\{♠k \,|\, k = 1, \ldots, 13\}, \{♡k \,|\, k = 1, \ldots, 13\}, \{♢k \,|\, k = 1, \ldots, 13\}, \{♣k \,|\, k = 1, \ldots, 13\})$. The submodel obtained from observing only the number is the algebra generated by the partition $(\{Sk \,|\, S = ♠, ♡, ♢, ♣\}, k = 1, \ldots, 13)$. Both are submodels of $(\Omega, \mathcal{P}(\Omega))$, but they are not submodels of each other.

Example 3.13. Flip a coin twice. Use $\Omega = \{hh, ht, th, tt\}$. A submodel obtained from observing only the number of heads is $\mathcal{A}_1 = \{\emptyset, \Omega, \{hh\}, \{ht, th\}, \{tt\}\}$. A submodel obtained from observing only the second coin flip is $\mathcal{A}_2 = \{\emptyset, \Omega, \{hh, th\}, \{ht, tt\}\}$. They are both are submodels of the full model $(\Omega, \mathcal{P}(\Omega))$, but they are not submodels of each other.

Example 3.14. Heart attack risk. In a certain population, We would like to define a probability measure describing the probability for a heart attack that distinguishes between people who had risk factors and those who did not. However, we only know that from 100 people who regularly see the doctor, 30 turn out to have risk factors for a heart attack and that 20 in 100 do see the doctor on a regular basis. Define events
D="see the doctor on a regular basis"
R="show risk factors for a heart attack"
Based on the information, one can define a probability measure on the algebra generated by the partition $D \cap R, D \cap R^C$ and D^C. There is no way to distinguish $D^C \cap R$ and $D^C \cap R^C$.

Ω: Stop this equally likely approach to probability! It's not realistic.

ω: But it's more fair to give everybody the same chance.

Ω: I didn't know you are a communist.

ω: I'm not a commo. . . comprumist, I mean I'm not a compunist, eh commponist eh I don't know what I am not. But I like equally when they are equally like, because:
All you need is counts, count, counts, is all you need.
All you need is counts (all together now),
All you need is counts (everybody).

3.3 Some results from combinatorics

3.3.1 Counting

> "I think you're begging the question," said Haydock, "and I can see looming ahead one of those terrible exercises where six men have white hats and six men have black hats and you have to work it out by mathematics how likely it is that the hats will get mixed up and in what proportion. If you start thinking about things like that, you would go round the bend. Let me assure you of that!"
> — Agatha Christie, The Mirror Crack'd

Question 3.15. Campus residences.
A campus residence unit consists of three two-storey buildings with four rooms in each storey. What is the total number of rooms in this unit?

Answer: 3 buildings times 2 floors times 4 rooms makes 24 rooms all together.

An easy way to illustrate this is by drawing a three dimensional lattice with one dimension for the buildings, one dimension for the floors and another dimension for the rooms. Another way to visualise this is by sketching a tree shaped graph: 3 branches for buildings with each of them growing 3 branches for the floors which each grow 4 branches for the rooms.

Let us now formally state the counting methods we have just applied intuitively.

Theorem 3.16. Fundamental Rule.
Given a set of n_1 distinct elements $a_1, \ldots, a_{n_1}$; a set of n_2 distinct elements $b_1, \ldots, b_{n_2}$; up to a set of n_v distinct elements $x_1, \ldots, x_{n_v}$; there exist $\prod_{i=1}^{v} n_i$ distinct ordered v-tupels $(a_{i_1}, \ldots, x_{i_v})$.

ω: That's a funny theorem. It only applies to tupels of lengths 2 or 24!

Ω: Don't take everything so literally, omi. They just want to avoid even more indices, but you can use it for whatever long tupels you like. See, $a, \alpha, \beta, \gamma, \delta, b, c, \ldots, x$ makes 29-tupels.

Proof. By induction. True for $v = 2$, just form a rectangular array with pair (a_i, b_j) at the intersection of the ith row and the jth column. Now suppose it is true for $v - 1$. Expressing the v-tupel as a pair $((a_{i_1}, \ldots, w_{i_{v-1}}), x_{i_v})$ and applying the $v = 2$ case with $m = n_1 \cdot n_2 \cdot \ldots \cdot n_{i_{v-1}}$ and $n = n_v$ shows the claim for v. □

Question 3.17. Group picture.
You want to take a picture of seven people arranged on chairs in a row. How many choices do you have?

Answer: Put the seven chairs in a row. The first person has the choice between seven chairs. For the second person, there are six chairs left. (Note: For the *number* of choices it does no matter which chair the first person chose.) The the third person, there a five chairs left. And so on. The sixth person only has a choice between two chairs, and the seventh person has no choice at all. So, the total number of arrangements is $7 \cdot 6 \cdot 5 \cdot \ldots \cdot 2 \cdot 1 = 5040$.

ω: But I want to be next to you!

Ω: Why don't you figure out how many possibilities we have when we restrict the arrangements to those where you sit next to me.

ω: I think that easier if we sit in a circle.

Definition 3.18. Permutation.
A permutation *of a finite set is any ordering of its elements in a list.*

Ω: Actually, this it not what I learned in my algebra class. They said, a permutation of a finite set is a bijective map of that set onto itself.

ω: But it's all the same, Omega!

Theorem 3.19. Total number of permutations.
A set with n elements has $n!$ different permutations.

Proof. Let $a_1, \ldots, a_n$ be the n distinct elements of the set. We compute the number of permutations following the idea of the calculation in Question 3.17. There are n choices for a_n to go, $n - 1$ choices for a_{n-1}, $n - 2$ choices for a_{n-2}, and so on, up to 2 choices for a_2 and just one option for a_1. Using the fundamental rule, that makes $n!$ different orderings. □

3.3.2 Sampling

Ω: I am collecting a collection.

Definition 3.20. Sampling types.
Suppose we sample r from n distinct elements in a pool. The sampling method *is called*

- with replacement *if, any element that was drawn is replaced and may be drawn again;*
- without replacement *if, once drawn, an element cannot be drawn again.*

The record (basic event) *is called*

- ordered *if the order of the draw is recorded (use notation* $(\ldots)$*);*
- not ordered *if the order of the draw is not recorded (use notation* $\{\ldots\}$*).*

Typically, we think of experiments involving tossing coins or rolling dice as sampling with replacement. Simply, because whatever side the coin or the die faces it will not come off but still be there for the next toss. In experiments involving drawing balls from a bag or tickets from a box, we have to specify whether or not we are going to throw them back or not.

Example 3.21. First, second and third prize. A first, second and third prize are given to three randomly selected people from a group of 200. This is a description in everyday language. To make it mathematically precise we state the implicitly made assumptions: The three prizes are different and each person can at most win one prize. In other words, this is an **ordered** sample of size $r = 3$ drawn at random **without replacement** from a pool of size $n = 200$. The number of possibilities can be calculated by considering the number of options in the stepwise process of selecting the winners, so the total number of possibilities amounts to $200 \cdot 199 \cdot 198 = 7,880,400$.

Example 3.22. Three prizes. Three identical prizes are given to three randomly selected people from a group of 200. We assume: Each person can win at most one prize. In contrast to before, we do not record which of the three winners got which of the three prizes. Hence this is a **not ordered** sample of size $r = 3$ drawn **without replacement.** In other words, any possibilities that only differ by the order would be identified. Since there are 3! possible ways to order the three winners, the total number of possibilities can be obtained by dividing the result from (i) by 3! which makes $1,313,400$.

Example 3.23. Tossing a coin three times. Toss a coin three times. A mathematically precise description would say that we take an **ordered** sample of size $r = 3$ at random **with replacement** from a pool of $n = 2$. The number of possibilities is $2 \cdot 2 \cdot 2 = 8$. This can be visualised by a binary tree.

ω**:** *(singing)* I do anything for you dear, anything...

Ω**:** Can you make me a coin that samples without replacement?

Example 3.24. Tossing three coins. Toss three coins at once. In contrast to the previous example, this description suggests that no record of the order is kept. This is a **not ordered** sample of size $r = 3$ at random **with replacement** from a pool of $n = 2$.

Example 3.25. Lottery. Draw 6 out of 49 numbered balls at random. The balls are drawn by a mechanical system one by one. Once drawn, a ball is not replaced into the machine. Whereas the mechanism of drawing the balls implies an order, it is not taken into account for the evaluation. To win the highest prize of the lottery one needs to have guessed all the 6 numbers correctly. Hence, this is a **not ordered** sample of size $r = 6$ from a pool of size $n = 49$. drawn at random **without replacement.** Following the reasoning from Example 3.22 we obtain that the total number of possibilities is $(49 \cdot 48 \cdot \ldots \cdot 44)/6! = 13,983,816$.

Ω**:** The chance to get this right is only 1 in 13,983,816. Why would you play the lottery?

ω**:** Because I am special, don't you see Omega?

Ω**:** What I see is that you may want to take the module *Games and Decisions.*

Theorem 3.26. Total number of samples.
Suppose we sample $r \geq 1$ from n distinct elements in a pool $\{e_1, \ldots, e_n\}$. Let $O(r,n)$ be the number of distinct ordered records and $N(r,n)$ be the number of distinct not ordered records.

(i) Suppose the sampling methods is with replacement. Then

$$O(r,n) = n^r \qquad \text{and} \qquad N(r,n) = \binom{n+r-1}{r}.$$

(ii) Suppose the sampling methods is without replacement and $r \leq n$. Then

$$O(r,n) = \frac{n!}{(n-r)!} \qquad \text{and} \qquad N(r,n) = \binom{n}{r}$$

Proof.

(ii) (without replacement):

For ordered records we have n choices for the first element, $n-1$ for the second, $n-2$ for the third and so on. Using the Fundamental Rule Theorem 3.16,

$$O(r,n) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1) = n!/(n-r)!$$

Each subset of r distinct elements can produce exactly $r!$ distinct ordered records, which proves the formula for $N(r,n)$.

(i) (with replacement):

For ordered records we have n choices for each of the elements we draw. Using the Fundamental Rule 3.16, $O(r,n) = n^r$.

To compute $N(r,n)$ we will use that there is a bijection between the following two sets

- the set of unordered records of the form $\{e_{i_1}, \ldots, e_{i_r}\}$
- the set of ordered n-tuplets $(N_1, N_2, \ldots, N_n)$, where N_i is the number of times element e_i is drawn for $i = 1, \ldots, n$.

It remains to compute the size $\tilde{N}$ of the latter set. Note that $N_1 + \ldots + N_n = r$, so $\tilde{N}$ equals the number of ways to place $n-1$ vertical lines between r dots. The dots to the left of the first line correspond to N_1, and for $i = 2, \ldots, n-1$ the dots between the $(i-1)$th and the ith line correspond to N_i. The dots to the right of the $(n-1)$th line correspond to N_n. To compute the number of possibilities, think of this as $r+n-1$ dots of which $n-1$ are chosen at random, not ordered and without replacement, to be changed to lines. According to (ii), this yields $\tilde{N} = \binom{r+n-1}{n-1} = \binom{n+r-1}{r}$. In the last step we use $n+r-1-(n-1) = r$ and $\binom{m}{k} = \binom{m}{m-k}$

□

Remark 3.27. To visualise the last steps look at an example: $r = 4, n = 6$. This make $r+n-1 = 9$ dots of which 5 are to be replaced by vertical lines. The pattern

$$\bullet \;|\; | \;\bullet\; | \;|\; \bullet \;\bullet\; |$$

correspondes to $N_1 = 1, N_2 = 0, N_3 = 1, N_4 = 0, N_5 = 2, N_6 = 0$, and the pattern

$$| \;\bullet\; \bullet \;|\; | \;|\; \bullet \;|\; \bullet$$

correspondes to $N_1 = 0, N_2 = 2, N_3 = 0, N_4 = 0, N_5 = 1, N_6 = 1$.

3.3.3 Balls in buckets

In how many ways can r balls be placed into n buckets? This is a popular way of asking about the number of diophantine solutions (i.e. in $\mathbb{N}$) to the equation $x_1 + \ldots + x_n = r$. A variation is to ask the same question under the additional condition that no bucket is empty.

Theorem 3.28. Occupancy theorem.
In allocating r indistinguishable elements to n coordinates of a record the number N of distinct allocations is

(i)
$$\binom{n+r-1}{r}.$$

(ii) If $r \geq n$ and no position to be left unoccupied,
$$\binom{r-1}{r-n}.$$

Proof. (i) For $i = 1, \ldots, n$ let N_i = number of balls in the ith bucket, so $\sum_{i=1}^{n} N_i = r$.
$(*)$ Take $r + n - 1$ dots and change $n - 1$ of them into lines.
N_1 = number of dots left of line 1.
N_i = number of dots between line $i - 1$ and line i $(i = 2, \ldots, n - 1)$.
N_n = number of dots right of line n.
N equals the number of ways in which $(*)$ can be done. How many are there?
This is sampling $n - 1$ from $r + n - 1$ without replacement. It is not ordered, because the order in which we pick the dots to be changed into lines does not matter. Hence this yields
$$\binom{r+n-1}{n-1} = \binom{r+n-1}{r}.$$
(ii) First of all, place n balls into n buckets to satisfy the condition that no bucket is to be left empty. For the remaining $r - n$ balls use (i) and get
$$\binom{n+(r-n)-1}{r-n} = \binom{r-1}{r-n}.$$

□

ω: What about eggs in baskets? Or students in sections? Or sweets into children, seeds into pits in mancala, or money into bank accounts, or...

Ω: Hold on! Tell me, why are people so keen on this *balls in buckets* or *balls in bin* metaphor? There is no reason take for granted that buckets or bins are ordered. In fact, the *raison d'être* of a bucket is that it can easily be moved around, in particularly with something in it! The occupancy theorem, however, is about allocating objects to coordinate positions of a record. I would like to see that renamed to something students in classrooms.

3.3.4 Combinatorics in statistical physics.

This section is just for people interested in physics and its connection to probability theory; it is not necessary for understanding probability theory, let alone for succeeding at

the exam. There is a number of models representing particles in a space describing their position and momentum. One possible starting point is a finite number r of indistinguishable particles a state space partitioned in a finite number n of distinct cells. The total number of arrangements of the particles over the cells is given by Theorem 3.28(i) using particles for balls and cells of buckets. The uniform distribution on the space of all such arrangements is called *Bose-Einstein statistics.* It has been experimentally verified for bosons, that is, particles with integer spin values such as photons and mesons.

Assume there are n_l cells at energy level l, $(l = 1, \ldots, L)$, so $\sum_{l=1}^{L} n_l = n$. The *microscopic state* of the system is a record of the states of all the individual particles; the *macroscopic state* is a record of the numbers of particles at the different energy levels. Imagine that the measurement technologies only enable us to observe the macroscopic state, even though the particles really are distinguishable. The distribution of their energy level allocation is affected by whether or not they are regarded as distinguishable or not.

Now assume the particles are distinguishable. However, they are still assumed to be similar in that they are equally likely to allocated to each of the cells. The outcome space for the unobservable microscopic state of the system is $\Omega_{\text{micro}} = \{1, \ldots, n\}^r$. In other words, a microscopic state is an arrangements of r distinguishable particles in n cells. For each particle a cell is chosen. This corresponds to sampling with replacement (several particles allowed per cell) of r cells from a set of n cells. By symmetry (there is no reason why any of the states should be preferred over another), the distribution is uniform on Ω_{micro}.

The outcome space for the observable macroscopic state is $\Omega_{\text{macro}} = \{1, \ldots, r\}^L$. In other words, the macroscopic state is a record of the numbers of particles allocated to different energy levels. For each of the r particles, on of the n cells is chosen it is recorded to which of the L energy levels that cell belongs. This corresponds to a multinomial distribution with parameters r for the number of draws (with replacement) and n_l/n $(l = 1, \ldots, L)$ for the likelihoods of the types. The validity of this distribution is a classical assumption in statistical mechanics named *Maxwell-Boltzmann statistics* referring to these scientists work on the topic.

The *Fermi-Dirac statistics* refers to the same set-up but incorporates *Pauli's exclusion principle* which says that there is at most one particle allowed per cell. This applies to Fermions, that is, particles with half-integer spin values such as electrons, proton and neutrons. The exclusion principle obviously requires $n \geq r$, and the outcome space for the microscopic states has to be adjusted according to the additional constraint. For each particle, a cell is chosen and no other particles are allowed in that cell. This corresponds to sampling without replacement of r cells from a set of n cells.

The outcome space for the macroscopic states remains unchanged. As before, for each of the r particles, on of the n cells is chosen it is recorded to which of the L energy levels that cell belongs, but now it is not allowed to use a cell more than once. This corresponds to a hypergeometric distribution with parameters r for the number of draws (without replacement) and n_l $(l = 1, \ldots, L)$ for the likelihoods of the types.

One might think that Pauli's exclusion principle is only relevant if there is a small number of cells per energy level. That intuition is actually backed up by a mathematical theorem which states that, for any $r \in \mathbb{N}$ fixed, the hypergeometric distribution with parameters r and n_l $(l = 1, \ldots, L)$ converges to the multinomial distribution with parameters r and p_l $(l = 1, \ldots, L)$ for $n \to \infty$ and $n_l/n \to p_l$ $(l = 1, \ldots, L)$.

4 Axioms of mathematical probability

The set-up for probability developed in Chapter 3 is build on top of a basis of equally likely events representing the classical thinking about probabilities until about the early 20th century. We will discuss the limitations of this approach in the next section and devote the following three sections to an introduction to the modern axiomatic approach to probability. We conclude this chapter by introducing cumulative distribution functions as a way of representing probability measures on the real line (and subsets of the real line).

The axioms of mathematical probability go back to the work of the young Russian mathematician Andrey Kolmogorov. They were laid out in his seminal work on the foundations of the theory of probability[1]. The text was written under the influence of modern set theory he had studied during his stays as Göttingen, Munich and Paris and published after his return to Moscow. In the same period of time, other people offered axiom system of a similar flavour; the time was ripe for what became the platform of modern probability theory, the a set theoretic formalisation for the already well developed and applied existing calculus of probability. The place of Kolmogorov's work in the philosophy and history of mathematics is reviewed in a recent article by Vladimir Vovk and Glenn Shafer[2].

4.1 Uses and limitations of the classical approach to probability

We will go through a few examples to discuss our current approach to probability.

Example 4.1. Tossing a coin three times. Standard outcome space is $\Omega_3 = \{\omega_1, \omega_2, \omega_3 \,|\, \omega_i = h, t\}$. Using Theorem 3.26 $n = 2$ and $r = 3$, with replacement and ordered, yields $|\Omega_3| = 2^3 = 8$. Standard model uses $\mathcal{A} = \mathcal{P}(\Omega_3)$, i.e. the algebra generated by the partition induced by all outcomes.

$$A_1 = \text{“all heads”} = \{hhh\}, P(A_1) = |A_1|/|\Omega_3| = 1/8$$
$$A_2 = \text{“second toss is heads”} = \{hhh, hht, thh, tht\}, P(A_2) = |A_2|/|\Omega)3| = 1/2$$

$|A_2|$ can be calculated using Theorem 3.26 for $n = 2$ and $r = 2$, with replacement and ordered, because one position being fixed leaves a reduced version of the original case ($n = 2$ and $r = 3$).

Example 4.2. Tossing a coin three times. Outcomes with the same number of heads are indistinguishable, so a natural outcome space would be $\Omega_h = \{0, 1, 2, 3\}$. However, these outcomes are not equally likely. We can still handle this situation by simply going

[1] Andrey Kolmogorov, “Grundbegriffe der Wahrscheinlichkeitsrechnung”, Berlin: Julius Springer, 1933. Translation: “Foundations of the Theory of Probability” (2nd ed.), New York: Chelsea, 1956, `www.mathematik.com/Kolmogorov/index.html`

[2] Vladimir Vovk and Glenn Shafer, “The Sources of Kolmogorov's *Grundbegriffe*, Statistical Science, Vol. 21 (2006), No. 1. `www.probabilityandfinance.com/articles/STS154.pdf`. Longer version: “Kolmogorov's contributions to the foundations of probability”, Problems of Information Transmission 39 (2003), `www.probabilityandfinance.com/articles/06.pdf`.

back to the outcome space Ω_3 from Example 4.1.

$$P(0) = P(\{ttt\}) = 1/8$$
$$P(1) = P(\{tth, tht, htt\}) = 3/8$$
$$P(2) = P(\{thh, hth, hht\}) = 3/8$$
$$P(3) = P(\{hhh\}) = 1/8$$

Example 4.3. First, second and third prize. There are N people from which to select three to give the prizes to. We assume this is without replacement, because no person is supposed to win more than once, and we assume the order matters, because the prizes are different. Hence we model it by $\Omega = \{(i, j, k) \,|\, i, j, k = 1, \ldots, N, i \neq j, j \neq k, k \neq i\}$ and $\mathcal{A} = \mathcal{P}(\Omega)$. $\Omega = \frac{N!}{(N-3)!}$ using Theorem 3.26 for $n = N$ and $r = 3$. Now assume $N = 10$ and 4 of the are in the last row.

$A =$ "all winners are from the last row" $\qquad P(A) = |A|/|W| = 4!/10 \cdot 9 \cdot 8 = 1/30$

Example 4.4. . Same assumptions as in Examples 4.3, except we assume the order does not matter, because the prizes are all the same. Again, to actually perform calculations we go back to the ordered model in Example 4.3.

Ω**:** We keep seeing examples were the records are not naturally ordered, but then go on to an ordered model. I don't understand this at all.

ω**:** But I do. My teacher does this. Whenever she wants to count us at a field trip, she shouts 'Don't move!!!'

Example 4.5. Biased coin. How can we model a coin that does not show heads and tails equally often? It would be natural to use $\Omega = \{h, t\}$ and $\mathcal{A} = \mathcal{P}(\Omega)$ as in the case of a fair coin, but this would automatically constrain us to the case of a fair coin.

Start with a simply case, a coin that shows heads twice as often as tails. Use $\Omega = \{h_1, h_2, t\}$, with h_1, h_2 both denoting heads. $\mathcal{A} = \mathcal{P}(\Omega)$ and $P(\omega) = 1/3$ for all $\omega \in \Omega$.

For any bias that can be expressed as a rational number we can proceed the same way. To model a coin that, on average, shows heads k times in l tosses, $k \leq l$, use $\Omega = \{h_1, \ldots, h_k, t_1, \ldots, t_{l-k}\}$ and $\mathcal{A} = \mathcal{P}(\Omega)$. Again, identify $h_1 = \ldots = h_k = h$ and $t_1, \ldots, t_{l-k} = t$ and set $P(\omega) = 1/l$ for all $\omega \in \Omega$. Any other bias, can be approximated arbitrarily closely by a rational number.

Example 4.6. Coin tossing until heads come up. A coin is tossed until heads come up. Any particular outcomes observed consists of a finite number of tails and one heads at the end, so we can choose $\Omega = \{h, th, tth, ttth, tttth, ttttth, \ldots\}$. Note that while there are infinitely many different outcomes, each individual outcome is finite. Ω is countable. To proof that you could use the one-to-one map from Ω to $\mathbb{N}$ given by the number of tails in each $\omega \in \Omega$. Examples for events are:

$$\text{"it takes 3 trials to get a head"} = \{tth\}$$
$$\text{"it takes at least 3 trials to get a head"} = \{tth, ttth, tttth, \ldots\}$$
$$\text{"it takes at most 3 trials to get a head"} = \{h, th, tth\}$$

Their are infinitely many outcomes. One way to get around that is to introduce and upper limit N of tosses allowed, e.g. $\Omega_N = \{h, th, tth, \ldots, t \ldots th\}$ with the the last outcome there having length N. In each of the W_N, however, the outcomes are not equally likely requiring a strategy along the lines of Example 4.5.

Example 4.7. Spinner. We have discussed finite models for the spinner in Example 3.10. This can be done arbitrarily fine by increasing the number N of segments of the circle. However, we end up distinguishing a finite number of classes for a continuum of possible angles.

In summary, the classical approach imposes too many constraints:

- Being tied to equally likely probabilities dictates the choice of the outcome space. In Example 4.2 it would be convenient of have a probability measure P defined on the most natural outcome space Ω_h rather than making the detour of going back to Ω_3 from Example 4.1.
- Unequal probabilities in examples with finitely many outcomes can be modelled in the classical approach along the lines of Example 4.5, but it potentially gets rather lengthy and may only result in an approximate model.
- Infinitely many outcomes can only be modelled by finite approximations, along the lines discussed in Examples 4.6 and 4.7.

We will now develop a generalised framework for mathematical probability.

4.2 Which collection of subsets?

A probability measure is defined on a collection of subsets (describing the events) of the outcome space. First we need a suitable structure of sets. This is going to be only a slight extension of the set algebra structure in Definition

Definition 4.8. σ-algebra of sets.
Let Ω be a set of points. A system $\mathcal{A}$ of subsets of Ω is called σ-algebra if

(A1) $\Omega \in \mathcal{A}$

(A2)* $A_i \in \mathcal{A}$ $(i \in \mathbb{N}) \implies \bigcup_{i\in\mathbb{N}} A_i \in \mathcal{A}$

(A3) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$

The pair $(\Omega, \mathcal{A})$ is called measurable space.

To see what is the point of extending (A2) in the definition of an algebra to the form (A2)* to cover countably infinite unions think of Example 4.6.

$$\text{“first heads in } i\text{th toss”} = A_i$$
$$\text{“it takes an even number of times until first heads”} = \bigcup_{i\in\mathbb{N}} A_i$$

Example 4.9. σ-algebra generated by a partition. A system B_i $(i \in \mathbb{N})$ of subsets of Ω is called partition if it is *exhaustive* (i.e. their union is Ω) and *mutually exclusive* (i.e. $B_i \cap B_j = \emptyset$ for all $i \neq j$). The σ-algebra $\sigma(B_i, i \in \mathbb{N})$ generated by this partition consists of the empty set and all possible unions of any number of elements of the partition.

In this situation, any set in the σ-algebra can be represented explicitly by elements of the partition. For $A \in \sigma(B_i, i \in \mathbb{N})$, there is an index set $I \subset \mathbb{N}$ such that $A = \bigcup_{i\in I} B_i$. Note that this includes the empty set by using $I = \emptyset$.

But what is the point in defining a structure like a σ-algebra when we could just use the power set $\mathcal{P}(\Omega)$, that is the set of *all* subsets of Ω?

One reason was already discussed in Section 3.2. It is sometimes more appropriate to use submodel of $(\Omega, \mathcal{P}(\Omega))$. In particular in situations where the probability is only known for a system of subsets.

The other reason for not always using $\mathcal{P}(\Omega)$, however, is that is that paradoxes, or at least surprising hurdles, are in store when dealing with a not countable Ω. Even in situations that look as harmless as the following one.

Let $\Omega = \{0,1\}^{\mathbb{N}}$. This is an outcome space for infinitely many coin tosses. If P was a probability measure on Ω modelling infinitely many tosses by a fair coin, some of the properties we would expect are:

- $P(\Omega) = 1$.
- P is additive, and the additivity should also work out for countably infinite numbers of events.
- The probability of an event does not change under an operation that flips the outcome of a certain toss. For example, $P($"forth toss heads and fifth toss heads"$) = P($"forth toss tails and fifth toss heads"$)$.

The Guiseppe Vitali proved in 1905, that there is no such P. The proof is rather elementary, but it does require the axiom of choice[1].

Similar issues arise in attempts to construct measures on other spaces. Felix Hausdorff showed in his classic text about the foundations of set theory[2] essentially the following: The surface of the unit ball in $\mathbb{R}^3$ can be decomposed in three parts A, B and C in a way such that each of them is congruent with each of the others (and can be moved into exactly the same position by a 120 degree rotation around a suitable axis), while each of them is also congruent to the union of the other two (and can be moved into the same position by a 180 degree rotation around a suitable axis). Hence, each of these sets is both a half and a third of the surface of the unit ball. Assuming there was a rotation invariant measure assigning the value 1 to the surface of the unit ball, the measure of the sets A, B and C would have to be both $1/2$ and $1/3$. Hence the construction of such a measure is not possible[3].

The result is known as *Hausdorff paradox.* More general and deaper results of the same flavour of have been published by Stefan Banach and Alfred Tarski[4] and have become widely known under the name *Banach-Tarski paradox.* As summarised in by John von Neumann[5], one palpable consequence of their results is that one could slice up a unit ball in $\mathbb{R}^3$ into 9 pieces such that, after some suitable rotations, five of them put together again

[1] See, for example, Hans-Otto Georgii, "Stochastics, Introduction to Probability and Statistics", de Gruyter, 2008

[2] "Grundzüge der Mengenlehre", Verlag Veit & Co, Leipzig, 1915. Reprints: Chelsea Pub. Co. (1949, 1965, 1978), also in Hausdorff-Edition Vol. II (2002), edited by E. Brieskorn, F. Hirzebruch, W. Purkert, R. Remmert, E. Scholz.

[3] From a summary in John von Neumann, "Zur allgemeinen Theorie des Masses", Fundamenta Mathematicae 13 (1929). `matwbn.icm.edu.pl/ksiazki/fm/fm13/fm1316.pdf`.

[4] Stefan Banach and Alfred Tarski, "Sur la décomposition des ensembles de points en parties respectivement congruentes", Fundamenta Mathematicae 6 (1924). `matwbn.icm.edu.pl/ksiazki/fm/fm6/fm6127.pdf`.

[5] As 3.

make a unit ball, and the remaining four make another unit ball.

ω: This is an arbitrage opportunity, Omega. We can get 1/2 for the price of a 1/3.

Ω: My new financial smart product *SphereShares* and *BallBash*. I am selling the right to slice up a unit sphere or cut up a unit ball and put them together again with the promiss they'll be bigger than. I'm hoping to pay back my student loan that way.

The results by Vitali, Hausdorff and Banach and Tarski remain a bit mysterious, because our imagination for subsets is somewhat limited. Anything subsets of $\mathbb{R}$ people typically can think of is contained in the σ-algebra generated by open intervals (named after Emile Borel) which is the default set system for defining probabilities on $\mathbb{R}$.

Example 4.10. Borel σ-algebra. The *Borel σ-algebra* on $\mathbb{R}$ is smallest σ-algebra on $\mathbb{R}$ that contains all open intervals. By Definition 4.8 it automatically includes all closed intervals (using (A3)), all countable unions of open and closed intervals (using (A2*), all countable intersections of open and closed intervals (using (A2*) and (A3)). Iterating those mechanisms one obtains a lot more, such as unions of intersections of unions of intersections of open sets and so on. In particularly, half open intervals, infinite intervals and sets with only one point in it:

$$[a,b) = \bigcup_{n=1}^{\infty} \Big[a, b - \frac{1}{n}\Big], \qquad [a,\infty) = \bigcup_{n=1}^{\infty} [a, a+n], \qquad \{a\} = \bigcap_{n=1}^{\infty} \Big(a - \frac{1}{n}, a\Big].$$

Short notation is $\mathcal{B}(\mathbb{R})$. By the same token, Borel σ-algebras can be defined on $\mathbb{R}^n$ by replacing open and closed intervals with open and closed n-dimensional balls.

4.3 Mathematical probability

Definition 4.11. Axioms for mathematical probability.
Let $(\Omega, \mathcal{A})$ be a measurable space. A set function $P : \mathcal{A} \longrightarrow [0,1]$ is called probability measure *if the following two axioms are fulfilled:*

(P1) $P(\Omega) = 1$.

(P2)* *P is* σ-additive, *i.e., for all $A_i \in \mathcal{A}$ $(i \in \mathbb{N})$ with $A_i \cap A_j = \emptyset$ for any $i \neq j$,*

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(A_i).$$

The triple $(\Omega, \mathcal{A}, P)$ is called probability space.

Example 4.12. Coin tossing (with bias) aka Bernoulli experiment. This is the simplest possible random experiment that is not trivial. There are two outcomes, often called 0 and 1 with the latter referred to as "success". There is a fixed $p \in [0,1]$, the probability for success. Use $\Omega = \{0,1\}$ and the σ-algebra generated by the partition $(\{0\},\{1\})$, that is, $\mathcal{A} = \{\,\emptyset, \{0\}, \{1\}, \{0,1\}\,\}$. Then $P(\{1\}) = p$ uniquely defines a probability measure by $P(\{0\}) = 1 - P(\{1\}) = 1 - p$.

ω: I want $p = 1$, because then I've won! I always win... always, always, always!

Ω: Obviously, this case is very boring, and so is the case $p = 0$. They are included in the model, but they are the extreme cases. Such experiments are not actually random, they are *deterministic*.

Example 4.13. Finitely many outcomes. Given weights $p_i \geq 0$ for $i = 1, \ldots, n$ with $p_1 + \ldots + p_n = 1$,

$$P(\{w_i\}) = p_i \qquad (i = 1, \ldots, n) \tag{4}$$

defines a probability measure on $\Omega = \{w_1, \ldots, \omega_n\}$ with σ-algebra $\mathcal{P}(\Omega)$.
If $p_i = 1/n$ for all $i = 1, \ldots, n$ then this is called the *uniform distribution.*
A few explicit situations where it arises naturally:

(i) An n-faced die with p_i specifying the probability for the ith face to show up. If $p_i = 1/n$ for all $i = 1, \ldots, n$ then the die is *fair.* Otherwise it is *loaded.* (This refers to methods of inserting small quantities of metal into some of the sides of the die to increase the likelihood it lands the other side up.)

(ii) Drawing a ball at random from a box with a finite number of coloured balls. With m different colours and N_i $(i = 1, \ldots, m)$ balls of colour i, $N = N_1 + \ldots + N_m$, the probabilities are $p_i = N_i/N$.

Example 4.14. Countable number of outcomes. Given weights $p_i \geq 0$ $(i \in \mathbb{N})$ with $\sum_{i=1}^{\infty} p_i = 1$,

$$P(\{w_i\}) = p_i \qquad (i = 1, 2, \ldots) \tag{5}$$

defines a probability measure on $\Omega = \{w_1, w_2, \ldots\}$ with σ-algebra $\mathcal{P}(\Omega)$.

ω: I need to invite infinitely many friends to my birthday party. Because, whenever I think I will invite these n friends I realise I forgot one! I want to have a game at the party where everybody has the same chance of winning. What are the p_i $(i \in \mathbb{N})$ to make that happen?

Ω: I'm afraid you can not find such a game. Also, you just can not have infinitely many friends. Look, at every given time you only have a finite number of friends. Actually, there are only finitely many people in the whole world, omi!

ω: But you forgot my imaginary friends, Omega.

Remark 4.15. No uniform distribution on a countable set. There is no uniform distribution on a countably infinite set. Why? If there was a uniform distribution than there would be a $c \in [0, 1]$ such that $P(\omega_i) = c$ for all $i = 1, 2, \ldots$ By the properties of a measure, $P(\Omega) = \sum_{i=1}^{\infty} c$ (that means adding up c infinitely often). If $c > 0$ the series diverges and if $c = 0$ the series equals 0. Both cases contradict the axiom that $P(\Omega) = 1$.

Ω: In Section 4.5 you will learn how to make this work for a continuum of friends.

ω: But I do not *have* a continuum of friends. Even if I added all my imaginary ones.

Ω: Try nextgenerationfacebook.

Here is a general method how to create a probability measure.

Lemma 4.16. Probability measure defined on a partition.
Let B_i $(i \in \mathbb{N})$ be a partition of Ω and $\mathcal{A} = \sigma(B_i, i \in \mathbb{N})$. Let $p_i \geq 0$ with $\sum_{i \in \mathbb{N}} p_i = 1$. Using the representation in Example 4.9, set $P(A) = \sum_{i \in I} p_i$.

Proof. We need to show (P1) and (P2). About (P1):

$$P(\Omega) = P(\bigcup_{i \in \mathbb{N}} B_i = \sum_{i \in \mathbb{N}} p_n = 1.$$

About (P2)*: Let $A_i \in \mathcal{A}$ $(i \in \mathbb{N})$ with $A_i \cap A_j = \emptyset$ for any $i \neq j$. Then there are $I^{(i)} \subset \mathbb{N}$ $(i \in \mathbb{N})$ with $A_i = \bigcup_{n \in I^{(I)}} B_n$, $(i \in \mathbb{N})$. Hence, using $I = \bigcup_{i \in \mathbb{N}} I^{(i)}$,

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = P\Big(\bigcup_{n\in I} B_n\Big) = \sum_{i\in I} p_n = \sum_{i\in\mathbb{N}} \sum_{n\in I^{(i)}} p_n = \sum_{i\in\mathbb{N}} P(A_i)$$

□

4.4 Deductions from the axioms

The axioms really come to live by all the useful rules that can be derived from them.

(i) **Empty sets:** $P(\emptyset) = 0$

(ii) **Finite additivity:** For all $A_i \in \mathcal{A}$ $(i = 1, \ldots, n)$ with $A_i \cap A_j = \emptyset$ for any $i \neq j$,

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = \sum_{i=1}^{n} P(A_i).$$

(iii) **Complement:** $P(A^c) = 1 - P(A)$

(iv) **Partition by another event:** $P(A) = P(A \cap B) + P(A \cap B^c)$

(v) **Difference:** $P(A \setminus B) = P(A) - P(A \cap B)$

(vi) **Subset:** $B \subset A \Longrightarrow P(A \setminus B) = P(A) - P(A \cap B)$

(vii) **Monotony:** $B \subset A \Longrightarrow P(B) = P(A)$

(viii) **Addition rule:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(ix) **Subadditivity:**

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) \leq \sum_{i\in\mathbb{N}} P(A_i)$$

Note that the last two statements are about the union of sets that are not assumed to be disjoint. There is also an addition rule for more than two sets (see exercise sheets).

Proofs:
(i) Apply σ-additivity to the mutually exclusive sets $A_i = \emptyset$ $(i \in \mathbb{N})$

$$P(\emptyset) = P\Big(\bigcup_{i\in\mathbb{N}} \emptyset\Big) = \sum_{i\in\mathbb{N}} P(\emptyset).$$

$c := P(\emptyset) \in [0, 1]$ by definition. If $c > 0$ then the series diverges, hence $c = 0$.

ω: This sounds like a joke about mathematicians. They come up with the most complicated proof for something that is obvious to everybody else!

Ω: And then they complain about headlines like "Mathematicians proved fact known to vegetable grocers for centuries" after Thomas Hales tackled Kepler's conjecture in 1998.

(ii) Define $\tilde{A}_i = A_i$ for $i = 1, \ldots, n$, and $\tilde{A}_i = \emptyset$ for $i \geq n+1$ and apply (P2)* (σ-additivity) and (i).

(iii) $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$, hence $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$ using (ii).

(iv) $(A \cap B) \cap (A \cap B^c) = \emptyset$ and $(A \cap B) \cap (A \cap B^c) = A$, hence $P(A) = P((A \cap B) \cap (A \cap B^c)) = P(A \cap B) + P(A \cap B^c)$ using (ii).

(v) $A \setminus B = A \cap B^c$ and use (iv).

(vi) Use (v).

(vii) Follows from (vi).

(viii) Using (iv),

$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B^c) \\ P(A) &= P(A \cap B^c) + P(A \cap B) \\ P(B) &= P(A^c \cap B) + P(A \cap B) \end{aligned}$$

Comparing terms completes the proof.

(ix) Construct mutually exclusive sets $B_1 = A_1$, $B_2 = A_2 \cap A_1^c$, and $B_i = A_i \cap A_{i-1}^c \cap \cdots \cap A_1^c$. Note that their union equals the union of the original sets A_i ($i \in \mathbb{N}$). Hence, using σ-additivity,

$$P\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} P(B_i).$$

Using monotony shows that $P(B_i) \leq P(A_i)$. □

4.5 Probability measures on $\mathbb{R}$ or subsets of $\mathbb{R}$

4.5.1 Random numbers on $[0, 1]$

How do you pick a digital number between 0 and 1 at random? Different approaches have been practiced to do that in practice:

- Use a spinner that creates a random angle which can be converted into a number between 0 and 1. Measuring the angle will work up to some degree of precision, so only finitely many numbers in $[0, 1]$ can be covered. This corresponds to the finite models discussed for the spinner in Example 3.10. However, they are evenly distributed over the entire interval and with increasing the precision of the measurement instruments will refine the outcome space.
- Toss a coin n times recording 1 for heads and 0 for tails to create a sequence $x_1, x_2 \ldots x_n$ with $x_i \in \{0, 1\}$ $(i = 1, \ldots, n)$ Interpret this as coefficients of a number in base 2, i.e. $x = \sum_{i=1}^{n} x_i 2^{-i}$ and convert that number into base 10. You will only flip the coin a finite number n of times, hence your outcome space will not actually be $[0, 1]$, but just $\{0, 1\}^n$.
- Between the 1920s and the 1950s random number tables were produced using a mixture of manuel and automatic procedures.
- With the rise of computer sciences, *pseudo-random number generators* replaced the tables. They use initial values together with iterations of certain suitable functions. The mathematical background for this is a branch of analytic number theory in concerned with uniform distributions mod 1. Operating on the basis of numbers with finitely many decimal places, computer based generation of random numbers will always be restricted to a finite outcome space.

All practiced methods are constraints to finite outcome spaces, though each may be tailored to cover $[0, 1]$ to any desired degree of precision. In these finite outcome spaces, each

outcome has the same probability. This is what we understand by "picking a number at random".

Now we build an idealised model using the outcome space $\Omega = [0, 1]$. For any $\omega \in \Omega$, $P(\{\omega\}) = 0$. This can be seen by the following argument (similar to what we used to prove that there is no equally likely probability on countably infinite outcome spaces in 4.15): If $P(\{\omega\}) = c$ then $P(\{\omega'\}) = 0$ for any other $\omega' \in \Omega$. Hence we can find a sequence ω_i $(i \in \Omega)$ with $P(\{\omega_i\}) = c$ for all $i \in \mathbb{N}$. Using σ-additivity of P,

$$1 = P([0, 1]) \geq P(\bigcup_{i\in\mathbb{N}} \{\omega_i\}) = \sum_{i\in\mathbb{N}} P(\{\omega_i\}).$$

If $c > 0$, the last sum would be infinite, hence $c = 0$.

It is therefore sensible to think of probabilities of intervals rather than of probabilities of individual outcomes. For example, what is the probability that the random number is contained in the interval $[0.2, 0.3]$. The intuitive answer is $1/10$. Why? Because $[0, 1]$ can actually be partitioned in 10 such intervals and each should be equally likely, as they only differ from each other by a translation. (The intersections in the boundary points can be ignored in this discussion, as we have just argued that the probability for single points.) Using additivity yields that the probability has to be $1/10$ for each.

By the same token, we can do that for other partitions of $[0, 1]$. Fix $n \in \mathbb{N}$, consider $[(i - 1)/n, i/n)$ $(i = 1, \ldots, n)$ and show that $P([(i - 1)/n, i/n)) = 1/n$ for all $i = 1, \ldots, n$.

Generally, the probability for the random number to be in an interval $[a, b] \subset [0, 1]$ is equal to its length $b - a$. If a and b are rational, a partition can be defined accordingly and $[a, b]$ can be represented as an element of that partition or as unions of some of them. Otherwise, $[a, b]$ can be approximated by intervals with rational endpoints (technical details omitted here). Note that we do not need to distinguish between open, half open and closed intervals, because the probability of the endpoint is 0 anyway.

I summary, the idealised process of picking a random number between 0 and 1 can be modelled as follows:

$$\Omega = [0, 1], \mathcal{A} = \mathcal{B}([0, 1]), P([a, b]) = b - a \text{ for all } a, b \in [0, 1], a \leq b. \tag{6}$$

This measure could be fully described by a function that takes tracks the accumulated likelihood as x moves up the interval from 0 to 1 :

$$F(x) = x \text{ for } x \in [0, 1] \tag{7}$$

This idea will be further developed in the next section.

4.5.2 Cumulative distribution functions

Let P be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. It can be shown that the function

$$F(x) = P\big((-\infty, x]\big) \qquad (x \in \mathbb{R}) \tag{8}$$

has the following three properties:

(i) F is nondecreasing, i.e., $x > y \Longrightarrow F(x) \geq F(y)$.

(ii) $F(-\infty) := \lim_{x\downarrow\infty} F(x) = 0 \qquad$ and $\qquad F(+\infty) := \lim_{x\uparrow\infty} F(x) = 1$

(iii) F is continuous from the right.

F is called *cumulative distribution function (c.d.f.)* for P.

A proof will be omitted here, but here is a sketch: All the properties of F follow from the axioms of P. (i) follows from the monotony of P (Section 4.4(vii)), because increasing the interval $(-\infty, x]$ in (8) results in increasing its probability. To show the second statement in (ii) it is practical to first show that σ-additivity of P implies a condition called σ-continuity of P with respect to increasing sequences of sets, which says that for a sequence of sets A_i $(n \in \mathbb{N})$ with $A_1 \subset A_2 \subset A_3 \ldots$ it yields $P\big(\bigcup_{n\in\mathbb{N}} A_i\big) = \lim_{n\to\infty} P(A_n)$. Then apply this to $A_n = (-\infty, n]$ $(n \in \mathbb{N})$ and use Definition 4.11(P1). For the first statement in (ii) show a corresponding condition for decreasing sets and their intersection, apply it to $A_n = (-\infty, -n]$ $(n \in \mathbb{N})$ and use Section 4.4(i). To show (iii) note that for continuity from the right it is suffices to prove that for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F(x + x_n) = F(x) \quad \text{for all sequences } x_n \ (n \in \mathbb{N}) \quad \text{with } x_n > x \text{ and } \lim_{n\to\infty} x_n = x$$

This follows from σ-continuity of P applied to the sequence of sets $A_n = (-\infty, x + x_n]$. □

If F is discontinuous in $x_0 \in \mathbb{R}$ then the *mass* in x_0 equals the height of the jump of F in x_0, i.e.

$$P(\{x_0\}) = F(x_0) - \lim_{x\to x_0, x<x_0} F(x) > 0$$

There is a one-to-one correspondence between probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and functions with the properties (i), (ii) and (iii) above. The connection between them is given by (8) and

$$P\big((a, b]\big) = F(b) - F(a) \qquad \text{for all } -\infty \le a < b < \infty \tag{9}$$

The proof of the one-to-one correspondence is beyond the scope of this lecture.

Probability distributions can be classified in three categories based on their cumulative distribution functions:

(i) **Discrete.** F is a step function. It has atoms x_i $(i \in I)$, where I is a finite or countably infinite index set, with $P(\{x_i\}) > 0$ for $i \in I$, and $P(\{x\}) = 0$ for all $x \in \mathbb{R} \setminus \{x_i | i \in I\}$.

(ii) **Continuous.** F is differentiable. Then the derivate of F can be used to describe P. It is called density and will be studied in more detail in Probability B.

(iii) **Mixed discrete and continuous.** Such situations do come up in examples such as traffic queue waiting times that have a positive mass in 0 but are otherwise continuous.

Example 4.17. General form of a c.d.f. for the discrete case. Let S be a finite or countable infinite subset of $\mathbb{R}$ and

$$p_s \ge 0 \quad (x \in S) \qquad \text{with} \qquad \sum_{s\in S} p_s = 1$$

Then this defines a probability measure via $P(\{s\}) = p_s$. The corresponding c.d.f. is a step function with the representation

$$F(x) = \sum_{s\in S, s\le x} p_s \qquad (x \in \mathbb{R})$$

Example 4.18. Pick a point at random from an interval. What is the probability measure that describes picking a random number from an interval $[a, b]$. The above discussion on $[0, 1]$ can be generalised to an interval $[a, b]$. It just needs an additional normalisation to insure that $P([a, b]) = 1$.

$$P([x, y]) = \frac{y - x}{b - a} \qquad (a \le x < y \le b)$$

For the c.d.f. this yields

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} & \text{if } a \le x < b, \\ 1 & \text{if } x \ge b. \end{cases}$$

Picking points from half open or open intervals can be done in the same way. Including endpoints or not does not make a difference for the probabilities, because individual points have probability zero under continuous measures like this one.

Example 4.19. Spinner. (Continuation of Example 3.10.) A precise model for the spinner corresponds to picking a random number of the interval $[0, 2\pi)$, which is covered by Examples 4.18. A model for the approximation of the actually observed angle by only finitely many cases is covered by Example 4.17. More specifically, $S = \{1, \dots, N\}$, $p_s = 1/N$ and

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{i}{N} & \text{if } x \in [\frac{i-1}{N}2\pi, \frac{i}{N}2\pi) \text{ for an } i \in \{1, \dots, N\} \\ 1 & \text{if } x \ge 2\pi \end{cases}$$

In all examples so far F is either a step function or a continuous function. There are also mixed cases such as this one.

Example 4.20. Waiting in traffic. The following c.d.f. is an example for a the waiting time in a traffic light queue with no waiting time for those who come in while it is green.

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x + \frac{1}{2} & \text{if } x \in [0, 1) \\ 1 & \text{if } x \ge 1 \end{cases}$$

5 Conditioning

Here are a few reasons for learning about conditional probabilities:

(1) **Getting on top of trickster games.**
There are three coins one of which is copper on both sides, one is black on both sides, and one is black on one sides and copper on the other side. You are asked to pick a coin at random, put it on the table and look at what colour it faces. It shows black. What is the probability the other side is black?

Many people would reason as follows: One side is black means it can not be the coin with two sides copper. Of the remaining two coin each is equally likely. So the answer is 1/2. This is wrong. Can you tell why?

You picked the coin at random. The fact that you have already seen a black side − and this is crucial! − means one of the following two options happened: You picked the black/black coin or you picked the black/copper coin *and* got the black side up. So, of the initial six sides you already narrowed it down to three sides. These three are all equally likely. In two of the three cases the other side will also be black. Hence the answer is 2/3.

Draw a tree illustrating all the possible paths of this experiment to make it more obvious. Actually, we have seen a similar reasoning in the goat problem in the introduction.

(2) **Updating probabilities.**
Here are some examples how additional information could alter a probability assignment:

(i) A medical test indicates you have kidney cancer with a 75% chance to survive at least another 10 years. A few months later, metastasis are located in another organ. Given that information, you chances of surviving the 10 years will decrease.

(ii) You invested in stocks of the company *BuyOurStuff*. They have been on a steady upwards trend during the last few years and you assume it is highly likely that they will continue that way. You actually need to sell them off by the end of next month to finance the downpayment for a house and you are about 90% confident they will at last stay at their current value, which is exactly what you need to rise for the downpayment. Then a systematic tax fraud is discovered in the companies records resulting in them having to pay large amount of taxes by the end of the year. How does this affect your chances to get enough money for your downpayment from selling the stocks?

(iii) You start studying mathematics and find the material rather difficult compared to the material you did in school. You give it about 10% to get an Upper Second in your first year overall. Then you receive your Winter Analysis and Foundation exam results. While the average was rather low, you actually got scores as hight as 71% and 75%. You now believe that your chances to get an Upper Second in your first year overall are well above 50%.

(3) **Modelling different degrees of knowledge, circumstances or attitudes.**
A random number between 0 and 1 is drawn. What is the probability that the number is smaller than 0.25?

- Ann has no further information. She answers $p = 0.25$.
- Ben got a message saying the number is between 0 and 0.5. He answers $p = 0.5$.
- Cem got a message saying the number is larger than 0.7. He answers $p = 0$.
- Deb was told the number is between 0.1 and 0.2. She answers $p = 1$.

If people are being given different information, they may come up with different probability assignments. Furthermore, in cognitive psychology that there is no such thing as an objective way to absorbe information. The processing of information includes, among other mechanisms, filtering, modification and emphasising. This is mostly motivated by a tendency to fit new facts into existing believe systems while minimising cognitive dissonance, that is, the feeling of discomfort associated with holding two conflicting beliefs. (Persistant holding of prejudices is explained by such processes.) As a result of such processes, the evaluation of information

is a subjective. In particularly, any probability assignments based on a piece of information are subjective. Even if two people are given the exact same piece of information, they may come up with different probability assignments.

Mathematically, alternative probability assignments can be represented in terms of conditional probabilities.

5.1 Conditional probability.

Definition 5.1. Conditional probability.
Let $(\Omega, \mathcal{A}, P)$ *be a probability space and* $A, B \in \mathcal{A}$ *events. Assume* $P(B) > 0$. *The* conditional probability *of* A *with respect to* B *is given by*

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

Remark 5.2. Conditional probability is a probability measure.
We call $P(A \,|\, B)$ conditional *probability*, but *is* it actually a probability as defined in Section 4? The answer is yes, because for any fixed B, the set function

$$P(\,\cdot\,|B) : \mathcal{A} \longrightarrow [0,1] \qquad \text{with} \qquad A \mapsto P(A \,|\, B),$$

defines a probability measure. The proof amounts to showing the two properties in $(P2)$.
(i) is obvious: $P\big(\Omega \,|\, B\big) = P(\Omega \cap B)/P(B) = 1$.
(ii) follows from basic set operations: For $A_i \in \mathcal{A}$ $(i \in \mathbb{N})$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$,

$$\begin{aligned} P\Big(\bigcup_{i\in\mathbb{N}} A_i \,\Big|\, B\Big) &= \frac{P\Big(\big(\bigcup_{i\in\mathbb{N}} A_i\big) \cap B\Big)}{P(B)} = \frac{P\big(\bigcup_{i\in\mathbb{N}}(A_i \cap B)\big)}{P(B)} \\ &= \frac{\sum_{i\in\mathbb{N}} P(A_i \cap B)}{P(B)} = \sum_{i\in\mathbb{N}} P(A_i \,|\, B). \end{aligned}$$

Note the the σ-additivity of P could be used, because $(A_i \cap B) \cap (A_j \cap B) = \emptyset$ for all $i \neq j$,

ω: Why does the left-hand side in Defintion 5.1 look symmetric, but then the right-hand side is not?!

Ω: $P(A \,|\, B)$ is just a short form for the right-hand side and it is *not* symmetric. But I agree, that the symbolic expression $(A \,|\, B)$ is misleading in terms of its symmetric appearance. But what actually is the relationship between $P(A \,|\, B)$ and $P(B \,|\, A)$?

Some implications of the definition are both easy to proof and very frequently used:

Multiplication rule

$$\begin{aligned} P(A \cap B) &= P(A \,|\, B) \cdot P(B) \qquad \forall A, B \in \mathcal{A} \text{ with } P(B) > 0 \qquad (10) \\ &= P(B \,|\, A) \cdot P(A) \qquad \forall A, B \in \mathcal{A} \text{ with } P(A) > 0 \qquad (11) \end{aligned}$$

"Flip around" formula

$$P(A \,|\, B) = \frac{P(A)}{P(B)} \cdot P(B \,|\, A) \qquad \forall A, B \in \mathcal{A} \text{ with } P(A) > 0, P(B) > 0 \qquad (12)$$

Averaging conditional probabilities

$$P(A) = P(A \,|\, B) \cdot P(B) + P(A \,|\, B^c) \cdot P(B^c) \qquad \forall A, B \in \mathcal{A} \text{ with } 0 < P(B) < 1 \qquad (13)$$

Bayes' rule

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid B^c) \cdot (1 - P(B))} \tag{14}$$
$$\forall A, B \in \mathcal{A} \text{ with } P(A) > 0, 0 < P(B) < 1$$

The first two statements are immediate consequences of Definition 5.1. The last two statements are special cases of more general Theorems 5.4 and 5.5 shown in the next section. Here is an example for the use of the first statement.

Question 5.3. Pick a box and pick a ticket.
There are two boxes. One has three tickets with the numbers $1, 3$ *and* 5*, the other one has tickets with the numbers* 2 *and* 4*. First toss a fair coin. If heads come up choose the box with the odd numbers, if tails come up choose the box with the even numbers. Then you pick a ticket at random from the chosen box. What is the probability you get a* 3*?*

We will present two solutions, a naive approach and one based on the multiplication rule.

Answer 1: There are five different outcomes defined by the number of the ticket drawn, so $\Omega = \{1, 2, 3, 4, 5\}$. Since the two boxes are equally like, $P(\{1, 3, 5\}) = 1/2 = P(\{2, 4\})$. Since the tickets within each box are equally like, using the addition axiom, $P(\{1, 3, 5\}) = P(\{1\}) + P(\{3\}) + P(\{5\}) = 3 \cdot P(\{3\})$. Putting this all together yields $P(\{3\}) = 1/6$.

Answer 2: Denote the outcomes by two-character combinations: b_o and b_e for boxes with odd and even numbered tickets in it, followed by the number on the ticket. So the outcome space is $\Omega = \{b_o1, b_o3, b_o5, b_e2, b_e4\}$. We want to find the probability for b_o3. Define events:
$B_o :=$"box with odd numbered tickets was chosen"$= \{b_o1, b_o3, b_o5\}$,
$B_e :=$"box with even numbered tickets was chosen"$= \{b_e2, b_e4\}$,
$T_i :=$"ticket with number i was drawn" $(i = 1, 2, 3, 4, 5)$.
By assumption, $P(B_o) = P(B_e)$, so both must be $1/2$. The only way to draw a ticket with the number 3 is by choosing the box with odd numbers first. If drawing from that box, chances for each of the numbers are equal; in other words, $P(T_i \mid B_o) = 1/3$ for $i = 1, 3, 5$. Using (10), $P(\{b_o3\}) = P(T_3 \mid B_o) \cdot P(B_o) = 1/3 \cdot 1/2 = 1/6$.

Theorem 5.4. Total probability
Let $(\Omega, \mathcal{A}, P)$ *be a probability space and* $B_1, \ldots, B_n \in \mathcal{A}$ *a partition of* Ω*. Assume that* $P(B_i) > 0$ *for* $i = 1, \ldots, n$*. Then*

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i) \tag{15}$$

Proof. Since $A \cap B_1, \ldots, A \cap B_n$ are mutually exclusive, $P(A) = \sum_{i=1}^{n} P(A \cap B_i)$. Applying Definition 5.1 transforms the addends to $P(A \mid B_i)P(B_i)$. □

5.2 Bayes theorem and applications

Theorem 5.5. Bayes
Let $(\Omega, \mathcal{A}, P)$ *be a probability space and* $B_1, \ldots, B_n \in \mathcal{A}$ *a partition of* Ω *with* $P(B_k) > 0$

for $k = 1, \ldots, n$. Then

$$P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{P(A \mid B_1)P(B_1) + \ldots + P(A \mid B_n)P(B_n)} \tag{16}$$

Proof. Fix $k \in \{1, \ldots, n\}$. Formula (12) yields $P(B_k \mid A) = P(A \mid B_k)P(B_k)/P(A)$, which results in (16) after replacing $P(A)$ by the right-hand side in Theorem 5.4. □

Terminology

$P(B_i)$ are called *prior probabilities* – that's because it's *before* knowing about A
$P(A \mid B_i)$ are called *likelihoods* – probabilities of A given the different options B_i
$P(B_i \mid A)$ are called *posterior probabilities* – that's because it's *after* knowing about A

ω: How comes a theorem that is so easy to proof is so famous?

Ω: It's not the technique that made it famous. The novelty was the sheer idea to reverse the perspective. The "flip around" formula and Bayes theorem then just tell you how to crunch the numbers accordingly. In many applications it's about using the information about the probability for A given a few different B_i's to reverse engineer which of the B_i's has happened in the first place. And it also helped that it would sort out answers for very practical questions.

Example 5.6. Balls in boxes. There are three boxes:
Box 1 has 1 white ball and 2 black balls.
Box 2 has 2 white balls and 1 black ball.
Box 3 has 3 white balls and 0 black balls.
You pick a ball from one of the three boxes at random. Given the ball is white, what is the probability is was from Box 1? from Box 2? from Box 3?

Let A be the even that the ball is white and B_i the event Box i was chosen ($i = 1, 2, 3$). Using total probability theorem,

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i) = \frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{1}{3} + \frac{3}{3} \cdot \frac{1}{3} = \frac{2}{3}$$

Using Bayes' Theorem,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)} = P(A \mid B_i) \cdot \frac{1}{3} \cdot \frac{3}{2} = P(A \mid B_i)/2$$

In particularly,

$$P(B_1 \mid A) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}, \qquad P(B_2 \mid A) = \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}, \qquad P(B_3 \mid A) = 1 \cdot \frac{1}{2} = \frac{1}{2}$$

Note that the three conditional probabilities add up to 1 (as they should).

Example 5.7. Medical diagnostics. A lab test for the disease *VerySick* is conducted on blood samples and has the potential outcomes "positive" and "negative". We consider a certain population that has an incidence rate of 1% for the disease. We have experience with applying the test in this population: In the past, 95% of the people with the disease produce a positive test result, and 2% of the people without the disease produce a positive test result.
What is the probability that an individual chosen at random from that population will have the disease given his or her test was positive?

Let D be the event the individual has the disease and let $+$ and $-$ be events that the individual has a positive or negative test, respectively.

Using Bayes theorem,

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)(1 - P(D))} = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.02 \cdot 0.99} \approx 32\%$$

Discussion of this result. In words, it means that given a positive test, the chance to have the disease is about 32%. Only 1 out of 3 people in this situation will be given the correct message.

Why is the number so low? The reason is the small *prevalence* $P(D) = 1\%$ of the disease. As a result, the nominator is small even if $P(+ \mid D) = 95\%$ is quite high. The denominator, on the other hand, is inflated by the factor $1 - P(D) = 99\%$.

What are the practical implications for conducting such tests? 2 out of 3 people will be given a wrong result. If *VerySick* is not curable or life-threatening this is a terrifying message. It is therefore questionable to test the people in the first place. The disadvantages of testing may be outweighed by the existing treatment options at early diagnosis or by precaution in the case of infectious diseases. In many cases, further methods can be used to obtain a more reliable diagnosis, but they may be costly or invasive (and have potential side effects or risks). Also, the 2 out of 3 people who were given the wrong diagnosis still have to live with these news during a many weeks or months long waiting time until further results are obtained. Typical examples are screening tests for cancer.

A particular extreme example is antenatal care, because every pregnant woman undergoes many routine tests about medical problems or risks throughout the 9 months of pregnancy. With each of them carrying some amount error, she can more or less expect to get one or the other wrong result. This is an example for what is known as multiple testing problem in statistics. Ironically, with anxiety and stress having negative impacts on the pregnancy, wrong diagnostic results may manifest in real physiological problems counteracting the intention of antenatal care. A particular emphasis therefore need to be put on the way such results are communicated.

A small disease *prevalence* is typical for applying a screening test in an unspecific population. However, things look very differently when it is know that the individual to be tested comes from a population at higher risk for the disease. For example, with the same test accuracy probabilities as above, if $P(D) = 30\%$, then $P(D \mid +) \approx 95.3\%$. That means less than 5 out of 100 such people were given the wrong diagnosis.

ω**:** I want to tell you something. I really very much hate word problems.

Ω**:** Why?

ω**:** Either I do math or I do writing.

Ω**:** You mean one part of your brain is in charge of math and the other part of your brain is in charge of the "real world", and the two don't mix?

ω**:** Yes, that's it.

Ω**:** You are basically saying that applied mathematics can not be done.

5.3 General multiplication rule and independence

If the occurrence of the event B has no impact on the occurrence of the event A then $P(A \mid B) = P(A)$. Using the multiplication rule (10), this can be expressed as $P(A \cap B) = P(A) \cdot P(B)$. Or it could be expressed as $P(B \mid A) = P(B)$. Note that in formulations using conditional probabilities we need to assume that the probability of the event that comprises the condition is not zero.

Definition 5.8. Independence of two events.
Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A, B \in \mathcal{A}$. A and B are called independent*, if*

$$P(A \cap B) = P(A) \cdot P(B) \tag{17}$$

and dependent *otherwise.*

ω**:** I see, I see, it means they have nothing to do with each other! If I want a coin to come up heads and you want it to come up tails, than you and me are independent.

Ω**:** I'm not sure really what this means. Why don't you check?

ω**:** $P(\{h\}) \cdot P(\{t\}) = 1/2 \cdot 1/2 = 1/4$ and $P(\{h\} \cap \{t\}) = P(\emptyset) = 0$, oh, it doesn't work out...

Ω**:** They are disjoint, but they are *not* independent.

ω**:** Ah I get it, by being disjoint they actually have a lot to do with each other. If it's heads *implies* it's not tails!

Example 5.9. Independent and dependent events.

Roll a die. In Example 3.2 define the events $Odd = \{1, 3, 5\}$ and $G_k = \{1, \ldots, k\}$. Then Odd and G_k are independent if k is even, otherwise they are dependent. Just check (17), for example: $P(Odd \cap G_2) = P(\{1\}) = 1/6$ and $P(Odd) \cdot P(G_2) = 1/2 \cdot 1/3 = 1/6$, so they are independent, but $P(Odd \cap G_1) = P(\{1\}) = 1/6$ and $P(Odd) \cdot P(G_1) = 1/6 \cdot 1/6 = 1/36$.

Flip two coins. See Example 3.1. The assumption of equiprobable outcomes implies independency of the events $H_k =$"k*th* toss is heads" $(k = 1, 2)$, because
$P(H_1 \cap H_2) = P(\{hh\}) = 1/4 = 1/2 \cdot 1/2 = P(\{hh, ht\}) \cdot P(\{hh, th\}) = P(H_1) \cdot P(H_2)$.
This technical property reflects our idea that, if the coin flipping is conducted properly, *the coin has no memory.*

Remark 5.10. Real world applications. Independency is an assumption very often made for one or more of the following reasons:

(i) It is justified by theoretical considerations.

(ii) It is empirically justified (i.e. justified by data).

(iii) It simplifies calculations.

Reason (iii) is fine as long as the "real world application" only serves as an inspiration for a mathematical mind's desire to create beautiful problem and subsequently solve them. Before applying results of such creative processes to make statements about the "real world" additional study is required. In particularly, the magnitude and the direction of the error introduced by the simplification need to be addressed. A recent example of substantial errors caused by inappropriate independence assumptions occurred in models of the default risks of mortgages (even when they were secured by houses located in socio-economically highly homogeneous neighbourhoods). Such oversimplified risk models lead

to the subprime mortgage crisis in the US real-estate market that is believed to have triggered the 2007 global financial crises.

Reason (i) is often the case in science or engineering applications. Independence may be grounded in scientific knowledge or it may reflect technological constructions (e.g. electrical circuits). However, it does bear the risk that partial knowledge or an inappropriate weighing of existing knowledge leads to incorrect independence assumptions.

Reason (ii) is convincing and sometimes the only choice. However, the collection of data and the testing of independence in data bear their own challenges. Among other issues, model development based on empirical knowledge only can be misled by chance variation or bias in data. A simple but striking example is the empirical "proof" that the area of a rectangle is a linear function of the sum of the lengths of its edges[1].

Ideally, independence assumptions are based on both theoretical and empirical justifications. In some applications, independence assumptions are wrong, but can still play a role. For example in obtaining upper or lower bounds of an unknown true probability.

The multiplication rule (10) can be generalized to more than two events. This will be used to define a notion of independence for more than two events.

Theorem 5.11. General multiplication rule.
Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A_1, \ldots, A_n \in \mathcal{A}$ with $P(A_1 \cap \ldots \cap A_{n-1}) > 0$. Then

$$P(A_1 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2 \,|\, A_1) \cdot \ldots \cdot P(A_n \,|\, A_1 \cap \ldots \cap A_{n-1}). \tag{18}$$

Proof. Exercise. (*Hint: The result can be shown by induction based on multiple application of (10).*)

Definition 5.12. Pairwise independence.
Let $(\Omega, \mathcal{A}, P)$ be a probability space and $A_1, \ldots, A_n \in \mathcal{A}$ events. The events are called pairwise independent, *if*

$$P(A_j \cap A_k) = P(A_j) \cdot P(A_k) \qquad \text{for all } j, k \in \{1, \ldots, n\} \text{ with } j \neq k. \tag{19}$$

Remark 5.13. Pairwise and mutual independence are different.
Mutual independence implies pairwise independence. The reverse is not true. For example, take two independent coin tosses and consider the events $H_k =$"kth toss is heads" and $B =$"both tosses are the same". Then
$P(H_1 \cap H_2) = P(\{hh\}) = 1/4 = 1/2 \cdot 1/2 = P(\{hh, ht\}) \cdot P(\{hh, th\}) = P(H_1) \cdot P(H_2)$,
so H_1 and H_2 are independent. Also
$P(H_1 \cap B) = P(\{hh\}) = 1/4 = 1/2 \cdot 1/2 = P(\{hh, ht\}) \cdot P(\{hh, tt\}) = P(H_1) \cdot P(B)$,
so H_1 and B are independent and, similarly, H_2 and B are independent. However,
$P(H_1 \cap H_2 \cap H_3) = P(\{hh\}) = 1/4 \neq 1/2 \cdot 1/2 \cdot 1/2 = P(H_1) \cdot P(H_2) \cdot P(B)$, so they are not mutually dependent.

[1] David Freedman, Robert Pisani, Roger Purves, "Statistics", third edition, Norton.

6 Repeated experiments

ω: I've won! I always win... always, always, always!

Consider *independent* repetitions of the same kind of experiment. In advanced probability modules such as stochastic processes you will encounter more general models. In particularly, there will be models of processes that allow future repetitions to depend on the present (Markov processes) or on the present and a finite window of the past (higher order Markov processes).

We consider a series of independent Bernoulli trials (see Example 4.12). This is an abstract version of a model for repeated independent coin tosses.

Model 6.1. Bernoulli trials process.
A Bernoulli trials process *is a finite or infinite sequence of random experiments with the following characteristics:*

(B1) Each experiment has two possible outcomes, referred to as "success" and "failure".

(B2) The probability $p \in [0,1]$ for "success" is the same for each experiment; in particularly, it is not affected by the outcome of any of the other experiments.

The probability $1-p$ for "failure" is often denoted by q.
The standard outcome space for a Bernoulli trials process of length n is $\Omega = \{0,1\}^n$, where 1 refers to "success" and 0 refers to "failure". The standard outcome space for an infinite Bernoulli trials process is $\Omega = \{0,1\}^{\mathbb{N}}$, but (countable) submodels are also used.

The following toy examples from "mathematical finance" illustrate the main kinds of events considered in the study of repeated trials.

Example 6.2. Can you spare some change? Penny needs some cash. She asks everybody who passes by for some change. Each person gives her £1 with probability p and nothing with probability $1-p$. Assume that people react to her request independently of each other. Compute the probabilities for the following events:

- She asks 12 people. $A =$"4 of 12 people them give her £1."
- She keeps asking until her first success. $B =$"She ends up asking 5 people until her first success."
- She keeps asking until she has £4. $C =$"She ends up asking 57 people until she has got £4."

A is specifies the *number of successes* (see Section 6.1), B specifies the *waiting time until first success,* (see Section 6.2) and C specifies the *waiting time until a certain number of successes* (see Section 6.3).

6.1 Number of successes

Consider a Bernoulli trials process of fixed length n with success probability $p \in [0,1]$ and failure probability $q = 1-p$ as described in Model 6.1. In this section we calculate probabilities for events defined by the *number of successes,* i.e.

$$S_k = \text{"}k \text{ successes in } n \text{ trials"} \qquad (k = 0, 1, \ldots, n) \tag{20}$$

They can be represented using the 0-1-coding representing failure and success. For $\omega \in \Omega$ let $\omega(i)$ be the ith place in the tuple. Then,

$$S_k = \Big\{ \omega \,\Big|\, \sum_{i=1}^{n} \omega(i) = k \Big\} \qquad (k = 0, 1, \ldots, n). \tag{21}$$

Illustrate the calculation of the probability of S_k by looking at the case $n = 3$.
PROBLEM SOLVING TECHNIQUE: *Consider a special case or an example.*

Example 6.3. Three independents Bernoulli trials.
Let $n = 3$. Using independence we observe

ω	111	110	101	100	011	010	001	000
$P(\omega)$	p^3	p^2q	p^2q	pq^2	p^2q	pq^2	pq^2	q^3

In other words, there are four different probabilities involved:

- $P(\{111\}) = p^3$
- $P(\{000\}) = q^3$
- $P(\{110\}) = P(\{101\}) = P(\{011\}) = p^2\, q$
- $P(\{001\}) = P(\{010\}) = P(\{100\}) = p\, q^2$

Return to the general case of a Bernoulli trial process of length n. Because of independence, the probability of an outcome depends only on the numbers of successes and failures. Given the total number of trials is fixed, it is enough to record only the number of successes. This implies

$$P(\{\omega\}) = p^k\, q^{n-k} \qquad \text{for all } \omega \in S_k.$$

To calculate the probability of the event S_k it remains to count the number of outcomes in S_k. In how many ways can we choose exactly k successes in n trials? According to Section 3.3, using the formula for "without replacement, not ordered", this is $\binom{n}{k}$, so $|S_k| = \binom{n}{k}$ and

$$P(S_k) = \binom{n}{k} p^k\, q^{n-k} \qquad (k = 0, 1, \ldots, n). \tag{22}$$

This defines a probability measure on the outcome space of Model 6.1. It can also be thought of as a probability measure on the set containing all possible numbers of successes.

Definition 6.4. Binomial distribution.
The probability measure on the measurable space $(\{0, \ldots, n\}, \mathcal{P}(\{0, \ldots, n\}))$ *with weights*

$$p_k = \binom{n}{k} p^k\, q^{n-k} \qquad (k = 0, 1, \ldots, n) \tag{23}$$

is called the binomial distribution *with parameters* n *and* p.

Remark 6.5. Special case 1/2.
For $p = 1/2$ formula (22) simplifies to the following form:

$$p_k = \binom{n}{k} \frac{1}{2^n} \qquad (k = 0, 1, \ldots, n) \tag{24}$$

Answer to Example 6.2 A:

$$P(A) = \binom{12}{4} p^4(1-p)^8.$$

Example 6.6. Library books. You go the the library with a list of 7 books you need for your class. You would like to know the probability you can get a least 5 of them. Answers to questions like this one can be computed, but some assumptions have to be made. Here, we assume that the probability that a book is checked out is 10%, for any one of the books and independently of any of the other books. (This is a simplification and whether or not such assumptions are at least approximately true would need to be verified for a particular library.)

The model for a Bernoulli trials process of length 7 with success probability 1/10 (see Model 6.1) is suitable for this situation and we can compute the probability of the events S_k for k successes using (22)). Here, "success" is being used for the book being checked out.

$$\begin{aligned} P(\text{“at least 5 are not checked out”}) &= P(\text{“at most 2 are checked out”}) \\ &= P\Big(\bigcup_{k\in\{0,1,2\}} S_k\Big) = \sum_{k=0}^{2} P(S_k) = \sum_{k=0}^{2} \binom{7}{k} \Big(\frac{1}{10}\Big)^k \Big(\frac{9}{10}\Big)^{7-k} \\ &= \frac{9^5}{10^7} \cdot (9^2 + 7 \cdot 9 + 21) \approx 0.9743 \end{aligned}$$

So the probability that you can get at least 5 of the books from your list is about 97.43%.

Some more examples that could be tackled with a binomial distribution:

- On a given day, 20 babies are born in Coventry. What is the probability that 10 of them are girls?
- What is the probability that there are 3 left-handers in this classroom?
- What is the probability to get two faulty light bulbs in a pack of 30?
- You guess answers in a multiple choice exam at random. What is the probability you get 20 out of 50 questions right?

6.2 Waiting time until success

In this section, we look at events defined by the number of trials it needs until the first success. Events are defined based on the *waiting time for the first success:*

$$W_k = \text{“it takes } k \text{ trials until first success”} \text{ for } k = 1, 2, \ldots \tag{25}$$

We work with the model for infinitely many Bernoulli trials (see Model 6.1). Rather than using the default outcome space $\Omega = \{0,1\}^{\mathbb{N}}$, we only record the relevant outcomes, that is, $\Omega = \{1, 01, 001, 0001, 00001, \ldots\}$. Note that Ω is countable. This can be shown, for example, by using the one-to-one map between Ω and $\mathbb{N}_0$ that is defined by counting the number of zeros in any outcome $\omega \in \Omega$.

To calculate the probability for W_k just look at what it means to wait exactly k trials until first success. The first $k-1$ trials are failures and the kth trials is a success. Since

the trials are independent,

$$P(W_k) = q^{k-1}\, p \qquad (k = 1, 2, \ldots). \tag{26}$$

According to Example 4.14, formula (26) does define a probability measure on $(\Omega, \mathcal{P}(\Omega))$. The weights $p_k = P(W_k)$ are non-negative, and their sum equals 1, because W_k $(k \in \mathbb{N})$ form a partition of Ω. This can also be interpreted as a probability measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$.

Definition 6.7. Geometric distribution.
The probability measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ *with weights*

$$p_k = q^{k-1}\, p \ (k = 1, 2, \ldots)$$

is called the geometric distribution *with parameter* p.

To get a feeling for this distribution it is helpful to look at the consecutive odds ratios (and they are useful to quickly sketch a histogram of the distribution).

$$\frac{P(\{k+1\})}{P(\{k\})} = q \qquad (k = 1, 2, \ldots). \tag{27}$$

Answer to Example 6.2 B:

$$P(B) = (1-p)^4 p$$

Example 6.8. Floods. You buy a house in a flooding area with the understanding that the probability it floods in a given year is $p = 1\%$. What is the probability for the event A_k that the first flood will occur in the kth year after your purchase? Using a geometric distribution,

$$P(A_k) = (1-p)^{k-1} p = \frac{99^{k-1}}{100^k}$$

What is the probability for the event B_N that there are at most N floods in 100 years? To answer this question we do not use a geometric, but a binomial distribution:

$$P(B_k) = \sum_{k=0}^{N} \binom{100}{k} p^k (1-p)^{100-k}$$

ω: Why do they think flooding is a success?!

Ω: Well, yes, that is a good question.

Some more examples that could be tackled with a geometric distribution:

- You walk around town trying to find a restaurant that has your favourite dish. What is the probability it takes no more than 8 restaurants until you find one?
- Every year, the probability for an earthquake in a certain town is p. What is the probability that it takes at at least another 50 years until the next earthquake?

ω: You are trying to guess Rumpelstilzchen's name. What is the chance it takes you at most three attempts?

6.3 Waiting time for multiple successes

We are interested in the probabilities for waiting a certain number of trials until the rth success. Keep $r \in \mathbb{N}$ fixed. Define events based on the *waiting time until the rth success:*

$$W_k^{(r)} = \text{“it takes } k \text{ trials until } r\text{th success”} \text{ for } k = r, r+1, \ldots \tag{28}$$

As in the last section, consider the model for infinitely many Bernoulli trials (see Model 6.1). Now keep trying until you have obtained r successes. The outcome space Ω_r consists of all 0-1-patterns that have exactly r 1's in it, one of them at the end.
For example, for $r = 2 : \Omega_r = \{11, 011, 101, 0011, 0101, 1001, 00011, 00101, 01001, 10001, \ldots\}$.

Lemma 6.9. *Ω_r is countable (for any $r \in \mathbb{N}$).*

Proof. There is no canonical candidate for a one-to-one map f between Ω_r and the integers as in the last section, but it is still easy to do. One could arrange the outcomes in a triangle and then define a map from the integers to Ω by enumerating the outcomes row by row. For example, for $r = 2$ use

11
011 101
0011 0101 1001
00011 00101 01001 10001
...

and define $f(1) = 11, f(2) = 011, f(3) = 101, f(4) = 0011, \ldots$
It works similarly for larger r, but the explicit construction of such a map becomes more cumbersome as r gets bigger. Here is another way to show that Ω_r is countable: Every $\omega \in \Omega_r$ can be described by the its length and the location of the first $r-1$ successes. This suggests a surjection from the countable set $\mathbb{N}^{r-1} \times \mathbb{N}$ onto Ω_r, which implies that Ω_r is countable. □

To calculate the probability for $W_k^{(r)}$ just look at the meaning: The kth trials is a success, and there are $r-1$ successes and $k-1-(r-1)$ failures among the first $k-1$ trials. For any outcome of this kind the probability is $p^{r-1}\, q^{k-1-(r-1)}\, p$. So, similarly to the derivation of (22) we obtain

$$P(W_k^{(r)}) = \binom{k-1}{r-1} p^r\, q^{k-r} \qquad (k = r, r+1, \ldots). \tag{29}$$

This probability measure on $(\Omega_r, \mathcal{P}(\Omega_r))$ can also be represented as a probability measure on the possible waiting times itself.

Definition 6.10. Negative binomial distribution.
The probability measure on $(\{r, r+1, r+2, \ldots\}, \mathcal{P}(\{r, r+1, r+2, \ldots\}))$ with weights

$$p_k = \binom{k-1}{r-1} p^r\, q^{k-r} \quad (k = r, r+1, \ldots)$$

is called the negative binomial distribution *with parameters p and r.*

To get a feeling for the probabilities of these events it is helpful to look at the consecutive odds ratios (and hey are handy to quickly sketch a histogram). A simple calculation yields

$$\frac{P(W_{k+1}^{(r)})}{P(W_k^{(r)})} = \frac{k}{k-1} \qquad (k = r, r+1, \ldots). \tag{30}$$

Answer to Example 6.2 C:

$$P(C) = \binom{56}{3}(1-p)^{53}\,p^4$$

Example 6.11. Cables. A rope consists of many cables. During a high overload, a cable may break. We assume that at most one breaks at any given overload, with probability p, and that they behave independently of each other. The robe has to be replaced when more than 3 of them are broken. What is the probability of the event A_l that the robe can stand exactly l overloads (and has to be replaced right after woods)?

$$P(A_l) = \binom{l-1}{2}(1-p)^{l-3}p^3.$$

Some more examples that may be tackled with a negative binomial:

- People come in one-by-one and choose one of the two waiting rooms A and B at random. What is the chance it takes at least 20 people until there are 5 people in room A.
- You type text messages. What is the probability it takes 100 characters until you produce the 7th spelling error?

Ω: If we're late more than 3 times a term out parents need to come in to see the head teacher. What is the chance that our parents need to see that head teacher in a six week term?

6.4 Occurrences of independent events

6.4.1 Poisson approximation to the binomial (for rare events)

Example 6.12. Road accidents are rare in the sense that the probability p that a person selected at random from the population has a road accident on a given day is very small. For example, historical data for the UK says that p is about $0.000,000,21$. However, with about 55 Mio inhabitants, the average number of incidents all over the UK is about 11. This number is not negligible. For an insurer, this is relevant.

Actuaries need to look at the events of the type

$$H_k = \text{“exactly } k \text{ accidents today”}$$

Modelling this as a series of 55 Mio independent Bernoulli trials with parameter p, they get

$$P(H_k) = \binom{55,000,000}{k} \cdot 0.000,000,21^k \cdot 0.999,999,979^{55,000,000-k}$$

Actuaries, especially before computers, needed a way to approximate this.

Theorem 6.13. Poisson approximation
Let $\lambda > 0$ and $(p_n)_{n\in\mathbb{N}}$ a sequence in $[0,1]$ with $np_n \longrightarrow \lambda$. Then, for all $k \in \{0,1,2,\ldots\}$,

$$\lim_{n\to\infty} P\big(H_{k(n,p_n)}\big) = e^{-\lambda}\cdot\frac{\lambda^k}{k!}$$

where $H_k(n,p_n) =$ "k incidents in n trials with probability p_n".

Proof.

$$\begin{aligned} P(H_k) =& \frac{n!}{k!(n-k)!}\,p_n^k(1-p_n)^{n-k} \\ =& \frac{1}{k!}\left[\prod_{i=0}^{k-1}(n-i)\right]\frac{1}{n^k}(np_n)^k\left(1-\frac{np_n}{n}\right)^n(1-p_n)^{-k} \end{aligned}$$

The last factor converges to 1, because p_n is of the order $1/n$. Since $(n-i)\frac{1}{n} = (1-\frac{i}{n})$,

$$\left[\prod_{i=0}^{k-1}(n-i)\right]\frac{1}{n^k} = \left[\prod_{i=0}^{k-1}\left(1-\frac{i}{n}\right)\right] \xrightarrow{n\to\infty} 1$$

and for the remaining factors in brackets we have

$$(np_n)^k \xrightarrow{n\to\infty} \lambda^k \qquad \text{and} \qquad \left(1-\frac{np_n}{n}\right)^n \xrightarrow{n\to\infty} e^{-\lambda}$$

which yields

$$P(H_k) \xrightarrow{n\to\infty} \frac{1}{k!}\lambda^k e^{-\lambda}$$

6.4.2 Poisson model

The values $e^{-\mu}/\mu^k k!$ ($k \in \{0,1,2,\ldots\}$) add up to 1 (series for e^{μ}) and are non-negative, so they actually define the weight of a probability measure.

Definition 6.14. Poisson distribution.
Let $\mu > 0$. The probability measure on $(N_0, \mathcal{P}(\mathbb{N}_0))$ with weights

$$p_k = e^{-\mu}\cdot\mu^k/k!$$

is called the Poisson distribution *with intensity parameter μ.*

ω: We use this for counting fish. And the Sauterelle distribution for counting grasshoppers and the de Chevalier's distribution for counting horsemen.

Ω: Oh you're being silly... But why don't you read this old story about Chevalier de Méré... e.g. at `www.ualberta.ca/MATH/gauss/fcm/BscIdeas/probability/DeMere.htm`.

Here is a traditional application of the Poisson distribution.

Example 6.15. Earthquakes. Earthquakes are very unpredictable. A common simple model is to assume that the number of earthquakes in a given time unit follows a Poisson distribution. Say for a region in the West of the US historical data says that there is about λ earthquakes per year. The model for the number of earthquakes in a given year is a Poisson distribution with intensity parameter λ. Let N be the number of earthquakes next year.

(i) What is the probability for at least 3 earthquakes next year?
To avoid dealing with an infinite series go via the complement.
$P(N \geq 3) = 1 - \sum_{k=0}^{2} P(N = k) = 1 - (2^2/2! + 2 + 1) \cdot e^{-\lambda} = 1 - 5e^{-\lambda}$.

(ii) What is the probability for exactly k earthquakes during the next m years?
This corresponds to changing the unit to m years by using the new intensity parameter $m\lambda$, hence the probability is $e^{-m\lambda}(m\lambda)^k/k!$.

More situations where Poisson distributions potentially provide suitable models:

- number of people who survive age 100
- number of floods in the UK per year
- number of misprint on a book page
- number of particles given off by 1-gram of radioactive material per second
- number of occurrences of a certain pattern in a stretch of DNA

6.5 A first taste of asymptotics

For a Bernoulli trials process as described in Model 6.1 let $\omega(i)$ be the outcome of the i trial and

$$S_n(\omega) = \sum_{i=1}^{n} \omega(i)$$

the sum of the first n trials, that is the *relative frequency of successes* for this particular ω.

What happens to this expression in the long run? In other words, what is the asymptotic behaviour of $\frac{1}{n}S_n(\omega)$?

Thinking of a fair coin, we would expect it to converge to 1/2. More generally, in converges to the probability p for success in the individual trials. The asymptotic behaviour of sequences of random variables is one of the main topics in Part B, but we will state the first result here

Theorem 6.16. Golden Theorem (Weak Law of Large Numbers for Bernoulli sequences). *Use the set-up Bernoulli(p) sequences as set up above, for all $\varepsilon > 0$,*

$$P\Big(\Big|\frac{1}{n}S_n - p\Big| < \varepsilon\Big) \xrightarrow{n\to\infty} 1.$$

What does that mean in plain English? Does it mean that the relative frequencies of heads computed along any sequence of coin tosses converge to 1/2?

No! This would be much stronger and is actually wrong. For example, you may end up tossing heads forever, even with a fair coin. (Of course, very unlikely, but theoretically not impossible.)

What about excluding those weird cases? Then this turns out to be true, as you can learn in the Module Random Events in the second year. It is called the Strong Law of Large Numbers and states convergence of $\frac{1}{n}S_n(\omega)$ to p for all $\omega \in \Omega \setminus N$, where $N \in \mathcal{A}$ is a set of measure 0.

The Weak Law talks just about probabilities of events on a finite horizon:

$$A_n = \left\{ \left| \frac{1}{n} S_n - \frac{1}{2} \right| < \varepsilon \right\}$$

describes the event that the relative frequency of heads in the first n tosses is within an ε of $1/2$. So, in the case of fair coin tosses the theorem says:

> *The probability that the relative frequency of heads in the first n tosses is within an ε of 1/2 converges to 1 for n to infinity.*

7 Back to the beginning: What is probability?

To conclude Probability A, let us briefly discuss a few of the approaches to defining and understanding probability. For more about the history of probability see, for example, the books by Ian Hacking[1][2]. A starting point for philosophy of probability could be the new collection of essays edited by Anthony Eagle[3].

7.1 Classical probability

In many physical world examples symmetrical structures can be exploited to define probabilities. This naturally leads to models with equally likely outcomes, though can be pushed to cover a bit more than that, too. Over many centuries, theory development in probability has in part been stimulated by questions arising in gambling situations. The betting on a "double six" in 24 throws of a pair of dice and other problems posed by the French nobleman Antoine Gombaud, Chevalier de Mr, around 1654 led to an exchange of letters between Blaise Pascal and Pierre de Fermat in which the fundamental principles of probability theory were formulated for the first time. Maybe it was for these applications that von Kries in his work on the principles of probability in the 1880s introduced probability spaces as *Spielräume* (*Spiel* is German for game, *Raum* is German for space). An idealised casino serves as a laboratory for the science of random patterns: *Coins, dice* and *cards* to the probabilist what *drosophila, arabidopsis* and *c. elegance* are to the biologist.

Pierre-Simon Laplace and the publication of his book *Théorie Analytique des Probabilités* in1812 has taken probability theory away from solving disputes between gambling towards a wide range of scientific, societal and practical applications. The 19th century has seen probabilistic thinking and modelling entering the theory of errors, actuarial mathematics and statistical mechanics. One of the first famous examples for the latter is the theory of idealised gas by Maxwell and Boltzmann. Probability theory is the key to understanding the interplay between microscopic behaviour of particles systems and observed macroscopic quantities. In particularly, it has deepened the understanding of central concepts such as entropy. Quantum mechanics is heavily based on probabilistic thinking. Which concepts of probability are best suited to describe the phenomena in this fields is still a matter of debated in physics.

[1] Ian Hacking (1975), The emergence of probability, Cambridge: Cambridge University Press
[2] Ian Hacking (1990), The Taming of Chance, Cambridge: Cambridge University Press
[3] Anthony Eagle (2010), Philosophy of Probability, Contemporary Readings: Routledge

Genetics has been providing a stream of ever changing questions stimulating research in both in probability theory, as beautifully presented in the classical probability textbook by William Feller, and statistics, as impersonalised by Sir Ronald A. Fisher who is claimed by both professions. Today, probability theory and statistics provide a methodology and a language for quantitative bioinformatics and genomics.

7.2 Relative frequencies

It is an old intuitive idea that the probability of an event should reflect the relative frequency. Imagine you find a die, roll it 600 times and get 209 sixes. What do you think is the chance that will get a six in the next roll? You could say 209/600. Now you roll it 6,000 times and obtain 1,987 sixes. So it should be 1,987/6,000, which is similar but not the same. The more often, the more your estimates for the "real" probability will resemble each other. The underlying understanding of this approach is that the "real" probability is the asymptotic relative frequency, i.e., the limit of the relative frequencies. This tradition has been shaped, among others, by John Venn, Hans Reichenbach and Richard von Mises.

This understanding builds on the existence of such a limit. It further needs the limit to be deterministic, i.e. even though for any given finite time horizon the relative frequency is a random number, the limit of these frequencies will always be the same number. But it will not. Rolling only sixes is extremely improbable, but not impossible. Now we got stuck because of some weird cases.

Common sense can be reconciled with theory by focussing on what essentially happens instead of getting derailed by exceptions. There are so-called Laws of Large Numbers, which are all different ways of stating some sort of convergence of the relative frequencies to a limit representing the hypothetical "real probability" of an event. The exact type of results depends on the structure of the experiment, most of all the amount of dependence involved in the process. We have seen one of these asymptotic results, the Golden Theorem or Weak Law of Large Numbers. In order to prove such these kinds of theorems, however, we need to have already defined probability. Hence the approach of defining probability risks to result in a cyclic thought process. This has been addressed by an introduction of suitable axioms.

7.3 Propensity interpretation

Like the relative frequency approach, it is about relating probabilities to some "real word" phenomenon. In this tradition probability is thought to be a "physical propensity, or disposition, or tendency of a given type of physical situation to yield an outcome of a certain kind, or to yield a long run relative frequency of such an outcome"[1].

This approach has mainly been developed by Karl Popper who took it further to sees in a probability the "propensity of a repeatable experiment to produce outcomes of that type with limiting relative frequency p" collapsing it almost with the frequentist approach[2]. However, the angle is different. Propensities are assumed *causes* of the observed relative

[1] "Interpretations of Probability", Stanford Encyclopedia of Philosophy, 2002 (revised 2009). `http://plato.stanford.edu/entries/probability-interpret/`.

[2] dito.

frequencies rather than the frequencies themselves. A main difference to the frequentist approach that only refers to large ensembles is that the propensity interpretation allows to talk about single-case probabilities. Main criticism of this approach is that the definition of propensity still seems to be work in progress.

7.4 Axiomatic approach

The development of set theory and measure theory at the end of the 19th and the beginning of the 20th century, has accelerated the transformation from probability calculus into an axiomatically based mathematical theory about random patterns. The axiomatic approach has enabled 20th century mathematicians to build probability theory as a coherent mathematical discipline using the same formalism for all random experiments. The central piece, Kolmogorov's Axioms, were explained in Section 4.

In this mathematical framework, events are subsets of an outcome space and probability is a measure. The sets are elements of a σ-algebra, a system of sets closed under countably infinite set operations. The measure is normed and obeys laws reflecting how probabilities should behave when events are linked.

As a direct consequence of the axiomatisation, probability theory has unfolded enormously as an abstract mathematical discipline. As a side-effect, probability theory got increasingly detached from its roots in the manifold fields of applications (see Sections 7.1) with lead to Kolomogov voicing a complaint in 1963 that his axioms had been so successful on the purely mathematical side that many mathematicians had lost interest in understanding how probability theory can be applied[3]. Ironically, Andrei Kolmogorov and other mathematicians of his time had own always minded the connections between mathematical probabilities and its counterparts in the "real world".

7.5 Subjective probabilities

At the heart of this perspective is the individual's personal judgement about the likelihood of an event. For example: A risk averse personality tends to associate a higher likelihood with activities involving dangers than a risk taking personality. An insider will have a different idea about the future development of a certain stock than a regular trader. This perspective goes back at least as far as Bernoulli who defined probability as a degree of subjective certainty in the sense of an epistemic judgement.

Subjectivists of the 20th century, such as Frank P. Ramsey, Bruno de Finetti, Leonard Savage and Richard Jeffrey, added more factors to Bernoulli's knowledge and information based concept. Including factors like personality, values and taste, they defined probability as a measure of rational degree of belief. Using Dutch book arguments along with the condition of coherence, they derived the same rules for calculating with probabilities as expressed by the axioms (except they considered only additivity rather than σ-additivity). The most important technical tool of the subjective approach is conditional probability.

A difficult task is the elicitation of subjective probabilities. Expert opinions and surveys

[3]Vladimir Vovk and Glenn Shafer, "The Sources of Kolmogorov's *Grundbegriffe*, Statistical Science, Vol. 21 (2006), No. 1. `www.probabilityandfinance.com/articles/STS154.pdf`.

including questions that extract probability judgement based on betting metaphors are common methods. Opened the door to real human's minds, however, has also revealed that they do not always perceive and process probabilities as the mathematical definitions and rules would suggest. We have seen two examples in Section 2.7.3 illustrating a confusion between regularity and randomness and a misconception of conjunction of events.

The conception of probabilities is particularly difficult for very small probabilities. By their very nature it is hardly possible, in a finite life time, to experience the corresponding frequencies in a way that reflects their "real probability". Hence, learning from experience does not provide a good strategy for becoming knowledgeable about rare events. In other words, the epistemological basis for statements about rare events can not be purely empirical but has to rely on theoretical models, too.

ω**:** I know an example for rare events! There is a blind sea turtle coming to the surface once every one hundred years. And there is a yoke with a single hole floating on the surface of the ocean. What is the probability that it happens to come up with his neck right through that whole?

Ω**:** Is that what you learn in Foundation stage these days?! It's a teaching of the Buddha. He compares your sea turtle probability to the likelihood for be (re-)born as a human being.

ω**:** On another note, are you going to take probability B?

Ω**:** I have to sort out some very fundamental doubts about probability theory first. And this starts with coin flipping. I just read in a paper by Andrew Gelman and Deborah Nolan that claims that it is not actually possible to make a biased coin[1].

ω**:** I don't believe this.

Ω**:** And there is this work by Perci Diaconis, Susan Holmes and Richard Montgomery[1][2]. When they tried to make multiple flips of the same coin independent, they only got as close as to demonstrate that the other side came up 49% of the time.

ω**:** The magician Diaconis?

Ω**:** He did have a bit of a career change...

ω**:** Okay, I'm going to help you sorting out your doubt. I am going to make a biased coin all on my own. And I'll test it all on my own. Even if it takes a very long time.

Ω**:** But look, this is not a very good time for that. First you have to attend all your module lectures properly. Hey, I will give you a sticker for every homework assignment you turn in! And then you need to study for the exams. You need to be efficient, you see?

ω**:** No! I won't do it. There is no point doing Probability B and any exams for probability if things as simple as biased coins don't even exist. I'm not trying to get a degree in science fiction, but in mathematics, you see? These grown-ups have been telling us for years we have use their methods and answer their questions. There has never been a good time to think about our own questions, if you ask them. You are not one of them, Omega, are you? You *will* help me, will you? Yes, you will? Let's start *right now*!

[1] Andrew Gelman and Deborah Nolan, "Teachers Corner: You Can Load a Die, But You Cant Bias a Coin", American Statistician 56 (2002). `www.stat.columbia.edu/~gelman/research/published/diceRev2.pdf`.

[1] Perci Diaconis, Susan Holmes and Richard Montgomery, "Dynamical Bias in the Coin Toss", Society for Industrial and Applied Mathematics Vol. 49 (2007), No. 2

[2] Esther Landhuis, "Lifelong debunker takes on arbiter of neutral choices", Stanford Report, June 7, 2004. `http://news.stanford.edu/news/2004/june9/diaconis-69.html`.

Synopsis of the lecture part B

(*Online notes will be posted on the module website shortly after term.*)

Week I

1) Random variables and their distributions, discrete and continuous case
2) Combinations of random variables, cumulative distribution functions
3) Indicators, Gamma family of distributions, special case exponential family

Week II

4) Memoryless property of exponential, Normal family, distribution of the transformation of a discrete r.v.
5) Distribution of the transformation of a continuous r.v., expectation for discrete random variables, illustration, expectation of a transformation of a discrete random variable
6) Linearity of expectation of a discrete r.v., expectation of a geometric r.v., method of indicators, expectation and prediction, median, mode

Week III

7) Expectation for continuous random variables, heuristics for consistency with concept of expectation for discrete r.v., linearity, variance, properties of variance
8) Joint distributions, discrete and continuous case, independent r.v., correlated r.v., uniform on a disc
9) More examples for joint distributions, expectation of a sum of finitely many r.v., covariance, independent r.v.

Week IV

10) Variance of a sum of finitely many r.v., phrase question about sums/averages of r.v. and their asymptotics,
11) Tail probabilities: Markov's and Chebychev's inequality, Weak Law of Large Numbers (WLLN)
12) Proof of WLLN, generating functions

Week V

13) Presentation by MMORSE project students about survey, generating functions (properties and examples)
14) Central Limit Theorem
15) Applications of the Central Limit Theorem