JACOBS UNIVERSITY

Herbert Jaeger

Principles of Statistical Modeling

Lecture Notes

Version 0.11, Apr 4, 2019

Master Program Data Engineering

Contents

Course Concept and Overview

“Data Engineering”, “Data Analysis”, “Data Science”, “Big Data”, such catchwords have rocketed to public prominence in the last few years. These days we witness an amazing explosion of data processing technology and data analysis methods. It is fuelled by a number of factors:

- an explosion of available data volumes in the first place, enabled by technological developments in internet services, mass data storage devices and database technology,
- a nonlinearly accelerating progress in mathematical data analysis methods which has led to unforeseeable and almost surreal innovations, like artificial visual hallucinations,
- disinhibitive shifts in social behavior that make people spill out personal data,
- fast-growing societal complexity and dynamics which makes national security agencies hunger for information,
- exploding complexity of manufacturing and commercial processes, itself enabled by novel data technologies, unleashing economical forces for ever-increasing data processing rates and data processing complexities,
- not to forget: self-reinforcing hype-cycle mechanics.

OK., understood, *data* there is plenty. But nobody wants just data — everybody wants *information*[1] instead. It’s like ore and gold: the raw material is rough and ragged and bulky, most of it useless waste, containing the precious substance but in traces, requiring energy and ingenuity to be extracted. Similarly with data and information. Raw data is

- voluminous,
- unstructured,
- replete with irrelevant material,
- disorganized,
- faulty and fragmentary,
- partial and biased,

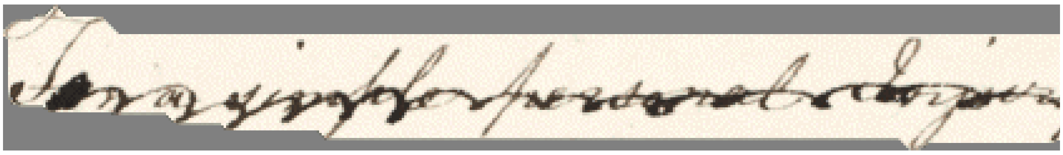
[1] I am using the term “information” in the commonsense meaning of “valuable-to-know stuff”, not the mathematical sense of information theory. I could also have used words like “insight”, “knowledge” or “intelligence”.

- unrepeatable: when you collect it again from the same source, you get something else.

Information, in contrast, you wish to be

- concentrated,
- pure,
- rendered in a homogeneous, transparent format,
- organized, accessible, searchable,
- reliable and repeatable,
- malleable, shapeable to fit your needs.

Dwelling on the mining metaphor a little longer — what in metallurgy is the mother rock, the cinders and the slag [I looked that up in a dictionary], for information processing corresponds to noise, randomness, irregularity, redundancies. That's what you need to get rid of, bringing to the fore the purified information content. In this course you will learn to appreciate how frightfully large is the relative amount of noise and irregularity in raw data, compared to the tiny fraction of clean information that can be extracted. Figure 1 shows an instructive example from the domain of "pattern recognition", a subarea of machine learning.



```
for a  simple personal Injury
For on simple personal injury
```

Figure 1: Extracting information from data. Raw data: photographic image of a historical handwriting (actually, the primary raw data was a photo of the entire page; part of the information extraction was already done by isolating the text line). Extracted information: the second printed text line. "Correct" information: the first printed text line. Extraction was done with a state-of-the-art historical document analysis tool developed by Planet GmbH, a data engineering company with whom I closely collaborate. Picture taken from Sanchez et al. [2015].

Metallurgy is based on the scientific insights of mineralogy, chemistry and physics. Extracting information from data is likewise an engineering task. It is a practical, technical enterprise that is based on scientific insight. The scientific foundation of data engineering is *probability theory.* Just like metallurgists use the theoretical-scientific insights of mineralogy, chemistry and physics to build practically useful machines like furnaces or chemical processing plants, data engineers use the mathematical theory of probability to design computational procedures that actually *do* the information extraction. Historically there have been two traditions of such practical data-to-information engineering, the first (and earlier) being *statistics* and the second (more recent one) being *machine learning.* Both fields share the same mathematical roots in probability theory, but they pursue interestingly different goals and have developed different tools and techniques. Today's "Data Science", "Data Analytics", etc., equally draws from both traditions.

In fact one could include a third engineering tradition in this picture, *signal processing*. However, signal processing is a bit outside of what is meant when people talk of "Data Science" etc. It is more connected to application domains like communication technology, electronics or robotics, therefore we will not consider it here. Students with an interest in signal processing may consider taking courses in the Electrical and Computer Engineering and in the Intelligent Mobile Systems programs.

This course is the basic theory course in the Data Engineering program. It is structured in four parts:

Part 1: Face to Face with Probability: Clear Concepts, Clean Notation. Here we introduce you to the "way of thinking" of probability. Randomness is an amazingly elusive thing, and it has taken philosophers and mathematicians several hundred years to agree on a consistent way of thinking about randomness. What is now the standard mathematical framing of "probability" has been consolidated only as late as in the 1930'ies – much later than most other subfields of mathematics. In Part 1 we will explain the intuitions of this standard conception of "probability". The mathematical formalism introduced in this part is minimal – just a few lines of axioms will be distilled at the end – but the underlying intuitions and conceptualizations are far from easy to assimilate. The aim is to give you a firm grasp on the concepts of a *probability spaces*, *random variables*, *distributions* and *samples*. Once these foundations are set, there will be a fast-forward refresher of all the elementary and useful derived concepts and formulas that you probably have learnt before in some introductory math course, like *conditional probability*, *marginal probability*, *density functions*, the *normal distribution* (and some more), *moments of a distribution*, *factorization of distributions*, *correlation* or *statistical independence*. What is maybe new: you will understand the fundamental intuitions underneath these concepts and tools — in typical introductory courses these concepts and tools are presented "mechanically" only.

Part 2: Introduction to statistics. Here we concentrate on *inferential statistics*: given an amount of observed data (a *sample*), what can be inferred from these about the underlying "true" properties of the mechanisms that generated those data? And with what degree of confidence can one make statements about "true" facts on the basis of data ridden by randomness? Indeed, what exactly does "confidence" mean in the first place? The methods of inferential statistics are thus the basis for *decision-making* based on data.

Part 3: Introduction to machine learning. Machine learning is a heterogeneous field with roots as diverse as artificial intelligence, cognitive science, neuroscience, signal processing, and, in fact, statistics. In this course we focus on machine learning as a set of techniques to find relevant structure in messy data, — to detect the regularities that are hidden in the random data. We will introduce a number of ways of how such "regularities" can be defined and formalized in the first place — in other words, we'll introduce a number of ways to distil *models* from data.

Part I

Face to Face with Probability: Clear Concepts, Clean Notation

Chapter 1

How statistical data come into being: the big picture in plain English

1.1 Five basic components of understanding "probability"

Before we start formalizing "randomness" and "probability" in a serious mathematical way, let us populate the stage by a number of simple and not-so-simple examples of random systems, using the word "probability" naively. All examples describe scenarios where statistical data are systematically collected. Each scenario will be described in terms of five components:

1. a circumscribed portion of reality in which we collect data — we will call this the *reality segment of interest* (RSOI);

2. within that reality segment of interest, a set of objects or events or moments in space-time from which observation data *could* be collected — let us call this the set of *observation opportunities* (OO);

3. an apparatus or a procedure which enables one to actually get data in the RSOI at every observation opportunity, like a measurement device in a lab or a questionnaire sent out by a public opinion polling team — we will call this the *observation procedure* (OP);

4. from among the observation opportunities, a finite subset of those opportunities where the observation procedure is actually set to action and data are recorded — the *observation acts* (OA);

5. and finally, the set of all possible results that the observation tool could deliver — we'll refer to this as the *data value space* (DVS). Since it is in most cases not obvious which observation values, exactly, are possible vs. impossible, the data value space can be generously specified as a possibly larger-than-necessary set of data values – the only important condition is that it must surely contain all possible values.

This may sound rather complex or hair-splitting. But, "probability" is one of the most elusive and controversial concepts of mathematics, philosophy and the sciences. It has taken mathematicians very long to convene on a reliable, useful and widely shared definition of "probability", completing this job only in the late 1930's — long after almost all other fundamental concepts of mathematics had safely been settled.

This terminology (reality segment of interest RSOI, observation opportunity OO, observation procedure OP, observation act OA, data value space DVS) is just my private plain-English way to name the main players in the probability game. You will find this terminology only in these lecture notes, not in other textbooks. Standard scientific terminology will be introduced as we go along.

1.2 Examples

We will now inspect a number of examples of our RSOI-OO-OP-OA-DVS setup.

1.2.1 Tossing a coin

This is the absolutely most classical example and contained in each and every textbook on probability or statistics. A person has a coin and can throw it as often as s/he wishes. From the outcomes of many throws — "head" (H) or "tail" (T), the person might want to determine whether the coin is fair or biased. Here is a description of our five components in this scenario:

Coin Tossing	
RSOI	A concrete person (for instance, *you*), with a concrete coin (say, the one Euro-cent coin that you always keep in your pocket for luck); you are sitting at your desk at home.
OO	Any moment in your life while you are sitting at that desk and have that coin at your disposal – in other words, any moment when you *might* toss the coin.
OP	The procedure here comprises the processes of you tossing the coin, recognizing the face that shows up, and writing "H" or "T" on a sheet of paper.
OA	Each action event when you actually do a toss under the abovementioned conditions, and make a note of the "H" or "T" outcome.
DVS	The set of symbols $S = \{H, T\}$.

Sneak preview (to be explained in much detail later): the sequence of "H" or "T" outcomes that eventually you have noted down on paper will be formalized as and called a *sample*. The *probability* of the coin to come up with "H" will be understood as the fraction of the "H" outcomes among all ("H" or "T") outcomes, assuming that *all* observation opportunities would have been realized. This probability is the "true" ratio of the coin (thrown by you in that room of yours) turning up a head. This probability is a real, physical property of that coin (when it is tossed by you in your room). The fraction of "H"-s in the actually recorded data on your sheet of paper is an *estimate* of that probability. The two probabilities of the "H" and "T" outcomes constitute the *distribution* of the H-T-value outcomes (in that specific RSOI).

Notice that the probability of "H" turning up is a physical-real property of the entire physical setup (you, your room, that coin). If any detail of the RSOI would be changed – for instance, you would be throwing the coin in the garden; or you would be doing the throwing in your room and at your desk but with another coin; or another person would throw the same coin in your room – then the physical situation would change and with it the probability of the coin showing “H” might change as well. Probabilities and everything else connected to it are defined always and only with respect to a given, specific reality segment of interest.

1.2.2 Throwing a die

A person has a die and can throw it as often as s/he wishes. From recording the numbers of outcomes “1”, “2”, ..., “6”. Here is the summary table:

Die Throwing	
RSOI	A concrete person (for instance, *you*), with a concrete specific die; you are sitting at your desk at home.
OO	Any moment in your life while you are sitting at that desk and have that die at your disposal – in other words, any moment when you *might* throw the die.
OP	Analogous to the coin tossing example.
OA	Whenever you actually do a throw under the abovementioned conditions, and make a note of the facing-up number.
DVS	The set of integers $S = \{1, 2, 3, 4, 5, 6\}$

1.2.3 Body weight

What can be said about the distribution of body weights of humans? To make this a well-defined question, a specific RSOI must be fixed. In the table below I invent one.

Body Weight	
RSOI	The set of all citizens of the EU, in January 2017.
OO	Each citizen makes for an observation opportunity. Compared to the coin and die throwing examples, the OOs here are objects, not moments in space-time. Another difference is that here the set of OOs is finite; in the coin and die throwing examples it was infinite (because your sitting time at the desk consists of a continuum of moments).
OP	The process here would be that a human operator places a citizen on a scale, reads off the indicated weight, and makes a note of the reading.
OA	The observation acts would comprise a suite of weight measurements that have actually taken place in January 2017, for instance funded and organized by a survey of the European Commission carried out in selected hospitals.
DVS	Possible human body weights are nonnegative real numbers. Also it can safely be assumed that no human weighs more than 500 kg. Thus one admissible choice for the data value space would be the real interval [0, 500]. But also the entire real line $\mathbb{R}$ would qualify.

This example exhibits a common problem in empirical statistics. If the goal of the European Commission is to get an overview of the distribution of body weights of European citizens, then it may be quite misleading to infer this distribution from the observation acts (the sample) carried out in a number of hospitals. This is because citizens that can be found in hospitals are typically ill, and ill people often suffer from weight loss; furthermore, typically patients in a hospital are older on average than the population average. Thus, one does not obtain an *unbiased* sample. Extrapolating from the actually carried-out observations to the global weight distribution among Europeans is therefore problematic. Since problems of this kind are ubiquitous, textbooks of statistics contain chapters that deal with methods to identify such "bias" in a sample; methods to compensate againts biased samples with computational techniques; methods to minimize the lopsidedness of samples by a careful *design* of a measurement campaign. At any rate, it is important in a scientific report which is based on statistical data to describe in some detail how the sample was actually collected – in our terminology, how it was decided when and where to carry out an observation act.

1.2.4 Determining the speed of light: single lab

Physicists believe that the speed of light in vacuum (customarily denoted by **c**) is a universal constant of nature, and they are interested in determining it with the greatest possible precision. One way to increase precision of an estimate of a constant of nature is to repeat a measurement procedure and use the average outcome instead of the value received from a single measurement. The following table gives an account of how repeated measurements in some physics lab can be modeled.

Speed-of-light I: Single Lab	
RSOI	A particular physics laboratory (say, Lab 1).
OO	Similar to the coin/dice throwing examples, the observation opportunities are moments in the lab lifetime.
OP	The observation procedure is based on a laser and mirror system and high-precision time measurement apparatuses, which are to be used in a way that is described in a thick handbook.
OA	The measurement apparatus is "fired" ten times per second for one hour, giving altogether 36000 observation acts. Each of these leads to the recording of a measured speed of light value.
DVS	Knowing that the true speed of light is about 300,000,000 m/sec, a safe data value space would be, for instance, the real interval $[200,000,000 \quad 400,000,000]$.

1.2.5 Determining the speed of light: multiple labs

In the sciences, experiments have to *reproducible* across different labs: other labs must be able to re-install the experimental set-up of Lab 1 and repeat the measurements. We face here a twofold repetition of measurements: first, each single lab will repeat the measurement in order to achieve a more precise estimate of the true speed of light; furthermore, such entire measurement series are repeated across different labs. Such twofold measurement repetitions are characteristic for the use of statistics in the sciences. When you read a research paper from the empirical sciences, you will find in it claims that are based on repeated measurements (a single measurement never gives a convincing support for a scientific claim in the empirical sciences). But since the empirical experiments leading to the claim should be reproducible in other labs, a statistical model of the full scientific situation must cover both the repeated measurements in a single lab, plus the reproductions of such repeated measurements in different labs. The way how this is conceptually captured in statistical modeling is really not easy to digest, as you will see from the involved and maybe a little non-intuitive RSOI-OO-OP-OA-DVS descriptions in the table below.

Speed-of-light II: Multiple Labs	
RSOI	Taking into account that natural science experiments should be reproducible by any lab, — even labs that will only be established in the future —, we take as the reality segment of interest the collection of *all* labs that host appropriate laser-timing apparatuses, considered in the year 2017. This collection of labs contains not only physically existing labs but also "virtual" labs that *could* be actually built or that could install the right kind of laser-timing device.
OO	The observation opportunities are now the labs `Lab-i`.
OP	The observation procedure in this multi-lab scenario comprises the execution of an entire series of measurements in a specific lab. The outcome of runningd this procedure is thus a sequence of real numbers (as many as the times that the apparatus has been fired in that lab).
OA	An observation act occurs when in an existing lab the observation procedure is actually carried out and a sequence of measurement values is recorded.
DVS	A technical hurdle with defining an appropriate data value space lies in the circumstance that the measurement value sequences returned from running the OP will be of different length. We need to formalize a collection of sequences of real numbers of variable length. Assuming that each single measurement value lies between 200,000,000 and 400,000,000, in a clean notation we could define the DVS as $S = \{(x_1, \ldots, x_{T_{\text{stop}}}) \mid T_{\text{stop}} \in \mathbb{N}, T_{\text{stop}} > 0, 200000000 \leq x_i \leq 400000000\}$.

1.2.6 Credit risk modeling

Jacobs Bank (JB) is active in handing out student loans. For their business it is important to know in advance whether a student will be able to pay back her loan. In order to predict the risk of credit failure, JB tries to learn as much as they can about their clients. In a questionnaire that every student customer has to fill in, they ask for items like age, gender, profession of parents, amount of monthly income, sources of monthly income, etc., etc. Furthermore, if the student also has her current account with JB, the daily transactions are recorded. Finally, and importantly, if the student fails to pay back the loan, a capital "F" is marked in a special field of the record; if the loan is payed back, this field is filled with "B". The bank attempts to predict and quantify loan failure risk on the basis of these data.

Credit Risk	
RSOI	The reality segment of interest here is the collection of all Jacobs student who have a loan contract with JB — both the students that actually did so in the past and for whom records exist, and “potential” students that will enroll in the future. Wishing our university a long life, the vast majority of these latter students has not been born at the time of writing (2018).
OO	The observation opportunities here are again all past, present and future students of Jacobs University with a JB loan contract. Here we have the (not uncommon) case that the RSOI and the OO coincide — the RSOI is construed as a set of OOs.
OP	The recording procedure is the multi-step action of JB to email the questionnaire to a student, of the student to fill it in and send it back, then of JB again to copy the questionnaire entries to their database.
OA	The observation acts comprise all those (past and present but not future) cases where in fact some student filled in the questionnaire and the questionnaire data were entered into JB’s database.
DVS	The observed values here are student records in JB’s database. Each such record is a list of fields filled with words (e.g. the name of the student), with yes/no choices, with numbers (e.g. age, parents’ income), with entire text documents in pdf (the company might require copies of legal documents), or just blanks (“missing values”). Such a record is a complex datastructure. The records of different students may have different fields and lengths. It is helpful to think of a student’s record as a row in spreadsheet. S is the set of all (hypothetically) possible such records. In mathematical formalism, such a record would be the cross product of the possible value ranges for each field in an Excel row. To make this more concrete, assume such a record starts with fields `FamilyName`, `GivenName`, `YearOfBirth`, and ends with a field `PayBackCompleted`. Then the mathematical format of a record would be the set $\{\text{A-Z,a-z}\}^* \times \{\text{A-Z,a-z}\}^* \times \mathbb{N} \times \ldots \times \{\text{B, F}\}$.

1.2.7 Text translation

A “text” is a finite string of letters from an alphabet, and it is read from left to right – there is a natural temporal ordering in a text. Statistical models of texts are used in many modern data analysis applications, for example in automated address recognition systems used by parcel delivery companies, or in spam filters, or reading aids for visually impaired persons, or in machine translation systems — which is the application that we will consider here, say from English to German. Today’s best machine translation systems mostly operate on the basis of individual sentences (i.e., they are blind to information carry-over from one sentence to the next in multi-sentence texts).

Sentence Translation	
RSOI	Assume the goal of the statistical study is to develop a translation engine from English to German, to be deployed in a smartphone app. A suitable reality segment then would be, for instance, all English-speaking countries in the year 2018.
OO	An observation opportunity in that RSOI is any situation where some person is awake and has a smartphone ready.
OP	The observation procedure is that a smartphone service provider records an English text that was typed by a user, makes the text available to a professional translator who translates that text to German, and saves the translation alongside with the English source text.
OA	The cases where the OP is actually carried out.
DVS	The set of pairs (s, s') where s is a sentence letter string in English and s' a correct translation. Alternatively one could also admit all pairs (s, s') where s and s' are *any* letter strings; correct English-German sentence pairs would form a small subset.

1.2.8 Speech recognition

This example is a very important one in today's machine learning world, where we are witnessing uprecedented progress in automated speech recognition systems. Such systems (like Apple's Siri) transform a microphone signal into a written text output (or into other useful formats or actions). To make such a system function, it must be *trained* by the factory that produces it on examples of speech recordings. To make such a system function *well*, it must indeed be trained on a truly gigantic set of speech examples — the bigger and the more diverse the better. The recent advances in speech recognition technology are owed to a large part to the progress in computing hardware and storage technology as enabling factor for processing very large amounts of training data.

Speech Recognition	
RSOI	All English speakers of the world (including dialects and non-native speakers), in all kinds of everyday life situations, in the decade 2011-2020.
OO	All wake moments in the lifes of English-speaking adult humans in the years 2011-2020.
OP	An engineer picks a person from the street, pays her a small honorary, let her read a written text into a microphone, writes the recorded sound signal to a file.
OA	All events where an Apple researcher actually does that.
DVS	The data value space S is the set of all pairs (x, y) where x is a digital microphone signal and y is a text.

1.2.9 Evolutionary trees

This example is lies at the heart of an important line of research in genetics and evolutionary biology, where researchers try to find out how evolution proceeded in the past, on the basis of genetic evidence collected at the present time. A particular instance of this scientific quest frequently makes it into the media: what can we learn from the genetic fingerprints of today's humans and the great apes (chimps, gorillas, orangs and some more) about the evolutionary history of the human species? is there a common ancestor? are we closer to chimps or to orangs? etc. I follow the leads of Mau et al. [1999], a highly cited reference that introduced a particular statistical method of analysis to reconstruct evolutionary trees of descendence from current DNA records.

Recall that a biological species is defined through its DNA sequence (quite a simplification — two individuals of the same species won't have exactly identical DNA, but we ignore that here), and a DNA sequence can be seen as a very long word written in the letters A, C, G, T. Let us consider humans, chimps, gorillas, orangs only; they are characterized by four DNA sequences that we denote by w_h, w_c, w_g, w_o. An *evolutionary tree* is a binary tree whose nodes are DNA sequences and whose links are labelled with numbers denoting evolutionary timespans. Figure 1.1 shows two hypothetical trees for the evolutionary history of the great apes.

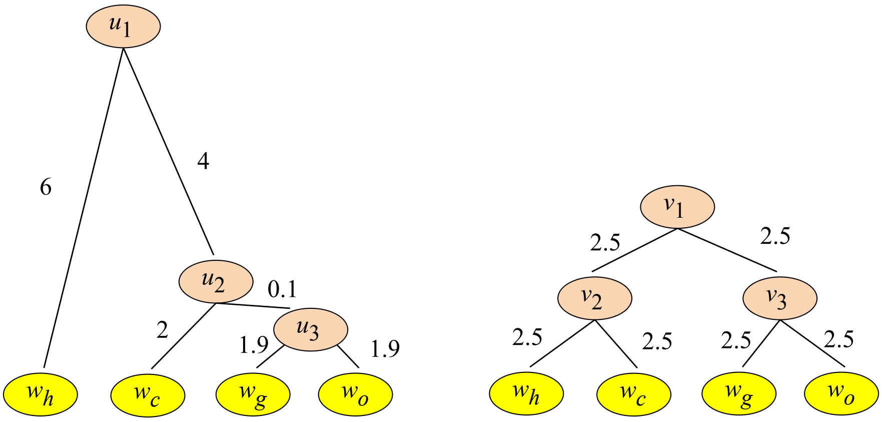


Figure 1.1: Two hypothetical evolutionary trees for the great apes. Yellow: currently existing species. Light orange: hypothetical ancestor species, now extinct. Each species is characterized by a DNA word u_i, v_i or w_x. Numbers at links indicate millions of years. In the left tree, humans split from the other apes a long time ago. In the right tree, humans are as closely related to chimps as gorillas to orangs. I made up these trees, no biological plausibility implied.

Evolutionary biologists attempt to reconstruct the "most probable" evolutionary tree on the basis of (i) the known DNA sequences of today (yellow in the figure) and (ii) assumptions about the laws of evolutionary change of DNA sequences (by not entirely random mutations).

Doing statistical analyses for historical or evolutionary processes faces a difficulty that seems insurmountable: statistical statements are based on the repetition of observations, but there is only ONE real world history. One cannot repeat the history of the earth or of human society in order to supply statisticians with nice data. This problem is solved by an audacious construction: one assumes that our universe with our earth in it is only one of innumerable other *possible worlds* in which the same laws of nature rule, but history/evolution proceeds along different random paths. Since one cannot obtain real physical observations from other worlds than ours, one must take resort to computer simulations of the other worlds to obtain "data".

Evolutionary Trees	
RSOI	The reality segment of interest is the (hypothetical) collection of all possible earths which have the same laws of nature as ours, and which in their year 2000 have evolved chimps, gorillas, orangs and humans with the same DNA sequences w_h, w_c, w_g, w_o as “our” apes.
OO	Each such hypothetical world is an observation opportunity.
OP	The procedure to “observe” an evolutionary tree is to run a computer simulation which implements the known (random) laws of evolutionary change. The outcome of such a simulation run is a labelled tree as in Figure 1.1.
OA	Each actually executed simulation run is an observation act.
DVS	S is the set of all possible binary trees with four leaves, where the four leaves are annotated with w_h, w_c, w_g, w_o, the internal nodes are annotated with any words over the alphabet $\{A, C, G, T\}$ and where the links are labelled with integers, such that the sum of integers on any path from the root to any leaf are equal.

Evolution theorists proceed as follows in order to determine the most probable evolutionary tree for the great apes. The first collect a large number of such trees, by running their evolution simulation engine many times. Then they check whether one type of tree (characterized by its branching topology, ignoring the duration labels) occurs in a significantly larger number than other types of trees. If that is the case, they publish an article claiming that humans are more closely related to chimps than to gorillas, etc., reporting the structure of the dominant tree type.

1.2.10 Weather forecasting

This possible-worlds idea is also used in weather forecasting (and financial forecasting and others). The physics of atmospheric dynamics are largely known and furthermore they are deterministic. Thus, forecasting weather amounts to run a simulation of these deterministic dynamics, starting from the current NOW state of the world’s atmosphere. The problem: this current state is only very imperfectly known, because weather stations and weather planes and satellites can only give exceedingly fragmentary data. How the atmosphere between the measured grid points is conditioned must be inferred by interpolation, and this is not exact. Many different earth atmospheres would agree with the currently known sparse measurements. Therefore, meteorologists run not a single, but dozens of simulation runs from different initial conditions, all of which agree with the currently known atmospheric measurements but use different interpolations (such a set of runs with slightly different initial conditions is called an *ensemble* in meteorology).

Weather Forecasting	
RSOI	The reality segment of interest is the collection of all atmospheric developments that start from a current NOW state which is compatible with the available physical measurements.
OO	Each atmospheric development that is compatible with known NOW data is an observation opportunity.
OP	The procedure is to run a (deterministic) simulation starting from one of the possible NOW states of the atmosphere, returning a global atmosphere "simulation video" up to a fixed time horizon, say ten days.
OA	Each actually executed simulation run is an observation act.
DVS	The data value space here is the collection of all 10-day simulation traces that are possible from any initial starting condition.

1.3 Discussion

The examples have been chosen to illustrate the great diversity of random systems, and to highlight some important issues. The good news up front: it is possible with the concepts and formalism of probability theory to treat all of these diverse cases in a unified fashion. Before we embark on the formalization, I discuss (still in plain English) some relevant issues that can be found in the examples.

First, a note on terminology. What we called a RSOI is often referred to in textbooks of statistics as a *population.* Using this word has historical reasons: the early development of statistical methods was very much driven by psychological research (specifically, IQ testing for military recruiting), and the the "observation opportunities" entities were, in fact, people. People still are important objects of statistical assessments in the modern world of Big Data — think of customer profiling or security & surveillance. However, many domains of statistical modeling are not populated by animate beings. Think of the Speed-of-light examples where the data-generating entities are physics laboratories. To convey the right idea of the abstractness and generality of statistical modeling, I used the term "reality segment of interest" instead of "population".

It is not easy to give a concise and general characterization of a RSOI. Here is an attempt:

A reality segment of interest is a circumscribed portion of reality, in which one has specified a set of opportunities for making observations.

Almost every part of this characterization is problematic and needs further comments.

"reality": There are varying degrees of how "real" a RSOI is. While the RSOI in our Die Throwing example is as real as you are (seize yourself by the arm and find out how real that is), the multiple labs (among which are all the ones that might still be built) in example Speed-of-light II is a collection where labs that actually currently do exist are mixed with "potential" labs. Similarly, in Credit Risk the piece of reality comprises students who might get born in the future. In Evolutionary Trees I lightheartedly included innumerable possible universes into the RSOI, and "observing" them meant simulating them. The definition given in Wolfram MathWorld (`http://mathworld.wolfram.com/Population.html`) for a population underlines this wide range of abstraction levels that is spanned by RSOIs:

> *"The word population has a number of distinct but closely related meanings in statistics.*
>
> 1. *A finite and actually existing group of objects which, although possibly large, can be enumerated in theory (e.g., people living in the United States).*
> 2. *A generalization from experience which is indefinitely large (e.g., the total number of throws that might conceivably by made in unlimited time with a particular pair of dice). [...]*
> 3. *A purely hypothetical population which can be completely described mathematically."*

"a circumscribed portion of...:" In any statistical data analysis, the underlying RSOI must be specified as precisely as possible. In the natural sciences, this typically means that a scientific paper must describe the experimental set-up and measurement procedures in enough detail to enable a reader of the paper to *reproduce* the experiment. In the social sciences, psychology, and clinical medicine, where actual human subjects are the source of data, the reference population from whom the analyzed data were obtained must be very well described in scientific publications. Readers of such papers should be aware that the results of statistical analyses reported in the paper only inform us about the particular "portion of humankind" that was available for the respective study. It is commonplace to complain that statistics lie, but in most cases it is not a lie that makes statistical claims dubious but a mismatch between the actual portion of reality where observation acts were carried out, and the intended RSOI which is decried in the resulting scientific publication (or in the second-hand commentaries about the original publication). For instance, scientific results in psychology are often based on populations consisting of undergraduate psychology students, because they can be easily recruited by the experimenters. Medical surveys are often based on patients that have been hospitalized. Then the reported results hold only for populations consisting of undergraduate psychology students and hospitalized patients, respectively — and they do *not* hold for humans in general. Figure 1.2 highlights what can come out of a mismatch between the RSOI a statistical study is based on, and the RSOI to which the findings are (wrongly) transferred.

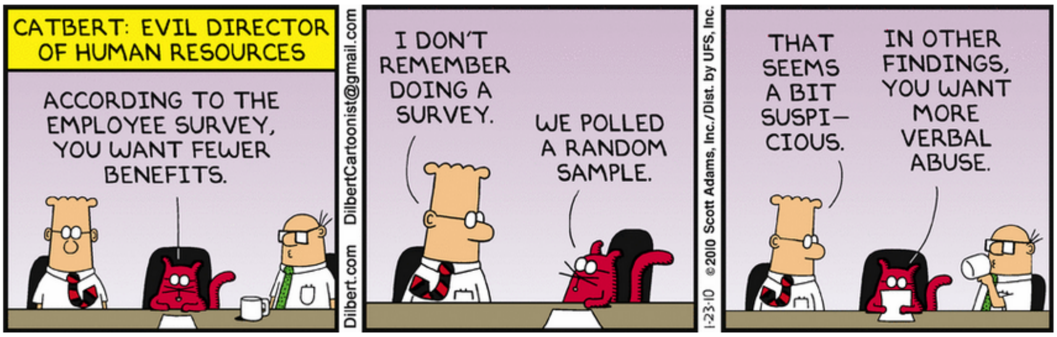

CATBERT: EVIL DIRECTOR OF HUMAN RESOURCES
ACCORDING TO THE EMPLOYEE SURVEY, YOU WANT FEWER BENEFITS.
Dilbert.com DilbertCartoonist@gmail.com
I DON'T REMEMBER DOING A SURVEY.
WE POLLED A RANDOM SAMPLE.
1-23-10 ©2010 Scott Adams, Inc./Dist. by UFS, Inc.
THAT SEEMS A BIT SUSPI-CIOUS.
IN OTHER FINDINGS, YOU WANT MORE VERBAL ABUSE.

Figure 1.2: Effects of a criminally chosen and under-documented RSOI. Taken from dilbert.com.

Another point worth noting: a RSOI is a piece of reality, and it usually cannot be described in full exactness. Plain English has to be used for the specification, and plain English always has some residual vagueness.

"... in which one has specified a set of opportunities for making observations": Here the critical, difficult word is "opportunity". It is difficult because depending on the case at hand, such "opportunities" can take on very different formats – they can be moments or intervals in space-time (as for example in Die Throwing), or (born or unborn) humans (Credit Risk), real or hypothetical physics labs (Speed of Light II), or possible states in an incompletely measured physical world (Weather Forecasting). – The crucial property of a RSOI which is relevant for statistical considerations is that it hosts a set of observation opportunties. For this reason, as we will soon see, in mathematical abstraction an RSOI is modeled exactly as that, just a *set* of opportunities where observations might be made.

In statistics and probability theory, an RSOI always comes with an observation procedure (OP). It is the very purpose of an RSOI to enable the recording of data. Like the RSOI, the OP is a piece of reality, subject to the same necessity to be specified as exactly as possible. Again, this typically means to sharpen your use of plain English, and provide a lot of detail. In the natural sciences, empirical papers have a "methods" section where among other things the OP is described. Often this drills down to indicate the manufacturer and type and version number of the measurement appartuses that were used. In psychology and the social sciences, the OP often is embodied in a questionnaire, which must be made accessible to the reader of the scientific study in some way for checking and re-use. But it is not only the questionnaire that makes the OP: when the questionnaire is distributed by (e)mail and has to be returned by (e)mail, which is a common procedure, this very procedure may have a critical impact on the study's validity because the recipients who return the questionnaire may be systematically different from the recipients who don't (the former are more collaborative, have more free time, need the honorarium more urgently, etc.). Describing and discussing these practical circumstances of the OP is part of a professional documentation of statistical analyses.

A given RSOI with its OOs is not tied to a specific OP. The "opportunities for making observations" provided by a RSOI can be exploited for making many kinds of observations. For instance, in Die Throwing, a throwing event might be observed in other ways than by recording the number of dots of the displayed face – a physicist might measure the impact energy of the throw, an audio scientist might record the sound of the die klicking away on the table, a psychologist might measure the heartbeat rate of the thrower... Or, in Credit Risk, the student could also have been asked other items besides the ones that figured on the handed-out questionnaire.

While the RSOI and the OP are real-world fragments (as far as "real" goes) and can only be specified with difficulties and in plain English, the data value space is a mathematical object and can be described with complete precision in mathematical terms. Formally the DVS is a *set*. This set can be finite (Coin Tossing, Die Throwing) or infinite (all other examples), it can contain integers or real numbers or labelled graphs or words or soundtracks or complex, heterogeneous data structures (like in Credit Risk). Note that a "soundtrack" can be mathematically described as a function from time points to sound amplitude values, or a "text" as a sequence of symbols from a finite alphabet. In fact, every kind of object that mathematicians have defined can figure as data values and be collected in a data value set.

Chapter 2

Where the randomness sits in the big picture: elementary events and random variables

So far we have introduced the main players in the statistical modeling game in informal terms: the reality segment of interest, the observation opportunities, the observation procedures, the observations acts, and the data value space. But — so far we haven't started talking about randomness! Hello - ?! Dear Probby (if I may call you like that) — where are you?

Randomness *appears* when the observation procedure is executed *repeatedly.* On each individual occasion when it is executed, one may obtain a data value that differs from the previously obtained ones, or the ones that one might obtain in the future. Which specific data value comes out of a particular observation act is "random".

The point has arrived to begin formalizing the big picture. Step by step we will develop a mathematical abstraction of the big picture, ending up with the very abstract axioms and fundamental definitions of *probability theory.* Probability theory is a branch of mathematics which affords us with a rigorous formal model of our big picture that we have introduced in plain English only so far. This formal model is very abstract, which makes it difficult to master on an intuitive level; on the other hand, its axioms and definitions are surprisingly short and innocently simple-looking, which makes it easy to memorize.

The way how mathematics abstracts away from concrete RSOIs is quite dramatic: in the formalism of probability theory, an RSOI is modeled just by a set, almost always denoted by Ω in the literature. The elements $\omega \in \Omega$ of this set represent the individual events where observations *could* be made — the observation opportunities (OO).

The mathematical terminology is not completely standardized. In different textbooks and papers, the set Ω is named *population*, *statistical population*, *universe*, *statistical universe* and probably by other names too. I will use "universe". Its elements ω are referred to as *elementary events.*

Side remark: I love the standard word for Ω in the German literature, which is "Grundgesamtheit". This literally translates to something like "Basic Totality" which sounds as comprehensive and abstract as it should be.

Please, never confound the elementary events $\omega \in \Omega$ with the data value outcomes. As explained above, it is helpful to conceive of elementary events as moments in (possibly hypothetical) space-time where some data *might* be recorded — *opportunities* to make some observation.

Re-iterating what I indicated earlier, an elementary event is not tied to a specific observation procedure. In Die Throwing, when the thrower throws a die (that is, some ω is realized), one could record other data besides the number of facing dots. For instance, one could also record the velocity of the die hitting the table. An elementary event is an occasion in space-time where *some*, or *any*, measurements can be made — *which* measurements are made, i.e. which particular recording procedure is carried out at the occasion ω, is not part of ω.

We will soon learn to appreciate the immense modeling power that comes out of this conception of elementary events being open to *any* kind of observation procedure. This non-commitment to a specific data recording procedure allows a formal model of a random system to become incrementally extended — new kinds of observations can be added to the picture as one deems fit. For example (this stupid textbook example again!), in statistical testing whether a die is fair, once a suspicion grows that it is not, an experimenter might record data from the throwing acts ω in more detail, e.g. by measuring throwing speed, impact-on-table energy, and what-not. More relevant examples: in a clinical follow-up study on the effectiveness of a new pharmaceutical substance, one can add novel physiological tests to the spectrum of medical indicator measurements that have been used in previous studies. Or, in robotics, sensor readings from a novel sensor that has been put on an old robot can be integrated into the statistical sensor data processing that is done by the robot's on-board control system ("sensor fusion" is a big topic in robotics, and the underlying mathematical methods make use of the idea that an observation opportunity admits many different kinds of observations, realized for instance by different sensors on board of the robot).

The next player to consider in our big picture is the observation procedure (OP). We have just seen that in fact one should think and speak of observation procedures (plural!) instead, because an elementary event ω can be observed in an essentially unlimited number of ways, that is, executing different OPs. We already established that data value spaces are formal mathematical objects — namely, sets S which contain all the possible outcomes of a particular recording procedure. A particular OP is always connected to a specific data value space S. If one adds more OPs to the picture, the DVS is extended by the new DVS of the new OP. Mathematically this is done by constructing *cross-products* of sets.

For instance, in Die Throwing, the initial basic OP is to just count the number of dots, giving $S = \{1, 2, 3, 4, 5, 6\}$. Then, at a later point when the checking of the suspicious die is done with more detail, also the speed of impact on the table is measured. Let the speed be measured in cm/sec, then a reasonable-looking measurement range might be the real interval $[0, 100]$. The DVS for the two-measurement procedure would be $S = \{1, 2, 3, 4, 5, 6\} \times [0, 100]$. Finally, also the loudness of the die hitting the table is recorded with a microphone, giving a reading in decibels. Assuming that a die's impact is never louder that listening to a large orchestra from sitting inside the orchestra – which is about 90 dB – a surely wide enough dB range for the little die is again $[0, 100]$. Adding sound measurement outcomes, S becomes $S = \{1, 2, 3, 4, 5, 6\} \times [0, 100] \times [0, 100]$. And so on.

Now let us become mathematicians. In mathematical abstraction, an OP is a *function* which turns elementary events $\omega \in \Omega$ into data values $s \in S$. Such functions are called *random variables* in probability theory. They are typically denoted by capital roman letters $X, Y, \ldots$. So here is our first piece of the mathematical abstraction of the big picture:

$$\boxed{X : \Omega \to S \qquad (2.1)}$$

I have spent about 20 pages to describe data generating environments, data recording procedures, and data value spaces. All of this is condensed in one of the tiniest mathematical expression you have seen in your life! Such is the power of mathematical abstraction.

A number of comments and further explanations:

- In order to emphasize that a given universe Ω of elementary events admits the use of many different recording procedures, it is maybe more revealing to write

$$X_i : \Omega \to S_i \qquad (i \in I), \tag{2.2}$$

  which expresses that the same underlying universe Ω can host an entire family of random variables that we index by indices i from an index set I. Each random variable X_i comes with its own data value space S_i.

- Terminology: the word "random variable" for the mathematical model of a OP is universally used by everybody. The mathematical set S that contains the possible data values returned by a random variable X is named differently by different authors. In statistics textbooks it is mostly called the *sample space*. As we will later see this is a somewhat unlucky naming — we will later formally define "samples" and we will see that samples are *not* the elements of S. In my various legacy machine learning lecture notes (online at `http://minds.jacobs-university.de/teaching/ln`) I mostly used the word "observation space" for S. However, in this lecture I bend to tradition and call S by the name of *sample space*.

- Random variables are arguably the most misleadingly named mathematical objects that we have to live with. A random variable X is neither a variable (it is a function!), nor is it random (it is reliably and deterministically returning the result $X(\omega) \in S$ when applied to the argument ω). The randomness in our big picture and its mathematical abstraction is not created or modeled by the random variables. It resides in the universe Ω which has the property that different elementary events $\omega \in \Omega$ will lead to different data values when the (deterministic!) function X is applied.

We will use the acronym "RV" for "random variable" henceforth.

Chapter 3

Basic operations on RVs: products, projections, transformations

From the perspective of maths, RVs are just functions. The standard operations that can be applied to any mathematical function can be applied to RVs, too. In the context of probability theory, it is interesting to take a closer look at these and discuss what these operations mean in our big picture.

3.1 Products and projections

Let us consider again Die Throwing. Assume the experimenter counts the number of facing dots *and* measures the throwing velocity, for each throw. In formal abstraction this means we have two RVs

$$\begin{aligned} X_1 : \quad & \Omega \to S_1 \\ X_2 : \quad & \Omega \to S_2, \end{aligned}$$

where X_1 models the counting of dots OP (hence $S_1 = \{1, \ldots, 6\}$) and X_2 models the velocity measurement (hence we may set $S_2 = [0, 100]$, for instance, assuming that nobody can throw a die with more than 100 m/sec). We can equivalently tie both RVs into a single one, getting a compound RV X

$$X : \Omega \to S_1 \times S_2, \quad X(\omega) = (X_1(\omega), X_2(\omega)) \in S_1 \times S_2.$$

The outcomes of this compound OP are pair values whose first component is the observed number of dots and the second component is the velocity measurement result. In mathematical terms, X is the *product* of X_1 and X_2, which is written $X = X_1 \otimes X_2$. The sample space $S_1 \times S_2$ is the product set of the sets S_1 and S_2.

Pair-building operations like $(X_1(\omega), X_2(\omega))$ or product operations like $\times$ or $\otimes$ occur frequently in probability theory. They can be defined in very general and abstract ways. Appendix A gives the mathematical background story.

The converse of creating products from components is to pick components from a product. We denote by π_i the operation to pick the i-th element from a tuple. Thus, taking again $X : \Omega \to S_1 \times S_2, \quad X(\omega) = (X_1(\omega), X_2(\omega)) \in S_1 \times S_2$, from $S = S_1 \times S_2$ we can recover the component sets S_1 and S_2 by

$$S_1 = \pi_1(S) \text{ and } S_2 = \pi_2(S).$$

We use the same notation π_i also for picking component RVs X_i from a product RV $X = X_1 \times X_2$:

$$X_1 = \pi_1(X) : \Omega \to S_1 \text{ and } X_2 = \pi_2(X) : \Omega \to S_2.$$

This generalizes in an obvious way to more than 2 components.

In the CREDIT RISK example a data value was a list of info items supplied by the student. To make this concrete, assume this list had 50 entries, with the first info item being the student's age (a number between 1 and 100); the second his/her gender (either "F" or "M" or "O"); the third his/her nationality (one of 200 or so 3-letter codes like ABK (Abkhazia), AFG (Afghanistan), ..., ZIM (Zimbabwe); then many more that we won't detail, and the 50th a copy of the student's visum for Germany, given in jpeg format (that is a binary document, formally a word of 0's and 1's). Formally, S here is the set of 50-tuples of mixed type

$$S = \{1, \ldots, 100\} \times \{F, M, O\} \times \{ABK, \ldots, ZIM\} \times \ldots \times \{0,1\}^*, \tag{3.1}$$

and $X : \Omega \to S$ models the OP when a student is asked to supply all the 50 info items and those become entered into the bank's database. Conversely, for $1 \leq i \leq 50$, $\pi_i(X)$ models the OP "record the i-th item of the bank's questionnaire".

These two examples demonstrated *finite* products of RVs, leading to tuple value spaces. It is also possible to bind together infinitely many RVs. In the SENTENCE TRANSLATION scenario, the pairs of sentences (s, s') would be modeled by two families of RVs $(X_n)_{n\in\mathbb{N}}, (Y_n)_{n\in\mathbb{N}}$, where X_n returns the n-th letter in the English sentence s and Y_n the n-th letter in the German translation s'. Each X_n, Y_n takes values in $S = \{a, \ldots, z, A, \ldots, Z, ,, ;, ., \_, \#\}$. We use the $\#$ symbol to "fill time" after the end of a sentence, that is, $X_n(\omega) = \#$ when n is greater than the length of the sentence. This is one way to deal with variable-length timeseries data. Binding together each of the two families X_n, Y_n gives

$$\begin{aligned} X &:= \bigotimes_{n\in\mathbb{N}} X_n, \\ Y &:= \bigotimes_{n\in\mathbb{N}} Y_n, \end{aligned}$$

where

$$\begin{aligned} X &: \Omega \to \prod_{n\in\mathbb{N}} S, \\ Y &: \Omega \to \prod_{n\in\mathbb{N}} S. \end{aligned}$$

Infinite products of RVs are denoted with the symbol $\bigotimes$ (see Appendix A for the pure math story of such products of functions). This would yield values of X, Y of the form

$$\begin{aligned} X(\omega) &= \text{I \_ b u i l t \_ a \_ h o u s e . \# \# \# ...} \\ Y(\omega) &= \text{I c h \_ b a u t e \_ e i n \_ H a u s . \# \# \# ...} \end{aligned}$$

For a slightly more involved case, consider our SPEED-OF-LIGHT II example. Recall that in this scenario we declared an "experiment" to consist of a repeated execution of speed-of-light measurement trials carried out in some lab. How many such measurements are done is not fixed, different labs can repeat trials different numbers of times. In our first plain-English rendering we captured this by setting $S = \{(x_1, \ldots, x_{T_{\text{stop}}}) | T_{\text{stop}} \in \mathbb{N}, x_i \in$

$\mathbb{R}^{\geq 0}\}$. The number N of how many trials are executed before the experiment is stopped varies between experiments. The standard way of how such a situation is modeled in probability theory is to turn T_{stop} itself into a RV, called a *stopping time.* The speed value measured in the n-th trial of an experiment is modeled as the value returned by a RV X_n. For a mathematically homogenous model, one does not restrict n to not exceed the stopping time, but instead "records" the last trial's speed value again and again for all times $n \geq T_{\text{stop}}$. Using stopping times is another way to deal with variable-length serial data. Concretely, for example, if in some experiment in some lab three trials are done with outcomes x_1, x_2, x_3, this would be formally modeled by an infinite sequence $x_1, x_2, x_3, x_3, x_3, \ldots$ of speed values, together with the fact that the stopping time was $T_{\text{stop}} = 3$. More abstractly one introduces a RV T_{stop} and an infinite sequence of RVs $X_1, X_2, \ldots$ with

$$\begin{aligned} T_{\text{stop}}: & \quad \Omega \to \mathbb{N}, \\ X_n: & \quad \Omega \to \mathbb{R}^{\geq 0} \qquad \text{for } n \in \mathbb{N}, \end{aligned}$$

and requires that

$$\forall \omega \in \Omega, \forall n > T_{\text{stop}}(\omega) : X_n(\omega) = X_{T(\omega)}(\omega).$$

Thus we could join all the RVs of our model of Speed-of-light II into a single product RV E (for "experiment") by

$$\begin{aligned} E & := T_{\text{stop}} \otimes X_1 \otimes X_2 \otimes \ldots \\ & = T_{\text{stop}} \otimes \bigotimes_{n \in \mathbb{N}} X_n, \end{aligned}$$

where

$$E : \Omega \to \mathbb{N} \times \prod_{n \in \mathbb{N}} \mathbb{R}^{\geq 0}. \tag{3.2}$$

Here $\prod_{n \in \mathbb{N}} \mathbb{R}^{\geq 0}$ denotes the $\mathbb{N}$-fold product of the set $\mathbb{R}^{\geq 0}$. Check Appendix A for the mathematical definition of infinite products of sets. Intuitively, $\prod_{n \in \mathbb{N}} \mathbb{R}^{\geq 0}$ is the set made of all right-infinite sequences of nonnegative reals. In our example where only three trials are made, one would concretely have

$$E(\omega) = (3, x_1, x_2, x_3, x_3, x_3, \ldots),$$

with x_1, x_2, x_3 the three measured speeds of light in this experiment ω.

Again, we may isolate the components of a product RV by projections. In our example we would get $\pi_1(E) = T_{\text{stop}}, \pi_2(E) = X_1, \pi_3(E) = X_2$, etc.

Further Comments:

- Finite and infinite products of RVs and sample spaces play a *major* role in probability theory and its applications in statistics and machine learning. Make sure you establish an unshakeable friendship with them.

- Just a side remark: probability scenarios that involve stopping times occur frequently and importantly in economics and game theory. The stopping times then model occurances and decisions like "game over", "bankruptcy", "decision to quit a bidding contest" or the like.

3.2 Transformations of RVs

In statistical data analysis, the “raw” data that are recorded by some data recording procedure are often too bulky, too noisy or too redundant to be useful. One therefore often submits them to some *preprocessing* which compresses them, or de-noises them, or in some other way makes the data format more easily analysable.

A very common preprocessing step is to reduce numerical accuracy of raw data. For instance, in our Credit risk example, one of the questionnaire fields might be “regular monthly income”. A student might fill in a value of, say, 782.45 Euro. For credit risk estimation such a degree of precision is irrelevant, so the raw data value will be rounded to a precision of 100 Euros, transforming the raw numerical value 782.45 into the range value $[700 - 799.99] =: b_7$. Such coarse data range segments are called *bins* and the process of truncating precision to bin values is called *binning*.

This is expressed in formal maths as follows. Assume, in this example, that the RV for this questionnaire field was $X_{20} : \Omega \to S_{20}$, with $S_{20} = \mathbb{R}^{\geq 0}$. The raw value $X_{20}(\omega) = 782.45$ is passed to a function, call it β, which returns the associated bin. For instance, $\beta(782.45) = b_7$ or $\beta(512.12) = b_5$. The domain of β is S_{20}, the codomain is $S'_{20} := \{b_0, b_1, \ldots\}$. Now consider the composite function

$$\beta \circ X_{20} : \Omega \to S'_{20}.$$

This composite function $\beta \circ X_{20}$ is again a random variable! Generally, if

$$X : \Omega \to S$$

is some RV and

$$f : S \to S'$$

is some function, then

$$f \circ X : \Omega \to S'$$

is again a random variable — a *transform* of X.

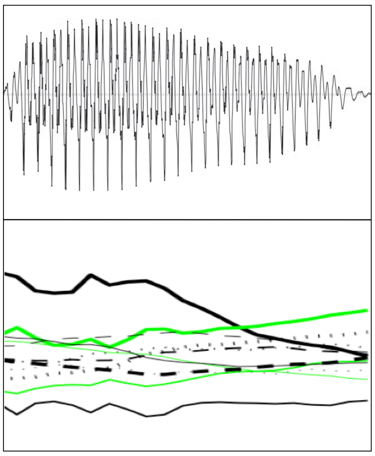


Figure 3.1: Transforming an audiosignal (top) into a “Mel-Cepstrum coefficient” representation (bottom) [schematic]. Upper picture taken from `aile.revues.org/4533`.

For a more involved case of a transformation, consider the example Speech recognition. The data recording procedure presumably involved a microphone connected to a computer, and the data value space was made of the sound recordings obtained with this apparatus, for instance in the format of a raw acoustic waveform recording (see upper panel in Figure 3.1). This format is not very practical for automated speech recognition algorithms. Instead of feeding such “raw” audiosignals into the recognizer algorithm, the raw data are *pre-processed* by some filtering procedure that transforms them into another representation format which is more suitable for further algorithmic processing. Figure 3.1 shows how this is standardly done in speech processing. The raw soundwave recording is submitted to a certain kind of sound frequency analysis, leading to a new representation that consists of time traces of 12 “Mel-Cepstrum coefficients”. As you can see in Figure 3.1, the transformed signal looks “simpler”, and indeed it is: storing it needs much less computer memory than storing the original soundrecording. Such a reduction of data size is an important function of preprocessing in many modeling applications. Another role of preprocessing, equally useful, is to get rid of “irrelevant” aspects of the raw data, leaving only what is maybe important for the particular modeling task (here: recognizing what has been said).

In formal terms, the original sound recording can be modeled by a RV $X : \Omega \to S$, where S is the space of all possible raw sound recordings. The preprocessing here is an algorithmic procedure β, actually a quite complicated one, which transforms a raw sound recording to the Mel-Cepstrum coefficient format.

Transformations of RVs can be iterated, always leading to further RVs. For instance, in the SPEECH RECOGNITION example, the Mel-Cepstrum coefficient format, which is made of 12 timeseries, can be further trimmed down to just a vector of 12 numbers, each of them being the mean value of the corresponding Cepstrum coefficient timeseries averaged over recording time.

Chapter 4

Modeling temporal data by stochastic processes

A stochastic process is the mathematical model of a system which evolves in time and shows some randomness in doing so. Examples: the erratic rises and falls of stock market indices; talking humans; electromagnetic signals picked up by an antenna; ... well, essentially everthing under the sun that evolves in time. And since essentially everything under the sun evolves in time, basically all real-world data are (or should be modeled as) time series. Such data obviously are super important in the data sciences. In this section I explain how such temporal random data are modeled by RVs.

When modeling a temporal system, the first item a modeler has to provide is a model of time. A main distinction is between *discrete* time and *continuous* time. In discrete-time models, time is progressing stepwise, leading to a sequence of time points. The can be separated by real-world timespans, like in daily recordings of stock markets (1 increment = 1 day), or like in sampling microphone signals for digital signal processing (1 increment = 1/20000 sec in 20 kHz sampling). Or they can be separated just by unit increments when a real-world duration is irrelevant, like when a typed text is seen as a temporal sequence of letters. In any case, the timesteps can be mapped to the integers, and one ends up with three kinds of sets T of time points that model discrete timelines:

- Finite discrete time: the set T of time points is a finite sequence of integers, $T = (0, 1, \ldots, N)$.
- Open-ended discrete time: the set of time points is the right-infinite sequence $0, 1, 2, \ldots$, that is, $T = \mathbb{N}$.
- Two-sided infinite discrete time: $T = \mathbb{Z}$.

In continuous-time models, time is conceived as a continuous flow of time points, with the analog main distinctions of finite intervals $T = [t_{\min}, t_{\max}\,] \subset \mathbb{R}$, right-infinite intervals like $T = [0, \infty) = \mathbb{R}^{\geq 0}$ or two-sided infinite time $T = \mathbb{R}$.

After this modeling decision has been made, i.e. after one of the six kinds of time point sets T just listed has been chosen, the core piece of a mathematical model of the respective random process is an indexed family of RVs

$$(X_t)_{t \in T},$$

with the interpretation of T as an (ordered) set of timepoints. Each RV X_t then delivers the data value of the random system at time t (for discrete time one often prefers n over t, giving $(X_n)_{n \in T}$).

Now comes the crucial part. All the X_t from such an indexed family share the same image set S, that is, each of them is a function $X_t : \Omega \to S$. A single elementary event $\omega \in \Omega$ thus gives an indexed family $(X_t(\omega))_{t\in T} \in \prod_{t\in T} S$ of observation values in S. Intuitively, $(X_t(\omega))_{t\in T}$ can be viewed as a timeline recording, and be visualized as a graph where the observation values at times t are plotted against T. The value sequence $(X_t(\omega))_{t\in T}$ is called a *path* or *realization* or *trajectory* of the stochastic process.

Figure 4.1 gives a discrete-time example where $T = (0, 1, \ldots, 4)$.

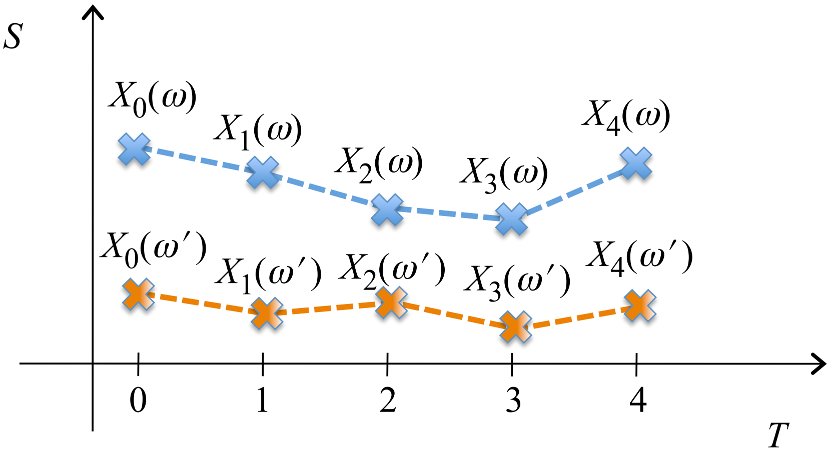


Figure 4.1: Two paths $(X_t(\omega))_{t\in T}$ and $(X_t(\omega'))_{t\in T}$ of a stochastic process defined for discrete time points $0, 1, \ldots, 4$ and $S = \mathbb{R}$.

Hold your breath: if one models a continuous-time system with paths like $(X_t(\omega))_{t\in\mathbb{R}}$, *uncountably many* RVs are needed! If you don't like that, you can equivalently group them all together and be left with a single RV $X = \bigotimes_{t\in\mathbb{R}} X_t$ whose values $X(\omega)$ then are functions $\varphi : \mathbb{R} \to S$.

Advice: don't continue reading before you feel sure you have a full grasp on these constructions and notations. You'll be using them heavily all over the place.

Chapter 5

Interim summary: formalizing the big picture

Let us assemble all the items that we have met so far in a take-home summary:

- The main real-world objects of statistical modeling are ensembles made of three items: a data-generating environment RSOI, a data-recording procedure OP, and a data value space DVS.
- RSOIs are systems in which data recording acts *can* be made. When defining a RSOI, one therefore often has to specify *potential* opportunities for data recording acts besides the data recording acts that have already materialized.
- These "opportunities for data recording acts" are not tied to a particular type of measurement or observation. In principle, at each such occasion, further data recording procedures could be carried out besides the ones that are initially included in the picture. This enables an incremental extension of statistical models.
- The mathematical abstraction of these three items are collected in the following table:

Real-world item	**Its mathematical abstraction**
RSOI, a collection of opportunities for data recording acts	A *universe* Ω made of *elementary events* ω
DVS, the set of possible measurement outcome values	*Sample space* S
OP, a (possibly compound) procedure to record data when an occasion materializes	A *random variable* $X : \Omega \to S$. It is often useful to establish X as a product $X = \bigotimes X_i$, where each $X_i : \Omega \to S_i$ provides a component of the compound values found in $S = \prod S_i$.

Chapter 6

Structure in randomness: events

So far, our mathematical model is just this, $X : \Omega \to S$. It allows us to to model that on different recording opportunities ω we get different observation values $X(\omega)$, but that's it. So far we cannot model what is meant when a data scientist says, "the probability of the die showing a 2 is larger than 1/6", or "the probability of student Dilbert Doolittle paying back his loan is estimated to be 0.8", etc. We will now add probability to the picture, step by step. The first step is to clarify what things can have a probability in the first place — that is, when we say, "the probability of ***", what is this "***"?

The "***" things are *events*, and understanding the nature of events is likely the biggest hurdle to making friends with probability.

6.1 Events

The key concept to formalize probability is the notion of an *event*. Here is a first definition of an important special type of event:

Definition 6.1.1 *Let $X : \Omega \to S$ be a RV X defined on the universe Ω, taking values in S. Let $A \subseteq S$ be a subset of S. Then the* event of X taking a value in A *is the set* $\{\omega \in \Omega \mid X(\omega) \in A\}$.

Events are always subsets of Ω. Here we have defined a special kind of events, namely events defined by the condition "X is taking a value in A". This is in fact the only kind of event that we need for statistics and machine learning. Other kinds of events (other subsets of Ω) are considered in abstract mathematical probability theory, but we will not have to deal with them. For all that matters to us, you can identify "event" with "set of elementary events where X is taking value in A".

By an abuse of terminology, we will also allow ourselves to call a set $A \subseteq S$ an event – although, rigorously taken, it is not an event itself but a set defining an event. Since we will be considering only events of the kind "X takes value in A", this does no harm.

Notation: the event $\{\omega \in \Omega \mid X(\omega) \in A\}$ is also written as $X^{-1}(A)$ or as $X \in A$. The latter notation is "dirty" because of course X is not an element of A, but still this notation is often used.

Let us inspect some examples.

- In Die throwing consider $A_1 = \{1\} \subseteq \{1, \ldots, 6\} = S$. Then $X^{-1}(A_1)$ is the set of all throws (realized or hypothetical) that yield a facing single dot. And for $A_2 = \{1, 2, 3\}$, $X^{-1}(A_2)$ is the event consisting of all throws whose outcome is 1 or 2 or 3 dots.

- In Speed-of-light I let us consider $A = [290000000, 300000000] \subseteq S$, the interval of real numbers between 290000000 and 300000000. Then $X^{-1}(A)$ is the set of all single-trial measurements in the modeled Lab where a value between these two limits is measured. Of course this again includes "potential" measurement trials yet to executed. (Btw., physicists have *defined* the length of a "meter" to be the distance that light travels in 1/299792458 sec. Therefore the speed of light is *exactly* 299792458 m/s, by definition. If deviations from this value would be found by high-precision experiments, then the length of a meter would need to be corrected, not the speed of light! What is measured in such Labs is actually not the speed of light, but the length of a meter! and don't ask me how *seconds* are defined by physicists... [you find an answer in `en.wikipedia.org/wiki/Speed_of_light`].)

- An observation value set A that defines an event can be quite richly structured when the sample space is richly structured. In Speed-of-light II consider the rendering of the sample space $S = \mathbb{N} \times \prod_{n \in N} \mathbb{R}^{\geq 0}$ that we introduced in Equation 3.2. This is an infinite product of sets. A subset worth considering might be the set of all elementary events (that is, measurement experiments) where at least one of the recorded values exceeds 300000000. That is, we would consider

  $$A = \{(n, x_1, x_2, \ldots) \mid \text{at least for one } i \in \mathbb{N}, x_i > 300000000\}.$$

  For later use we note that this set A could also be formalized as follows. Clearly, $(n, x_1, x_2, \ldots) \in A$ if and only if $x_1 >$ 3e8 **or** $x_2 >$ 3e8 **or** $x_3 >$ 3e8... Denote by G the right-infinite interval $(30000000, \infty) \subset \mathbb{R}^{\geq 0}$ of reals exceeding 3e8. Since logical **or** translates to set union, our set A can also be written as a countably infinite union:

  $$A = \bigcup_{i \in \mathbb{N}} \left(\mathbb{N} \times \prod_{n=0,\ldots,i-1} \mathbb{R}^{\geq 0} \times G \times \prod_{n>i} \mathbb{R}^{\geq 0} \right). \tag{6.1}$$

  Then $X^{-1}(A)$ is the event of all experiments (physically executed or hypothetical) where at least one measurement trial gives a value greater than 3e8.

- In Credit risk let us assume, for concreteness, that the bank collects data records $X(\omega)$ consisting of the 50 items indicated in Equation 3.1. The bank's analyst may be interested to model clients that are **not** German citizens. To this end, the analyst inspects the event

  $$A = \{(x_1, x_2, \ldots, x_{50}) \mid x_3 \neq \text{'GER'}\} \subset S.$$

  Again for later use we note that, since the logical **not** is captured by set complement, A can also be written as

  $$A = S \setminus (\{1, \ldots, 100\} \times \{F, M, O\} \times \{GER\} \times \ldots \times \{0,1\}^*). \tag{6.2}$$

  Then $X^{-1}(A) \subset \Omega$ is the event that contains all the potential query acts imposed by the bank on a student where the student fills in another code than GER in the nationality field.

Note that elementary events ω are not events of the type "event of X taking value in A". There are two reasons why not. First, a formal one: An elementary event ω is an *element* of Ω, not a *subset*. If one wishes to consider elementary events ω as events in our new sense, then one would have to re-interpret them as singleton sets $\{\omega\}$. The second reason is more substantial. Even when we consider singleton subsets $\{\omega\} \subset \Omega$, and have a RV $X : \Omega \to S$, then usually there exists no subset $A \subseteq S$ such that $\{\omega\} = X^{-1}(A)$. This is because different elementary events ω, ω' may yield the same outcomes under X, i.e. $X(\omega) = X(\omega')$ is typically possible — RVs are usually not injective. But then, for any $A \subseteq S$, when $\omega \in X^{-1}(A)$, also $\omega' \in X^{-1}(A)$ — in other words, ω is indistinguishable from ω' observation-wise.

6.2 A brief look at sigma-fields

In Definition 6.1.1 I referred to "a" subset A of S. However, not any arbitrary subset $A \subseteq S$ is admissible in probability theory. If one would admit any subset $A \subseteq S$ for generating an event, probability theory would run into a host of technical difficulties that would cripple it; most theorems of today's probability theory would not hold or would even be ill-defined. It has taken mathematicians centuries to work out an appropriate set of constraints on subsets $A \subseteq S$ such that the resulting *probability theory* is mathematically manageable on the one hand, and rich enough on the other to cover all phenomena of interest. The resulting theory of *σ-fields* (also called *σ-algebras*) rests on simple axioms (which we will present), but these simple axioms spawn a very intricate field of mathematics called *measure theory* (which we will not touch).

Technically, given a set S, a σ-field $\mathcal{F}$ over S is a subset of the power set of S, subject to certain constraints. Thus $\mathcal{F} \subseteq \mathbf{Pot}(S)$, where $\mathbf{Pot}(S)$ denotes the power set of S, i.e. the set of all subsets of S. In words, σ-fields over S are certain collections of "admissible" subsets of S.

In the examples given in the previous subsection I pointed out that certain relevant sets $A \subseteq S$ were obtained by set complement (Equation 6.2) and by countably infinite union (Equation 6.1). This was related to the use of **or** and **not** arguments. These logical operations occur naturally in statistical reasoning. Statistical modelers want to be able to say things like, "there is a high probability that the human species is closely related to chimps **or** gorillas", or "I wonder what is the probability that this loan will **not** be repaid". (The logical **and** can be expressed in terms of **or** and **not**, so we don't have to consider it separately). Logical and statistical reasoning are not that far away from each other! This intimate connection between logical reasoning and reasoning about statistical events is directly captured in the formal definition of a σ-field:

Definition 6.2.1 *A collection $\mathcal{F} \subseteq \mathbf{Pot}(S)$ of subsets of a set S is a σ-*field *over S if it satisfies the following conditions:*

1. *$S \in \mathcal{F}$,*
2. *if $A \in \mathcal{F}$, then $A^{\mathrm{c}} = S \setminus A \in \mathcal{F}$ (closure under complement),*
3. *if $A_1, A_2, \ldots$ are all elements of $\mathcal{F}$, then $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$ (closure under countably infinite union).*

Note that closure under countably infinite union includes closure under finite union. To see this, write $A_1 \cup A_2 = A_1 \cup A_2 \cup A_2 \cup A_2 \cup \ldots$.

Two immediate consequences: a σ-field $\mathcal{F}$ always contains the empty set (because $\emptyset = S^{\mathrm{c}}$), and $\mathcal{F}$ is also closed under countable intersection (follows from de Morgan's law). Note that intersection of sets corresponds to logical **and**. In sum, the sets occurring in a σ-field $\mathcal{F}$ admit all kinds of iterated "Boolean" re-combinations, and the sets thus obtained are also in $\mathcal{F}$.

Given a set S which contains more than one element, there exist more than one σ-field over S. It is an easy exercise to deduce from Definition 6.2.1 that the smallest σ-field over S is given by $\mathcal{F} = \{\emptyset, S\}$, and the largest is $\mathcal{F} = \mathbf{Pot}(S)$. The σ-fields that are possible for a given S are by no means unique. A statistical modeler must choose the σ-field that is best suited for the modeling purpose at hand.

There are two special kinds of sample spaces S which occur very often and whose σ-fields are almost always chosen in the same way:

Finite S. When the sample space is finite (as in DIE THROWING where $S = \{1, \ldots, 6\}$), one typically uses the power set σ-field $\mathcal{F} = \mathbf{Pot}(S)$.

Continuous S. When the sample space S is $\mathbb{R}$ or a part of $\mathbb{R}$ (like in SPEED-OF-LIGHT II where it was $\mathbb{R}^{\geq 0}$), then one typically uses a specific σ-field known as the *Borel σ-field.* We denote it by $\mathcal{B}(\mathbb{R})$. The Borel σ-field is a very richly structured thing, and we cannot explore it here in all the detail that it deserves. For us it is enough to know that the Borel σ-field $\mathcal{B}(\mathbb{R}) \subset \mathbf{Pot}(\mathbb{R})$ contains all intervals of the real line, including singleton intervals and infinite intervals. Concretely, for real numbers $a < b$, the following intervals are contained in $\mathcal{B}$: $\{a\}$ (the "point interval" consisting only of a); $[a, b], (a, b], [a, b), (a, b)$ (the closed, half-open, and open intervals with limits a, b); the half-infinite intervals $[a, \infty), (a, \infty), (-\infty, a], (-\infty, a)$ and, of course, $(-\infty, \infty) = \mathbb{R}$ itself.

Without proof I mention the following

Theorem 6.2.1 *If $(\mathcal{F}_i)_{i \in I}$ is a family of σ-fields over a set S, then*

$$\bigcap_{i \in I} \mathcal{F}_i = \{A \subseteq S \mid \forall i \in I : A \in \mathcal{F}_i\}$$

is a σ-field over S, too.

The proof is easy, do it. In math terminology, this theorem says that σ-fields are closed under arbitrary intersections.

Theorem 6.2.1 paves the way to *define* interesting non-trivial σ-fields. Consider a collection $\mathcal{G} \subseteq \mathbf{Pot}(S)$ of subsets of S. Then we can define the σ-field *generated by $\mathcal{G}$* as the smallest σ-field over S which contains all the sets from $\mathcal{G}$:

Definition 6.2.2 *For $\mathcal{G} \subseteq \mathbf{Pot}(S)$,*

$$\sigma(\mathcal{G}) := \bigcap_{\mathcal{F} is\ a\ \sigma-field\ over\ S and\ \mathcal{G} \subseteq \mathcal{F}} \mathcal{F}$$

is the σ-field generated by $\mathcal{G}$.

The Borel-σ-field $\mathcal{B}(\mathbb{R})$ is, by one of its many possible definitions, the σ-field over $\mathbb{R}$ which is generated by all the open intervals (a, b) in the real line $\mathbb{R}$. Because some familiarity with the Borel-σ-field is required for working in probability theory, we now exercise our skills a little by investigating some sets that are in $\mathcal{B}(\mathbb{R})$:

1. $\mathbb{R} \in \mathcal{B}(\mathbb{R})$ because, by the definition of a σ-field, $\mathbb{R}$ is contained in every σ-field over $\mathbb{R}$, hence it is in $\mathcal{B}(\mathbb{R})$ too.

2. Every right-infinite open interval (a, ∞) is in $\mathcal{B}(\mathbb{R})$ because it can be written as a countable union of sets from $\mathcal{G}$:

$$(a, \infty) = \bigcup_{n \in \mathbb{N}} (a, a+n),$$

hence every σ-field $\mathcal{F}$ containing $\mathcal{G}$ must also contain (a, ∞) due to Definition 6.2.1, which implies by Definition 6.2.2 that also $\mathcal{B}(\mathbb{R})$ contains (a, ∞). Similarly, the left-infinite open intervals $(-\infty, a)$ are in $\mathcal{B}(\mathbb{R})$.

3. Every closed interval $[a, b]$ is in $\mathcal{B}(\mathbb{R})$ because each such interval can be written as

$$[a, b] = ((-\infty, a) \cup (b, \infty))^{\mathsf{c}}.$$

Since we already know that $(-\infty, a), (b, \infty) \in \mathcal{B}(\mathbb{R})$ and σ-fields are closed under finite unions and complements of sets, this gives us $[a, b] \in \mathcal{B}(\mathbb{R})$.

4. Every singleton set $\{a\}$ is in $\mathcal{B}(\mathbb{R})$ because such sets can be written as $\{a\} = ((-\infty, a) \cup (a, \infty))^{\mathsf{c}}$.

5. Every countable set $\{a_1, a_2, \ldots\}$ of points $a_i \in \mathbb{R}$ is in $\mathcal{B}(\mathbb{R})$ because it is a countable union of singleton sets, which we know are in $\mathcal{B}(\mathbb{R})$.

6. Every set $A \subseteq \mathbb{R}$ that is open in the standard metric topology of $\mathbb{R}$ is in $\mathcal{B}(\mathbb{R})$ because, by a theorem of topology, every such set can be written as a countable union of open intervals.

7. Likewise, every closed subset of $\mathbb{R}$ is in $\mathcal{B}(\mathbb{R})$ because these sets are the complements of open sets.

We have seen that one frequently combines sets into set products. If on two sets S_1, S_2 there are two σ-fields $\mathcal{F}_1, \mathcal{F}_2$, the natural construction to define a σ-field on $S_1 \times S_2$ is the *product* of $\mathcal{F}_1$ and $\mathcal{F}_2$:

Definition 6.2.3 *Let S_1, S_2 be two sets equipped with σ-fields $\mathcal{F}_1, \mathcal{F}_2$. The* product *$\mathcal{F}_1 \otimes \mathcal{F}_2$ of $\mathcal{F}_1$ and $\mathcal{F}_2$ is the σ-field on $S_1 \times S_2$ which is generated by all sets $A \times B$, where $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$.*

This is not a very practically useful definition, because it is based on all sets in $\mathcal{F}_1$ and $\mathcal{F}_2$, which in infinite σ-fields are virtually unmanageably complex. The following theorem allows us to construct product σ-fields from generating sets, which are usually much simpler:

Theorem 6.2.2 *Let S_1, S_2 be two sets equipped with σ-fields $\mathcal{F}_1, \mathcal{F}_2$. Let $\mathcal{F}_1, \mathcal{F}_2$ be generated by $\mathcal{G}_1, \mathcal{G}_2$, respectively. Furthermore, let $\mathcal{G}_1$ contain a countable family $(A_n)_{n \in \mathbb{N}}$ such that $\bigcup_n A_n = S_1$, and let similarly $\mathcal{G}_2$ contain a countable family $(B_n)_{n \in \mathbb{N}}$ such that $\bigcup_n B_n = S_2$. Then $\mathcal{F}_1 \otimes \mathcal{F}_2$ is generated by $\mathcal{G}_1 \times \mathcal{G}_2$.*

For an illustration, we make use of this theorem to get an impression on the Borel-σ-field $\mathcal{B}(\mathbb{R}^2)$ on $\mathbb{R} \times \mathbb{R}$. We know that $\mathcal{B}(\mathbb{R})$ is generated by the open intervals (a, b) in the real line. This generator satisfies the conditions of Theorem 6.2.2. The Borel-σ-field $\mathcal{B}(\mathbb{R}^2)$ on $\mathbb{R} \times \mathbb{R}$ is hence generated by all open rectangles $(a, b) \times (c, d)$ in the two-dimensional plane, that is,

$$\mathcal{B}(\mathbb{R}^2) = \sigma(\{(a, b) \times (c, d) \mid a < b \text{ and } c < d\}) =: \sigma(\mathcal{H}).$$

For an exercise in "σ-thinking", we demonstrate that the line $L = \{(x,x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$ is in $\mathcal{B}(\mathbb{R}^2)$. Denote the rectangle $(x, x+a) \times (x, x+a)$, where $n \in \mathbb{N}$, by $R(x,a)$. Clearly each $R(x,a)$ is in $\mathcal{H}$. For $n \geq 0$ consider the set

$$T_n = \bigcup_{j \in \mathbb{Z}} R(j \cdot 2^{-(n+1)}, 2^{-n}).$$

Since each T_n is a countable union of sets from the generator $\mathcal{H}$, $T_n \in \mathcal{B}(\mathbb{R}^2)$. But it is easy to see (compare Figure 6.1) that

$$L = \bigcap_n T_n,$$

that is, L is a countable intersection of elements from $\mathcal{B}(\mathbb{R}^2)$, hence $L \in \mathcal{B}(\mathbb{R}^2)$.

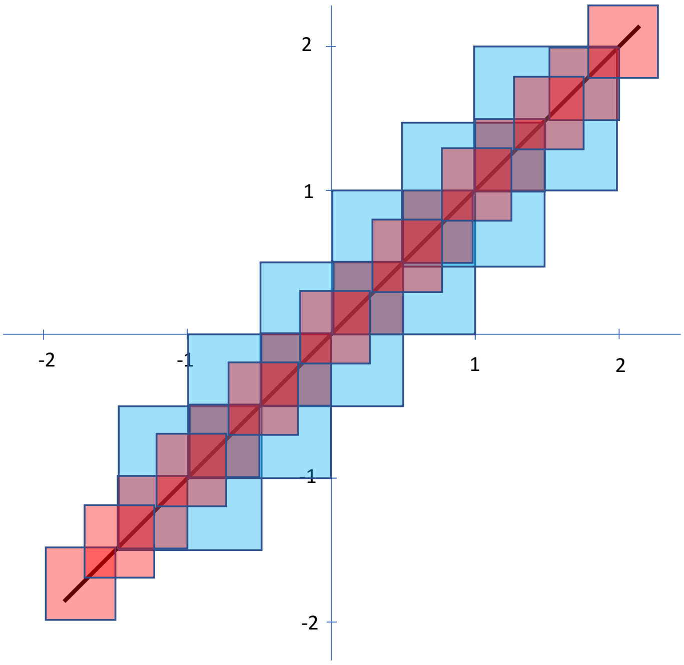


Figure 6.1: One way to see why the line of points (x,x) is in $\mathcal{B}(\mathbb{R}^2)$: the set of points (x,x) can be written as a countable intersection of sets T_n, where each T_n is in $\mathcal{B}(\mathbb{R}^2)$ because it is itself a countable union of sets $R(j \cdot 2^{-(n+1)}, 2^{-(n)})$ which are in $\mathcal{B}(\mathbb{R}^2)$ because they are open squares. The graphics shows some of the rectangles $R(j \cdot 2^{-1}, 2^0)$ contributing to T_1 (light blue) and some $R(j \cdot 2^{-2}, 2^{-1})$ contributing to T_2 (light red).

Like in the 1-dimensional case, there is a theorem that every topologically open and every closed set $A \subseteq \mathbb{R}^2$ is in $\mathcal{B}(\mathbb{R}^2)$. Since the line L of all (x,x) points is topologically closed, showing that $L \in \mathcal{B}(\mathbb{R}^2)$ could be done just by citing that theorem. But our grassroot derivation is more instructive.

Figure 6.2 shows some more sets that are in $\mathcal{B}(\mathbb{R}^2)$. It is, in fact, difficult (and would be beyond our means) to specify a set $A \subseteq \mathbb{R}^2$ that is NOT in $\mathcal{B}(\mathbb{R}^2)$. The powers of generating new sets by repeated countable unions and complements is beyond our imagination and renders the study of σ-fields a rather difficult enterprise. Usually this material is only taught to core mathematics students. I nonetheless wanted to afford you a glimpse, because it gives you a feeling of the wonders of probability.

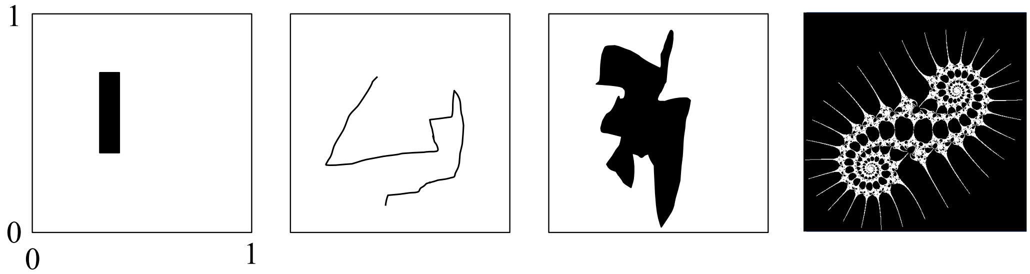


Figure 6.2: Some sets in the Borel σ-field on the unit square. From left to right: rectangles, lines, filled shapes, fractal sets. Image in last panel taken from `http://www.ijon.de/mathe/julia/some_julia_sets_4_en.html`.

If you want to dig a little deeper into σ-fields, I recommend the Wikepedia article `en.wikipedia.org/wiki/Sigma-algebra`.

Although σ-fields are very intricate objects, the reason why they arise inevitably and naturally in probability theory is simple:

- We want to be able to say, the probability that *anything* will happen is equal to 1. This "anything" is the event that includes all possible observation outcomes, that is, it is just S itself. This is condition 1 in Definition 6.2.1.

- We want to be able to speak of the probability that something does **not** happen. Since probabilities are assigned to events, which we defined through subsets of S, we have to deal with complements of sets - the second ingredient in the definition of σ-fields.

- Finally, we want to be able to speak of the probability that this **or** that can happen. Even more, we want to be able to speak of the probability that this **or** this **or** this **or** this ... may happen (we need countably infinite **or**'s because, for instance, we want to talk about the probability that an accident happens now **or** within a minute **or** within two minutes **or** ...). This leads to the third condition in the definition of σ-fields.

6.3 Measurable spaces and measurable functions

We round off our excursion into σ-fields with a few definitions and insights that are the "small coin" of texts on probability theory — these items are just taken for granted by anybody seriously talking probability.

When doing statistical modeling, the sample space S that we use is *always* equipped with a σ-field $\mathcal{F}$. This leads to a mathematical object known as a *measurable space*:

Definition 6.3.1 *A pair $(S, \mathcal{F})$, where S is a set and $\mathcal{F}$ is a σ-field over S, is called a* measurable space.

Often the choice of an appropriate σ-field is natural and the same σ-field is used in virtually all uses of a given sample space. For instance, the standard default σ-field for finite sample spaces S is the power set of S, and the default σ-field for real-valued sample spaces (subsets $S \subset \mathbb{R}^n$) is the Borel σ-field $\mathcal{B}(\mathbb{R}^n)$ which is generated by the open n-dimensional intervals in S. Since real-valued observation outcomes are ubiquitous, the Borel σ-fields $\mathcal{B}(\mathbb{R}^n)$ are a constant companion for statisticians.

When one has a measurable space $(S', \mathcal{F}')$, a set S and a function $\varphi : S \to S'$, this function creates a σ-field on S (compare Definition 6.2.2):

Definition 6.3.2 *Let $\varphi : S \to (S', \mathcal{F}')$ be a function from some set S to the measurable space $(S', \mathcal{F}')$. Let $\varphi^{-1}(\mathcal{F}') = \{\varphi^{-1}(A) \subseteq S \mid A \in \mathcal{F}'\}$ be the set of all pre-images of φ of elements of $\mathcal{F}'$. The σ-field $\sigma(\varphi^{-1}(\mathcal{F}'))$ on S is called the σ-field* induced *on S by φ.*

In fact, the σ-field induced by φ just is the collection of all pre-images of φ:

Proposition 6.3.1

$$\sigma(\varphi^{-1}(\mathcal{F}')) = \varphi^{-1}(\mathcal{F}').$$

The proof is left as a homework exercise.

Functions between measurable spaces abound in probability theory — after all, random variables and transformations of RVs are functions between measurable spaces. Such functions are always required to satisfy a certain soundness condition — they must be *measurable*:

Definition 6.3.3 *Let $(S, \mathcal{F}), (S', \mathcal{F}')$ be two measurable spaces and let $\varphi : S \to S'$ be a function. Then φ is said to be $\mathcal{F} - \mathcal{F}'$-*measurable *if for all $A' \in \mathcal{F}'$, $\varphi^{-1}(A') = \{a \in A \mid \varphi(a) \in A'\}$ is in $\mathcal{F}$.*

Another, equivalent way to characterize measurability of a function $\varphi : (S, \mathcal{F}) \to (S', \mathcal{F}')$ is to require that $\varphi^{-1}(\mathcal{F}') \subseteq \mathcal{F}$.

Notice the similarity of the concept "measurable function" with the concept of a "continuous function" used in calculus and topology. By its general definition in topology, a function φ from a topological space A to a topological space B is said to be continuous if the pre-image of any open set in B is an open set in A. The elementary calculus textbook definition of a continuous function $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is a special case of this general topological definition (recall that textbook definition: $\varphi : \mathbb{R}^n \to \mathbb{R}^m$ is continuous in $a \in \mathbb{R}^n$ if for all $\delta > 0$ there exists an $\varepsilon > 0$ such that if $\|a - a'\| < \varepsilon$, then $\|\varphi(a) - \varphi(a')\| < \delta$).

The most common use of all of these concepts occurs with random variables. When one has a sample space S equipped with a σ-field $\mathcal{F}$, and a RV $X : \Omega \to S$, this RV induces a σ-field $X^{-1}(\mathcal{F})$ on Ω. In intuitive terms: a RV "back-projects" the structure (= the σ-field) of a sample space into the universe. In our initial, private terminology: the structure we can see in a RSOI is the the structure of our observation spaces (= $(S, \mathcal{F})$), seen "backwards" through our observation procedures.

6.4 Interim summary 2: The big picture again, enriched with events

Let us assemble what we have got so far. Our mathematical model of a RSOI, OP, DVS now consists of Ω (abstract model of RSOI), a family $(X_i)_{i \in I}$ of RVs (abstract model of the OP), and an associated family $((S_i, \mathcal{F}_i))_{i \in I}$ of measurable spaces (models of the DVS subspaces that we use). Each set $A \in \mathcal{F}_i$ induces an event $X_i(A) \subseteq \Omega$.

Figure 6.3 illustrates this situation. You see that by virtue of the involved RVs, the universe Ω becomes populated by numerous events. In fact the schematic illustration in Figure 6.3 gives a much too weak impression of the richness of events that are induced in the universe by RVs. I invite you to unleash your powers of imagination, and contemplate the following, virtually mind-stunning points:

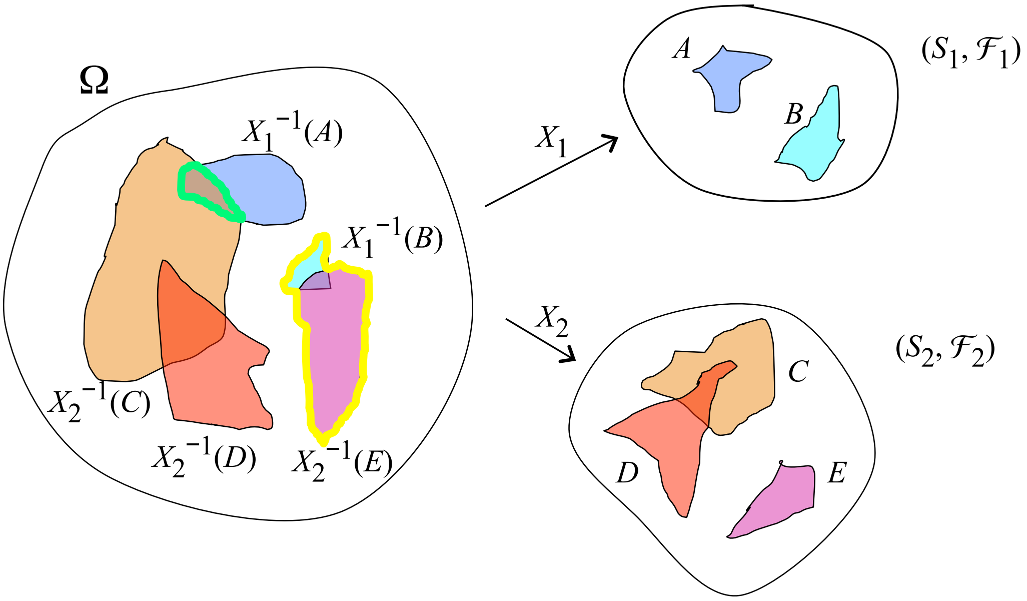


Figure 6.3: Two RVs X_1, X_2 defined on a universe Ω (schematic). A, B are some elements of $\mathcal{F}_1$, C, D, E some elements of $\mathcal{F}_2$. In Ω some events induced by X_1, X_2 are shown. The set highlighted by a bright green boundary is the event $X_1^{-1}(A) \cap X_2^{-1}(C)$ – that is, the event "X_1 takes value in A *and* X_2 takes value in C". The event highlighted by a yellow boundary is $X_1^{-1}(B) \cup X_2^{-1}(E)$ – that is, "'X_1 takes value in B *or* X_2 takes value in E"

- When some sample space S_i is continuous and equipped with the Borel σ-field $\mathcal{B}_i$, this σ-field $\mathcal{B}_i$ contains uncountably many sets. Showing only two or three of them the figure therefore does not give credit to the enormous number of sets that are often admissible in a sample space. And each of these sets $A \subseteq S_i$ leads to an event $X_i^{-1}(A)$ that adds to the event population in Ω.

- Many statistical models involve infinitely many RVs — even uncountably many. This occurs, for instance, always when open-future temporal developments of a random system are modeled by a stochastic process. Each of these RVs X_i may have an uncountable σ-field $\mathcal{F}_i$. And all of the elements A of all of these $\mathcal{F}_i$ lead to events in Ω!

- Events can be defined by combining the workings of several RVs. In Figure 6.3 two simple instances are indicated. The set $X_1^{-1}(A) \cap X_2^{-1}(C)$ is the event that could be paraphrased as "all occurances ω where measurement X_1 gives an outcome in A **and** X_2 a value in C" (set intersection corresponds to Boolean **and**). Similarly, $X_1^{-1}(B) \cup X_2^{-1}(E)$ is an event defined by combining outcomes from X_1, X_2 with logical **or**. Furthermore, the set complement operation includes the logical **not** into the picture (not shown in the figure).

- In probability theory one generally admits countably infinite event unions (**or**) and intersections (**and**). These arise more often than you would guess from the simple Figure 6.3. A strong motif for including such infinite combinations is that this enables a formal treatment of probabilistic statements which include limits, as in "as time progresses, the probability that the bank ends in bankruptcy approaches 1".

In spite of this complexity, there is an intuitive way to capture the essence of all of the events in Ω:

The grand intuition: an event is any subset of the universe that can be characterized by combining information about observation outcomes.

"Combining information" here means to use the logical **or**, **and**, and **not** operations, even (countably) infinitely often.

Finally, consider again a family of RVs $(X_i)_{i\in I}$, where $X_i : \Omega \to (S_i, \mathcal{F}_i)$. The set of all events that are induced by the family of RVs $(X_i)_{i\in I}$, that is the set $\bigcup_{i\in I} X_i^{-1}(\mathcal{F}_i)$, generates a σ-field $\sigma\left(\bigcup_{i\in I} X_i^{-1}(\mathcal{F}_i)\right)$ on Ω. This σ-field is denoted by $\mathfrak{A}((X_i)_{i\in I})$, or simply by $\mathfrak{A}$ when the RVs are clear from context. $\mathfrak{A}((X_i)_{i\in I})$ is the smallest σ-field on Ω which makes all the $(X_i)_{i\in I}$ measurable.

Thus we conclude this section with the following abstraction of the big picture, enriching our initial account from Equation 2.1:

$$\boxed{X_i : (\Omega, \mathfrak{A}) \to (S_i, \mathcal{F}_i) \qquad (6.3)}$$

Chapter 7

Finally... calling probability on stage!

Now (after about 40 pages...) we have set up the scene such that we can usher in her majesty, PROBABILITY, and install her on her throne. This throne is $\mathfrak{A}$, the σ-field on Ω which contains all the events we ever can identify through observations. We assign to each such event $E \in \mathfrak{A}$ a *probability*, that is a number between 0 and 1, written $P(E)$.

7.1 Probability spaces and THE probability space

At this point I will become more general than necessary for our purposes. In mathematical probability theory, "probability" is introduced in a very abstract way, without reference to statistical modeling of real-world phenomena. I present that abstract definition, and will show how it is to be applied in our modeling scenario, after giving the formal definition.

Definition 7.1.1 *Let M be a non-empty set and $\mathcal{F} \subseteq \mathbf{Pot}(M)$ a σ-field on M, that is, $(M, \mathcal{F})$ is a measurable space. A* probability measure *on $(M, \mathcal{F})$ is a function $P : \mathcal{F} \to [0, 1]$ which assigns to each set $A \in \mathcal{F}$ a value $0 \leq P(A) \leq 1$, subject to the following conditions:*

1. *$P(M) = 1$.*

2. *For all sequences $A_1, A_2, \ldots$ of pairwise disjoint elements of $\mathcal{F}$:*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{7.1}$$

*The second condition is called σ-*additivity*. The triple $(M, \mathcal{F}, P)$ is called a* probability space*.*

The conditions stated in Definition 7.1.1 have some immediate consequences (easy exercise to prove them):

$$\begin{array}{rcl} P(\emptyset) & = & 0 \\ P(A^{\mathsf{c}}) & = & 1 - P(A) \\ A \subseteq A' & \Rightarrow & P(A) \leq P(A') \end{array}$$

For us and our goal of formalizing statistical models of real-world phenomena, one specific probability space is key. Namely, we endow the universe Ω and its σ-field $\mathfrak{A}$ — which is the core of our grand intuition — with a probability measure, obtaining

$$\boxed{(\Omega, \mathfrak{A}, P)}$$

as our one-and-only fundamental probability space.

An almost philosophical comment is in place. In the picture that I have presented here, probability "lives" in the universe Ω. The universe Ω is the abstract model of a reality segment of interest (RSOI), a piece of reality. *We thus pictured probability to be a property of the real world.* According to this way of modeling random systems, probability is a "real", "physical" property of certain events to occur with certain probabilities. For instance, it is a physical property of a loaded die to come up with a facing "6" with a probability of $P(X = 6) = 1/5$. This is not the only way how philosophers — and engineers and other down-to-earth people — have conceived of probability. Specifically, one can also start from a notion of probability as a subjective degree of belief ("I believe it will rain tomorrow with a probability of more than one half"). This is the approach of *Bayesian* statistics. Chapter 15 says a little more about Bayesian statistics. The formal tools forged by abstract mathematicians (σ-fields, measurable spaces, random variables, probability spaces) are used by Bayesian statisticians, too — but there these formal constructs are interpreted with different background intuitions, and are mapped to reality in a very different way.

One could call the interpretation of probability that we embraced, an "empirical", "realistic", or "objectivistic" view, in contrast to the "subjectivistic" view of Bayesians. For the latter, probability is not a physical property of real-world objects, but a subjective "degree of belief".

The abstract tools for modeling probability (σ-fields, measurable spaces, random variables, probability spaces) were moulded into their present modern form by the Russian mathematician Andrey Nikolaevich Kolmogorov in a book [Kolmogorov, 1956] first published 1933 in German.

7.2 Notation: the basic format of probability statements

Different authors and textbooks use different notation styles for probability formulas, and it is not easy for a student to discern the invariant core behind these variable conventions. In turn, students often have serious difficulties to just write down formulas involving probability in a clean and consistent way. Now I will declare the notation that I will be using in this course. It is the rigorous and unambiguous notation used by mathematicians. In Chapter 14 I comment on other notation styles.

First thing: when I use the symbol P, it always means the probability measure that is declared on the fundametal probability space $(\Omega, \mathfrak{A}, P)$, which in turn is our model of a data-generating environment, a reality segment of interest.

P is a function with domain $\mathfrak{A}$ and codomain $[0, 1]$. The fundamental way of using the symbol P is therefore to employ it in formulas of the form

$$P(E) = p,$$

where $E \in \mathfrak{A}$ is an event and p is a real number from the interval $[0, 1]$.

I explained earlier that the σ-field $\mathfrak{A}$ should be thought of as being induced by the random variables that one wants to include in a probabilistic model, that is, $\mathfrak{A} = \mathfrak{A}((X_i)_{i \in I})$. All events $E \in \mathfrak{A}$ that we will practically consider are specified in terms of the outcomes of RVs. In the simplest case, only a single RV $X : \Omega \to (S, \mathcal{F})$ is involved in specifying E, via $E = X^{-1}(A)$, where $A \subseteq S$ is an element of a suitably chosen σ-field $\mathcal{F}$ on S. A common alternative notation for $X^{-1}(A)$ is $X \in A$. This leads to the following notation for such events E:

$$P(X^{-1}(A)) = p \text{ or } P(X \in A) = p, \tag{7.2}$$

where the second notation is used more often than the first.

More complex events in $\mathfrak{A}$ are induced by combining outcome information from several RVs. Consider the case of two RVs $X_i : \Omega \to (S_i, \mathcal{F}_i)$ where $i = 1, 2$. For $A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$ we get two events $X_1^{-1}(A_1), X_2^{-1}(A_2) \in \mathfrak{A}$. Since the σ-field $\mathfrak{A}$ is closed under intersection, also $X_1^{-1}(A_1) \cap X_2^{-1}(A_2) \in \mathfrak{A}$. In words, we are dealing with the event "X_1 returns an outcome in A_1 **and** X_2 returns an outcome in A_2." Therefore it is correct to write $P(X_1^{-1}(A_1) \cap X_2^{-1}(A_2)) = p$, but this is commonly written instead as

$$P(X_1 \in A_1, X_2 \in A_2) = p, \tag{7.3}$$

and it is called the *joint probability* that X_1 takes value in A_1 and X_2 takes value in A_2. This extends in an obvious way to cases of the joint probability of a finite family $X_1, \ldots, X_N$ of RVs.

A particular kind of event occurs when $A = \{a\}$, where $a \in S$, is a singleton. Then one writes $P(X = a)$ instead of $P(X \in \{a\})$.

Remark: when S is a continuous space, for example $S = \mathbb{R}$, then usually $P(X = a) = 0$. For example, the chance that the speed of light is *exactly* equal to 299792458.31415926535897932384626433832 is zero (after the decimal dot I started to list the digits of π, for best effect please continue reading all the decimals beyond the page margin). This happens in all cases where the probability can be described by a pdf. In contrast, in finite sample spaces one typically has nonzero probabilities for exact-value events. For example, assuming a fair die in Die throwing, $P(X = 1) = 1/6$.

7.3 Conditional probability

Assume $X : \Omega \to (S, \mathcal{F}),\ Y : \Omega \to (S', \mathcal{F}'),\ A \in \mathcal{F},\ B \in \mathcal{F}',\ P(Y \in B) > 0$. One *defines*

$$P(X \in A \mid Y \in B) := \frac{P(X \in A, Y \in B)}{P(Y \in B)} \tag{7.4}$$

and calls $P(X \in A \mid Y \in B)$ the *conditional probability* of the event $X \in A$ given $Y \in B$. Figure 7.1 gives a graphical illustration. The intuition behind this concept of conditional probability is that one wishes to restrict the original data generating environment to those elementary events that have an outcome in B when observed by Y. This new, restricted universe is $\Omega' = Y^{-1}(B)$. The σ-field $\mathfrak{A}'$ on Ω' is derived from the original σ-field $\mathfrak{A}$ on Ω by $\mathfrak{A}' = \{Y^{-1}(B) \cap E \mid E \in \mathfrak{A}\}$. On this new σ-field $\mathfrak{A}'$ a new probability measure P' is given by

$$P'(Y^{-1}(B) \cap E) := \frac{P(Y^{-1}(B) \cap E)}{P(Y \in B)}, \tag{7.5}$$

which generalizes (7.4) to arbitrary events in the restricted σ-field $\mathfrak{A}'$. The probability measure P' is also written as $P_{Y \in B}$. The triple $(\Omega', \mathfrak{A}', P') = (Y^{-1}(B), \mathfrak{A}', P_{Y \in B})$ is again a probability space.

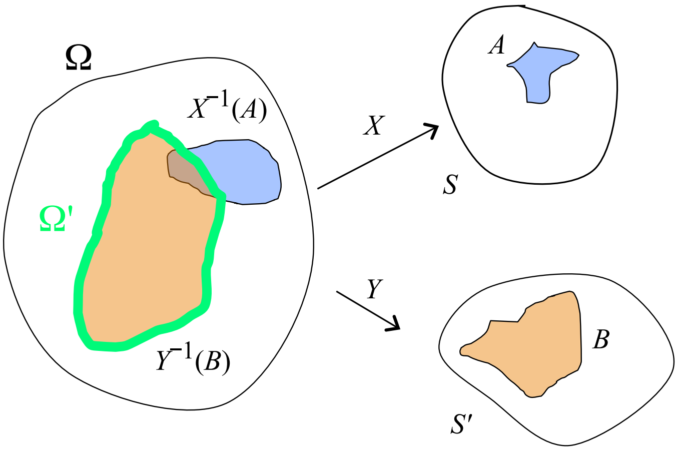


Figure 7.1: Illustrating conditional probability. Compare text.

It is allowed that $X = Y$. For a simple example, consider in Die throwing the RV X which counts dots. Let us assume a fair die. Then the probability to get a "1" dot count given that one already knows the dot count is uneven is $P(X = 1 \mid X \in \{1, 3, 5\}) = 1/3$.

Conditional probabilities arise almost everywhere in statistical analyses, but they are *always* present in stochastic process investigations. In fact, conditional probabilities are the very core of modeling random temporal systems. The eternal question, "what will happen next?", essentially is a conditional probability question. Consider the Sentence translation example, where $(X_n)_{n\in\mathbb{N}}$ are the RVs that yield the n-th letters in typed sentences. In order to predict the next letter in a sentence, it is necessary to know the beginning of the sentence up to that point. A standard way to characterize the probabilities that rule such temporal sequences is to specify the conditional probabilities for the next timestep observation, given the observations made up to that time. That is, a statistical model would specify in some way or other all the *transition probabilities*

$$\{P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n) \mid n \in \mathbb{N}, (x_0, \ldots, x_{n+1}) \in S^{n+2}\}.$$

Conditional probabilities are also the key for *classification* algorithms, which in turn are one of the core objects of interest in machine learning. A classification algorithm (or just "a classifier") is an algorithm which gets a *pattern* as input and returns a *class label* as output. For instance, a handwritten digit classifier would get pictures of handwritten digits as input and would output one of the 10 possible class labels "0", "1", ..., "9"; or a single-word speech recognition algorithm would receive a short sound recording as input and would output a typed word. To be good at their task, such classifier algorithms must internally compute conditional probabilities of the kind $P(\mathsf{ClassLabel} = c \mid \mathsf{InputPattern} = u)$, where $\mathsf{ClassLabel}$, $\mathsf{InputPattern}$ are random variables. The companion lecture "Machine Learning" treats this situation in some detail.

7.4 Bayes' formula

From Equation 7.4 it is straightforward to derive

$$P(X \in A \mid Y \in B) = \frac{P(Y \in B \mid X \in A)\, P(X \in A)}{P(Y \in B)}, \tag{7.6}$$

which is (one version of) *Bayes' formula.* Bayes' formula allows one to reverse the direction of conditioning. This is central in *diagnostic reasoning.* For an instructive example, interpret the event "$X \in A$" as "patient ω has cancer", and "$Y \in B$" as "blood test of patient ω gave high leucocyte count". Then $P(X \in A \mid Y \in B)$ is the probability that a patient has cancer, given that the blood test yielded this high leucocyte (white blood cells) count. This is clearly a probability of great interest to the patient. However, clinical studies of cancer patients yield the reverse kind of information, namely estimates of probabilities $P(Y \in B \mid X \in A)$ that a blood test has this-and-that outcome, given that the patient has cancer. The probability $P(X \in A)$ is the base rate probability of a person drawn at random from the population having cancer, and $P(Y \in B)$ is the base population probability of a person having low leucocyte counts. These numbers can be estimated from medical population screenings. Plugging these three items into (7.6) gives the crucial diagnostic information of the probability that the patient has cancer, given the blood test result.

The base-rate probability $P(X \in A)$ is called the *prior probability* or simply *the prior.* The conditional probability $P(X \in A \mid Y \in B)$ is called the *posterior* probability of X taking value in A. The prior probability measures the general knowledge about the occurrence rate of an event $X \in A$ in the population, before any symptoms Y have been observed. The the posterior probability is an "update" of that prior probability after the *evidence* $Y \in B$ has been factored in.

More generally speaking, Bayes' formula makes it possible to infer back from symptoms to causes. Read the left-hand side of (7.6) as "the probability of this underlying cause A, given that we have observed a symptom B". The core term $P(Y \in B \mid X \in A)$ in the right-hand side of (7.6) is "the probability that if this cause A is present, symptom B is observed". Such conditional probabilities of "effects given causes" are the essence of *causal* models of parts of reality. Causal models are what scientists are after, and experimental investigations in the sciences typically lead to causal models. Diagnostic reasoning is "causation explored backwards" from effects to causes, and Bayes' formula is the key to exploit available causal models for diagnostic reasoning.

A terminology confusion alert: Bayes' formula is just a little, super-practical formula that is used in all branches and corners of ML and statistics. In its basic form it is *not* connected to "Bayesian statistics". Bayesian statistics is a particular approach to statistical modeling whose conceptual roots lie in subjectivist conceptions of statistics. We will take a closer look at this approach in Section 15.

7.5 Where does probability come from, and how can we measure it?

So far we have described only how to write down probability statements in a clean notation. But what does "$P(X \in A) = p$" *mean*? What statement about reality is made with such a formula?

This is a difficult, ultimately philosophical question and there is no generally accepted unique answer. Philosophers, scientists and mathematicians have proposed a variety of answers. These answers can be broadly grouped into *objectivistic* and *subjectivistic* interpretations of "probability". According to the objectivistic view, probability resides physically in the real world — it is a phenomenon of nature, as fundamentally a property of physical reality as for instance "time" or "energy". According to the subjectivistic view, probability describes an observer's subjective opinion on something observed — a degree of belief, of uncertainty, of plausibility of judgement, or missing information etc.

The objectivistic view is adopted by virtually all textbooks of statistics and mathematical probability theory, it is the way how natural scientists look at randomness, and it is the view adopted in this tutorial too.

When probability is regarded as a phenomenon of nature, there should be ways to *measure* it. The standard proposition of how one can measure probability is by *relative frequency counting.* For instance, in a Die Throwing scenario, a loaded die may have the physical property that $P(X = 6) = 1/5$ (as opposed to $P(X = 6) = 1/6$ for a fair die). This property of the die could be experimentally measured by *repeating* the die-throwing act many times. This would give a sequence of outcomes $X(\omega_1), X(\omega_2), X(\omega_3), \ldots = x_1, x_2, x_3, \ldots$. After N such throws, an estimate of the quantity $P(X = 6)$ is calculated by

$$\hat{P}_N(X = 6) = \frac{\text{number of outcomes } x_i = 6}{N}, \tag{7.7}$$

where N is the number of throws that the experimentalist measuring this probability has carried out.

We generally use the angular hat symbol $\hat{\cdot}$ on top of some variable to denote a numerical estimate based on limited observation data. The true probability $P(X = 6) = 1/5$ would then become "measurable in principle" by

$$(^*) \qquad P(X = 6) = \lim_{N \to \infty} \hat{P}_N(X = 6). \tag{7.8}$$

Equation (7.8) embodies what is known as the *frequentist* interpretation of probability:

The frequentist view of probability: The probability of an event $X \in A$ is the relative frequency by which repeated measurement acts $X(\omega_n)$, where $n = 1, 2, 3, \ldots, \infty$ lead to an outcome in A.

If one looks at this "definition" critically one will find it is loaden with difficulties.

First, it defines a "measurement" process that is not feasible in reality because one cannot physically carry out infinitely many observation acts $X(\omega_n)$. This is maybe not really disturbing because *any* measurement in the sciences (say, of a voltage) is imprecise and one gets measurements of increasing precision by repeating the measurement, just as when one measures a probability.

Second, it does not inform us about how, exactly, the elementary events ω_n are "chosen". The events ω_n should be picked from Ω absolutely "at random" — but what does that mean in terms of experimental procedures? This is a very critical issue. To appreciate its impact, consider our the Credit Risk example. The bank can only use customer data collected at elementary events ω_n in the *past*, but wants to base creditworthyness decisions for *future* customers on those data. The reality segment of interest, modeled by Ω, thus rightfully comprised both past and future customers. Picking only past $\omega \in \Omega$ to base probability estimates on hardly can qualify as an "absolutely random" picking of elementary events, and in fact the bank may grossly miscalculate credit risks when the general customer body or their economical conditions change over time. These difficulties have of course been recognized in practical applications of statistics, and textbooks and courses contain instructions on how to create "random samples" or "unbiased samples" as well as possible.

Third, if one repeats the repeated measurement, say by carrying out one measurement sequence based on $\omega_1, \omega_2, \ldots$ and another one based on $\omega'_1, \omega'_2, \ldots$, the values $\hat{P}_N$ from

Equation (7.7) are bound to differ between the two series. The limit indicated in Equation (7.8) must somehow be robust against different versions of the $\hat{P}_N$. Mathematical probability theory offers several ways to rigorously define limits of series of probability quantities which we do not present here. Equation (7.8) is suggestive only and I marked it with a (*) to indicate that it is not technically correct and complete.

Among these three difficulties, only the second one is really problematic. The first one is just a warning that in order to measure a probability with increasing precision we need to invest an increasingly large effort, — but that is the same for other measurables in the sciences. The third difficulty can be fully solved by a careful definition of suitable limit concepts in mathematical probability theory. But the second difficulty is fundamental and raises its ugly head whenever statistical assertions about reality are made.

In spite of these difficulties, the objectivist view on probability in general and the the frequentist account of how to measure it in particular is widely shared among empirical scientists.

In Chapter 15 I will comment a little more on subjectivistic views.

Chapter 8

Samples and estimators

In statistics as well as in machine learning, one attempts to make assertions about "true" probabilities $P(X \in A)$ based on a finite collection of empirically observed data values $(X(\omega_1), \ldots, X(\omega_N)) = (x_1, \ldots, x_N) \in S^N$. Such finite collections of observed data values are generally called *samples.* But this term is mostly used in a rather loose way and there are two non-equivalent ways to make it precise.

Unfortunately there seem to be no separate words for the two definitions of "samples". In a German handbook of statistics [Müller, 1983] I found the (tentatively assigned) wordings "simple sample" (*einfache Stichprobe*) and "mathematical sample" (*mathematische Stichprobe*), which I will use here, but this is not a generally adopted terminology.

8.1 Simple samples

Let $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ be a random variable with its underlying probability and sample spaces. A (simple) sample of size N is the vector of N data points $(X(\omega_1), \ldots, X(\omega_N)) = (x_1, \ldots, x_N) \in S^N$ obtained from N elementary events $\omega_1, \ldots, \omega_N$. Notes:

- The notion of a (simple) sample does not specify how the elementary events ω_n are chosen. They may be chosen "at random" in which case one also speaks of a *representative sample* or a *random sample*, but there may also be some systematic distortion factor in the choice of the elementary events, then one speaks of a *biased sample* or *unrepresentative sample.* If one knows the nature of the systematic distortion one may attempt in practice to "normalize" the sample in some way to make it approximately random.

- The values x_n in the data vector $(x_1, \ldots, x_N)$ may be complex mathematical objects, for instance labelled trees in our Evolutionary tree scenarios.

- I will use the word *data point* for any such single value x_n.

- In the context of machine learning, the sample $(x_1, \ldots, x_N)$ is often called *training data.* The idea here is that training data are used as input for a "learning" algorithm which distils some model of reality from these data.

- Occasionally one finds that an author also calls an individual data point $X(\omega_n)$ a "sample".

8.2 Mathematical samples

The other conception of "samples" is more complex but also more insightful than the notion of simple samples. A good starting point for an explanation is the Speed-of-Light II scenario. Recall that in this scenario we considered multiple labs (existing and hypothetical ones) where series of repeated measurements were carried out. An elementary event ω gave rise to such an entire series of measurements, a circumstance that we formally modeled in Section 3.1 by random variables X_n, where $X_n(\omega)$ represented the n-th measurement in a series. When such a series comprises N measurements (being stopped after N repetitions), the obtained values $(X_1(\omega), \ldots, X_N(\omega)) = (x_1, \ldots, x_N)$ are again naturally called a sample.

To cast this in a more general frame, a (mathematical) sample of size N builds on a sequence $X_1, \ldots, X_N$ of RVs, all taking value in the same sample space S, such that each elementary event ω yields a vector of data points $(X_1(\omega), \ldots, X_N(\omega)) \in S^N$. One usually requires that the X_n are *identically and independently distributed* — for short, they are "i.i.d.":

Definition 8.2.1 *1. Two RVs* $X, Y : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ *are* identically distributed *if* $\forall A \in \mathcal{F} : P(X \in A) = P(Y \in A)$.

2. A finite family $(X_n)_{n=1,\ldots,N}$ *of RVs, where* $X_n : (\Omega, \mathfrak{A}, P) \to (S_n, \mathcal{F}_n)$*, is* independently distributed *if for all* $A_1 \in \mathcal{F}_1, \ldots, A_N \in \mathcal{F}_N$*, for all* $1 \leq i \leq N$*:*

$$\begin{aligned} P(X_i \in A_i \mid X_1 \in A_1, \ldots, X_{i-1} \in A_{i-1}, X_{i+1} \in A_{i+1}, \ldots, X_N \in A_N) = \\ = P(X_i \in A_i) \end{aligned} \qquad (8.1)$$

The second condition of independent distribution captures the intuition that in a series of repeated measurements, each measurement should be completely decoupled from the others, with no transfer of information whatsoever between the individual measurements. Notes:

- X, Y being identically distributed does *not* mean $\forall \omega \in \Omega : X(\omega) = Y(\omega)$.
- The standard textbook definition of $(X_n)_{n=1,\ldots,N}$ being independently distributed differs from ours, but is equivalent. It goes like this: $(X_n)_{n=1,\ldots,N}$ are independently distributed if for all $A_1 \in \mathcal{F}_1, \ldots, A_N \in \mathcal{F}_N$:

$$P(X_1 \in A_1, \ldots, X_N \in A_N) = P(X_1 \in A_1) \cdot \ldots \cdot P(X_N \in A_N). \qquad (8.2)$$

- For independence it is not enough to require that the RVs X_n are *pairwise* independent (one finds this erroneous specification sometimes in the literature). For an illustration that one has to prohibit all interactions simultaneously consider a die throwing scenario with a fair die with three observations $X_1, X_2, X_3 : \Omega \to \{0, 1\}$, where an elementary event ω means to throw the die twice, specified by
  - $X_1(\omega) = 1$ if the outcome of the first throw is uneven,
  - $X_2(\omega) = 1$ if the outcome of the second throw is uneven,
  - $X_3(\omega) = 1$ if the sum of outcomes of both throws is uneven.

  It is easy to see (see it!) that these three RVs are pairwise independent, but not jointly in the sense of (8.1) or (8.2). This example is taken from Bauer [1978].

The exact usage of the word "sample" in this second meaning is not uniform in the literature. Sometimes the (multi-)set or vector of data points $(X_1(\omega), \ldots, X_N(\omega))$ is called "sample", sometimes the vector $(X_1, \ldots, X_N)$ or its product $X = X_1 \otimes \ldots \otimes X_N$ is called "sample", or the elementary events ω are called "samples".

8.3 Estimators

The great advantage of the notion of mathematical samples, as opposed to what we called simple samples above, lies in the fact that a sample $X = X_1 \otimes \ldots \otimes X_N : \Omega \to S^N$ is itself a RV. This insight is the starting point for an important subfield of statistics called *estimation theory.* Abstractly speaking, an *estimator* is a function φ of a (mathematical) sample $X = X_1 \otimes \ldots \otimes X_N : \Omega \to S^N$. Usually an estimator is devised to yield an estimate of a statistical characteristic of the probability space $(\Omega, \mathfrak{A}, P)$.

One of the most frequently encountered estimators is the *sample mean.* It is defined when the sample space of each X_n is $S = \mathbb{R}^n$, i.e. when we are dealing with numerical measurements. The sample mean simply computes the average of the sample data points. When $X = X_1 \otimes \ldots \otimes X_N : \Omega \to (\mathbb{R}^n)^N$ is a sample of size N, the sample mean is the function

$$\begin{array}{cccc} \varphi : & (\mathbb{R}^n)^N & \to & \mathbb{R}^n \\ & (x_1, \ldots, x_N) & \mapsto & \frac{1}{N}(x_1 + \ldots + x_N). \end{array} \tag{8.3}$$

The function $\varphi \circ X : \Omega \to \mathbb{R}^n$ is a RV. Its values can be interpreted as data-based *estimates* of the true expected (average) value of the measurements X_n in the universe, the average taken over all $\omega \in \Omega$. These estimates depend on the outcome (values $X_1(\omega), \ldots, X_N(\omega)$) of the elementary events ω. The function φ is an *estimator* for the true population mean.

Seen abstractly, an estimator is just a function that for its input takes the values of a (mathematical) sample, and returns a number. Such functions also occur outside estimation theory and in general are called *statistics.* Thus, in general, if $X = X_1 \otimes \ldots \otimes X_N : \Omega \to (S^N, \mathcal{F}^N)$ is a size-N sample, and $\varphi : S^N \to \mathbb{R}^n$ is any $\mathcal{F}^N - \mathfrak{B}(\mathbb{R}^n)$-measurable function, then

$$\varphi \circ X : \Omega \to \mathbb{R}^n$$

is called a statistic. A statistic is itself a random variable; what makes it special is that it piggybacks on another RV which describes a (mathematical) sample.

For a given characteristic θ of a population, there usually exist many estimators. They will have different qualities. Some of them may be easier to compute than others, some of them may lead to more accurate estimates $\hat{\theta}$ than others, etc. The theory of estimators has developed systematic ways to define and compare the relative merits of different estimators for the same population characteristic. In part II of this lecture we will take a more in-depth look at estimator theory, since it is a core part of the field of statistics.

To round off this very preliminary excursion into the world of estimators, consider another estimator for the population mean of a numerical measurement:

$$\begin{array}{cccc} \varphi' : & (\mathbb{R}^n)^N & \to & \mathbb{R}^n \\ & (x_1, \ldots, x_N) & \mapsto & x_1. \end{array}$$

This "estimator" simply returns the value of the first measurement in a series. It is obviously cheaper to compute than the sample mean estimator, and obviously not as good in the sense that it ignores valuable information. But *on average* (across all $\omega \in \Omega$), this cheap estimator φ' will give the same result as the sample mean.

To make another leap into abstraction: every machine learning algorithm turns sample values (called "training data" then) into some "model". The models in machine learning are usually far more complex than simple average values of population measurables. But such "models" can still be regarded as data points in formal "model spaces", and a "learning algorithm" can abstractly be regarded as an estimator for an assumed "true" model. This renders estimation theory a fundamental tool for the analysis of machine learning algorithms, making it possible to specify and quantify the relative merits of different machine learning algorithms that are proposed for the same learning task.

Chapter 9

Distributions

I have remarked (emphatically) that the one-and-only fundamental source of probability is the probability space $(\Omega, \mathfrak{A}((X_i)_{i\in I}), P)$. Making this formal object the anchor of everything else reflects the objectivistic view on probability which places it into the real world (of which $(\Omega, \mathfrak{A}, P)$ is the formal model). I promised that when I write the symbol P it always will refer to this probability measure defined on $(\Omega, \mathfrak{A}((X_i)_{i\in I}))$.

In practical computations however you will find that you don't actually have a data structure $(\Omega, \mathfrak{A})$ available on your computer. The universe Ω is a very hypothetical set. It is useful for mathematicians to develop their abstract mathematical probability theory. Also $(\Omega, \mathfrak{A}((X_i)_{i\in I}))$ is good as a basis for intuitive or even philosophical thinking, because it formalizes "where probability comes from" — but it is unfit for doing practical calculations with.

A practician who wants to calculate and get concrete numbers as results needs something more effectively manageable than $(\Omega, \mathfrak{A}((X_i)_{i\in I}))$ and P. Such concrete objects that can be represented in computer data structures and can be manipulated with algorithms indeed exist: the measurable sample spaces $(S_i, \mathcal{F}_i)$. The objects collected in a sample space, which I called data points, are of the familiar kinds that computer scientists know: alphanumeric symbols, integers, Booleans, reals (approximately represented on the computer as floats of a certain precision), vectors, arrays, database records, trees, graphs, or any other data structures. If we can pull probability into the sample spaces, we can start doing calculations.

9.1 Distributions: formal definition and notation

Given a RV $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ we now *define* a probability measure on the measurable space $(S, \mathcal{F})$. Since this probability measure comes into existence through the random variable X we denote it by P_X:

Definition 9.1.1 *Let $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ be a random variable. The function $P_X : \mathcal{F} \to [0, 1]$ defined by*

$$\forall A \in \mathcal{F} : P_X(A) = P(X \in A) \tag{9.1}$$

is called the distribution *of X.*

It is easy to verify (do it) that P_X is a probability measure on $(S, \mathcal{F})$ according to Definition 7.1.1. The triple $(S, \mathcal{F}, P_X)$ is a probability space. Notes:

- A "distribution" is always a distribution *of a random variable.* Whenever a distribution is mentioned, there *must* be a RV behind it, even when it is not explicitly mentioned.

- Some authors express this by saying, "a RV X *transports* the underlying probability measure P to the sample space."

- By an abuse of terminology, also the elements $A \in \mathcal{F}$ are called "events".

- In most modeling situations one uses a multitude of RVs $X_i : (\Omega, \mathfrak{A}, P) \to (S_i, \mathcal{F}_i)$, where $i \in I$ for some index set I. Each X_i comes with its own distribution $P_{X_i} : \mathcal{F}_i \to [0,1]$. However, statisticians are not really interested in these individual distributions P_{X_i}. The main object of interest is the *joint distribution* of all the X_i together. This is the distribution of the product random variable $\bigotimes_{i \in I} X_i : (\Omega, \mathfrak{A}, P) \to (\prod_{i \in I} S_i, \bigotimes_{i \in I} \mathcal{F}_i)$. That is, statistical analyses revolve around $P_{\bigotimes_{i \in I} X_i}$. (Remark: here I wrote down a generic product of σ-fields, namely $\bigotimes_{i \in I} \mathcal{F}_i$. In Section 6.2 I briefly introduced finite products of σ-fields. Defining arbitrary products of σ-fields is possible but beyond the scope of this tutorial.)

- Notation: instead of $P_{\bigotimes_{i \in I} X_i}$ one may also write $P_{(X_i)_{i \in I}}$ or simply P_X if $X = \bigotimes_{i \in I} X_i$ is understood.

- Practical work in statistics and machine learning operates on distributions P_X, not on the original underlying probability measure P.

9.2 Distributions: concrete ways to "write them down"

Distributions P_X can be very complex and large-size objects, especially when $X = \bigotimes_{i \in I} X_i$ is a joint distribution involving many (and possibly many different sorts of) RVs X_i. Making a (joint) distribution manageable and amenable to algorithmic processing has led to a large variety of formalisms for *representing* distributions. Such formalisms carry exotic names like "mixtures of Gaussians", "Gaussian processes", "kernel machines", "neural networks", and many more. In a sense, machine learning can be seen as the art of managing complex distributions on computers. In the machine learning part of this course we will get to know some of those methods.

There are however two elementary formalisms to represent two elementary kinds of distributions, which are eminently useful and which appear as building blocks in many other, more complex formalisms. I will briefly describe these two.

9.2.1 Representing discrete distributions by probability mass functions

When the sample space $S = \{s_1, \ldots, s_N\}$ is finite, or if the sample space $S = \{s_1, s_2, \ldots, \}$ is countably infinite, one speaks of a *discrete* space and a distribution over a discrete space is called a *discrete distribution.*

Side note: Much more generally, *discrete mathematics* comprises all those branches of mathematics where one deals with at most countably infinite sets or structures made from countable components – this includes number theory, many subfields of graph theory, much of logic and algebra – and is opposed to calculus and linear algebra, where one deals with "continuous" objects like curves and vector spaces, which are uncountable sets). And again generally speaking, the formal methods and indeed the very ways of thinking are very different in discrete versus continuous mathematics.

For discrete sample spaces S, the σ-field that is almost always invoked is the power set $\mathcal{F} = \mathrm{Pot}(S)$. A distribution P_X on $\mathrm{Pot}(S)$ is fully specified by the probabilities of the singleton elements of $\mathrm{Pot}(S)$, that is, by the function

$$p : S \to [0,1], \quad s \mapsto P_X(\{s\}). \tag{9.2}$$

Such functions p are called *probability mass functions* (pmf). They are defined by the following condition:

Definition 9.2.1 *Let S be a finite or countably infinite, nonempty set. A function $p : S \to [0,1]$ satisfying*

$$\begin{aligned} p(s) &\geq 0 \\ \sum_{s \in S} p(s) &= 1 \end{aligned}$$

is called a probability mass function*.*

In case that S is finite, the value vector $(P_X(\{s_1\}), \ldots, P_X(\{s_N\}))$ is called a *probability vector.* Obviously, probability vectors are just the vectors that are non-negative and sum to 1.

For simplified notation one often drops the set braces and writes $(P_X(s_1), \ldots, P_X(s_N))$. When $A \subseteq S$ is any element of $\mathrm{Pot}(S)$, the probability of A is computed by summing its elements' probabilities

$$P_X(A) = \sum_{s \in A} P_X(s).$$

When S is finite and small enough, the probability vector representing the distribution can be explicitly represented in the computer program that one uses. Often however S will be finite but very (veeery!!) large — this happens easily when S is a product space obtained from combining numerous finite small factor spaces S_i. Our earlier Credit risk example was of this kind. In this situation, the (joint) distribution still is mathematically represented by a probability vector, but it is impossible to "write it down" or store it in computer memory because of its size. Another problem that arises from very large sizes of S is numerical underflow: most $P_X(s_i)$ will necessarily be closer to zero than machine precision can handle. The standard escape from that problem is to use log probabilities $\log P_X(s_i)$ instead of the unscaled probabilities $P_X(s_i)$. In fact, many theorems and algorithms in machine learning are directly formulated in terms of log probabilities.

9.2.2 Representing continuous distributions by pdfs and cdfs

Another basic, frequently encountered type of distributions arises when the sample space S is $\mathbb{R}^n$ or a subset thereof (for instance the unit hypercube). Then one can often use *probability density functions* (pdf's) to represent these distributions:

Definition 9.2.2 *Let $X : \Omega \to \mathbb{R}^n$ a RV taking values in $\mathbb{R}^n$, and let $p : \mathbb{R}^n \to \mathbb{R}$ be a function. If the distribution of X satisfies the condition*

$$P_X([a_1, b_1] \times \ldots \times [a_n, b_n]) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} p(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \tag{9.3}$$

for all intervals $\int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1}$, then p is called a probability density function *of X. It is denoted by p_X.*

Not every continuous-valued RV admits a representation of its distribution by a pdf. We will however not meet with such RVs in this lecture.

A more compact notation for the integral (9.3) is

$$\int_D p(\mathbf{x})\, d\mathbf{x},$$

where D denotes the k-dimensional interval $[a_1, b_1] \times \ldots \times [a_n, b_n]$ and $\mathbf{x}$ denotes vectors in $\mathbb{R}^k$.

For many frequently occurring distributions, analytical formulas for their pdf are known. The most famous and most widely encountered distribution is the *normal distribution* (the "Gaussian"). In the 1-dimensional case, its classical bell-shaped pdf is given by $p(x) = (2\pi\sigma^2)^{-1/2}\ \exp(-(x-\mu)^2/2\sigma^2)$, where μ is the mean and σ the standard deviation. I am sure you have met this beauty before. Figure 9.1 shows a plot of its pdf and depicts how the probability of an interval event corresponds to the pdf integral over this interval.

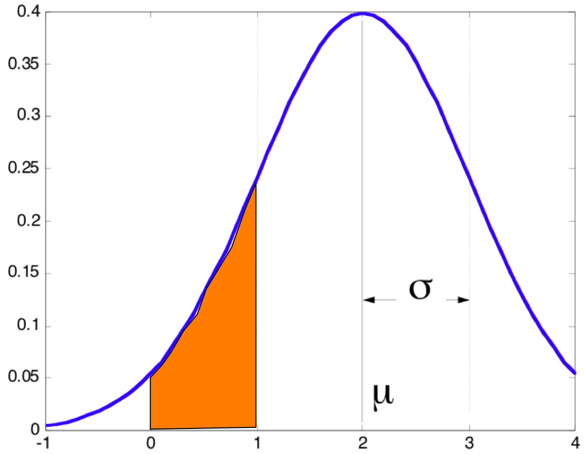


Figure 9.1: The pdf of the 1-dimensional Gaussian distribution with mean $\mu = 2$ and standard deviation $\sigma = 1$. The orange area gives the probability of the event $[0, 1]$.

When the sample space S is the 1-dimensional real line, and a distribution P_X on S is described by its pdf p, then there is another possibility to describe P_X which can be derived from the pdf, called the *cumulative density function* (cdf), denoted by F. It is defined by

$$F(\alpha) = \int_{-\infty}^{\alpha} p(x)\, dx. \tag{9.4}$$

This generalizes to higher dimensions, though cdf's of higher-dimensional distributions are rarely used. I only give the definition for the 2-dimensional sample space $S = \mathbb{R}^2$ on which a distribution $P_{X,Y}$ with a pdf $p(x, y)$ is given:

$$F((\alpha, \beta)) = \int_{-\infty}^{\alpha} \int_{-\infty}^{\beta} p(x, y)\, dx\, dy. \tag{9.5}$$

Some comments:

- A probability density function is *defined* to be a function which allows one to compute probabilities of value intervals as in Equation (9.3). For a given continuous RV

$X : \Omega \to \mathbb{R}$ over the reals this pdf, if it exists, is essentially unique. More precisely, "essentially unique" means that two pdfs p_X, q_X for the same distribution may only differ from each other on a *null set* (compare Chapter 13).

- Any pdf $p_X : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$ has the property that it integrates to 1, that is, $\int_{\mathbb{R}^n} p_X(\mathbf{x})\, d\mathbf{x} = 1$.
- Be aware that the values $p(x)$ of a pdf are not probabilities! Pdf's turn into probabilities only through integration over intervals. Values $p(x)$ can be greater than 1, again indicating that they cannot be taken as probabilities.

For another example, consider the uniform distribution on $[0, 2]$. Figure 9.2 shows its pdf and its cdf.

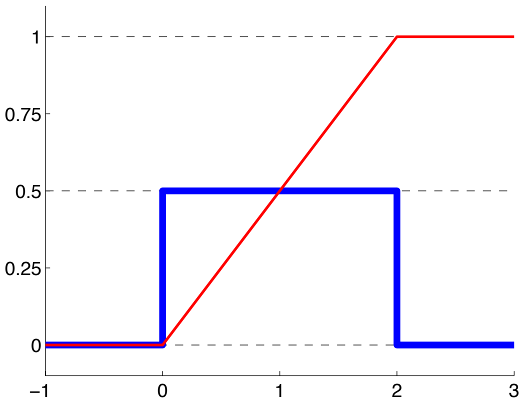


Figure 9.2: The pdf (blue) and cdf (red) of the uniform distribution on $[0, 2]$.

For the 1-dimensional $S = \mathbb{R}$, the cdf gives an immediate answer to the question "how probable is it that the RV X takes a value less or equal to α?":

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

This property of the cdf is frequently exploited in statistical decision making, as we will see in a few weeks.

In Equation (7.4) we introduced the conditional probability $P(X \in A \mid Y \in B)$. In that definition it was essential that $P(Y \in B) > 0$. In distributions that admit pdf's, one can also define conditional probabilities for cases where $P(Y \in B) = 0$. For getting the idea, consider a pdf $p_{X,Y}$ on $\mathbb{R}^2$. Then, if certain conditions concerning the continuity of $p_{X,Y}$ are given which we will not discuss, one can define the *conditional density function* $p_{X|Y=y} : \mathbb{R} \to \mathbb{R}^{\geq 0}$ by

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{\int_{\mathbb{R}} p_{X,Y}(x,y)\, dx}, \tag{9.6}$$

from which conditional probabilites of X given $Y = y$ can be obtained by

$$P_{X|Y=y}([a,b]) = \int_a^b p_{X|Y=y}(x)\, dx.$$

The conditions necessary to make this work are warranted, for instance, when the joint pdf $p_{X,Y}$ is differentiable.

Notice that for continuous distributions $P_{X,Y}$ one usually has $P(Y = y) = 0$, so we have a true extension of generality here compared to the basic definition (7.4).

9.3 Marginals

Computing marginal distributions from a joint distribution is one of the most frequently used operations in theory and practice of data analysis. Since a good intuitive grasp on this concept is a real empowerment for you, I will spend some space on explaining it carefully.

We first consider the case of two discrete RVs. For instance, an online shop creating customer profiles may record from their customers their age and gender (among very many other items), giving two RVs that we will denote by Age and Gender. The marketing optimizers of that shop are not interested in the exact age but only in age brackets, say $a_1 = \mathsf{at\ most\ 10\ years\ old}$, $a_2 = 11 - 20$ years, $a_3 = 21 - 30$ years, $a_4 = \mathsf{older\ than\ 30}$. Gender is roughly categorized into the possibilities $g_1 = \mathsf{f}, g_2 = \mathsf{m}, g_3 = \mathsf{o}$. From their customer data the marketing guys estimate the following probability table:

$P(X = g_i, Y = a_j)$	a_1	a_2	a_3	a_4
g_1	0.005	0.3	0.2	0.04
g_2	0.005	0.15	0.15	0.04
g_3	0.0	0.05	0.05	0.01

(9.7)

The cell (i, j) in this 3×4 table contains the probability that a customer with gender g_i falls into the age bracket a_j. This is of course the joint probability of the two observation values g_i and a_j. Notice that all the numbers in the table sum to 1.

If we are interested only in the age distribution of customers, ignoring the gender aspects, we sum the entries in each age column and get the *marginal probabilities* of the RV Age. Formally, we compute

$$P(\mathsf{Age} = a_j) = \sum_{i=1,2,3} P(\mathsf{Gender} = g_i, \mathsf{Age} = a_j). \tag{9.8}$$

Similarly, we get the marginal distribution of the gender variable by summing along the rows. The two resulting marginal distributions are indicated in the table (9.9).

	a_1	a_2	a_3	a_4	
g_1	0.005	0.3	0.2	0.04	**0.545**
g_2	0.005	0.15	0.15	0.04	**0.345**
g_3	0.0	0.05	0.05	0.01	**0.110**
	0.01	**0.5**	**0.4**	**0.09**	

(9.9)

Notice that the marginal probabilities of age $0.01, 0.5, 0.4, 0.09$ sum to 1, as do the gender marginal probabilities.

Marginal probabilities are also defined when there are more than two RVs. For instance, the online shop marketing people also record how much a customer spends on average, and formalize this by a third random variable, say $\mathsf{Spending}$. The values that $\mathsf{Spending}$ can take are spending brackets, say $s_1 = \mathsf{less\ than\ 5\ Euros}$ to $s_{20} = \mathsf{more\ than\ 5000\ Euros}$. The joint probability values $P(\mathsf{Gender} = g_i, \mathsf{Age} = a_j, \mathsf{Spending} = s_k)$ would be arranged in a 3-dimensional array sized $3 \times 4 \times 20$, and again all values in this array together sum to 1. Now there are different arrangements for marginal probabilities. For instance the probabilities $P(\mathsf{Gender} = g_i, \mathsf{Spending} = s_j)$ are the marginal probabilities obtained by *summing away* the Age variable:

$$P(\mathsf{Gender} = g_i, \mathsf{Spending} = s_j) = \sum_{k=1,2,3,4} P(\mathsf{Gender} = g_i, \mathsf{Age} = a_k, \mathsf{Spending} = s_j)$$

The general formula for marginal distributions obtained from the joint distribution $P_{X_1,\ldots,X_N}$ of N discrete RVs with sample spaces $S^1,\ldots,S^N$ goes like this. Let $\{i_1,\ldots,i_K\} = J \subset \{1,\ldots,N\}$ be the indices of the RVs whose collective marginal distribution we want to get, and let $\{j_1,\ldots,j_{N-K}\} = \{1,\ldots,N\} \setminus J$ be the indices of the RVs that we want to sum away. Then

$$P(X_{i_1} = s^{i_1},\ldots,X_{i_K} = s^{i_K}) = \tag{9.10}$$
$$\sum_{s^{j_1}\in S^{j_1},\ldots,s^{j_{N-K}}\in S^{j_{N-K}}} P(X_{i_1} = s^{i_1},\ldots,X_{i_K} = s^{i_K}, X_{j_1} = s^{j_1},\ldots,X_{j_{N-K}} = s^{j_{N-K}}).$$

In an analog fashion, when we are dealing with real-valued RVs whose joint distribution can be described by a pdf, the pdf's of marginal distributions are obtained by *integrating away* the RVs that one wants to omit. I give only the formula for two RVs $X, Y : \Omega \to \mathbb{R}$, whose joint pdf $p_{X,Y}$ can be integrated to obtaine the pdf p_X for X in a way that is structurally the same as we saw in (9.8):

$$p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x,y)\,dy. \tag{9.11}$$

Plugging this finding into the formula (9.6) of conditional density functions, you get the simpler-looking

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)}. \tag{9.12}$$

9.4 Connecting joint, marginal and conditional probabilities

The basic formula of conditional probability can be rearranged in interesting ways:

$$P(X \in A \,|\, Y \in B) = \frac{P(X \in A, Y \in B)}{P(Y \in B)}, \text{ or} \tag{9.13}$$
$$P(X \in A, Y \in B) = P(X \in A \,|\, Y \in B)\,P(Y \in B), \text{ or} \tag{9.14}$$
$$P(Y \in B) = \frac{P(X \in A, Y \in B)}{P(X \in A \,|\, Y \in B)}, \tag{9.15}$$

where the conditional, joint and marginal probabilities become expressed in terms of the respective others. When you memorize one of these formulas (I recommend the second one), you have memorized the very key to master probability arithmetics and you will never get lost when manipulating probability formulas.

Chapter 10

Expectation, variance, standard deviation, covariance

When the sample space S of a RV X is *numerical* – that is, S is a subset of $\mathbb{N}^n, \mathbb{Z}^n, \mathbb{R}^n$ or $\mathbb{C}^n$ – , there are two quantities which characterize this RV in an elementary way, and which are considered or computed by any statistician or machine learner almost by reflex: the expectation and the variance.

The *expectation* of X is the "average" value that is measured through X. It is denoted by $E[X]$. If S is discrete with pmf p, then

$$E[X] = \sum_{s \in S} s \cdot p(s), \tag{10.1}$$

and when S is continuous with pdf p, it is

$$E[X] = \int_S s \cdot p(s)\; ds, \tag{10.2}$$

provided that the (possibly infinite) sum or the integral exists. When the sum or the integral is infinite, the expectation is not defined.

It is important to keep apart the concepts of expectation from the concept of a *mean.* The expectation is a property of a distribution of a (numerical) RV, and a RV has only one unique expectation (if it is defined). In contrast, the mean is a property of a finite sample of numerical measurements. If $X_1, \ldots, X_N : \Omega \to S$ are N numerical RVs, then the mean of a (mathematical) sample is

$$\text{mean}(\{X_1(\omega), \ldots, X_N(\omega)\} = \frac{1}{N} \sum_{i=1}^{N} X_i(\omega).$$

For different elementary events ω one may get different sample means.

The *variance* of a scalar (that is, 1-dimensional) random variable X is the expectation of the *centered* (that is, subtracting the expectation, yielding a version of the RV that has zero expectation), squared RV:

$$\text{Var}[X] = E[(X - \mu)^2] = \begin{cases} \sum_{s \in S} (s - \mu)^2 \cdot p(s) & \text{for discrete, scalar RV's} \\ \int_{\mathbb{R}} (x - \mu)^2 \cdot p(x)\; dx & \text{for scalar RV's with pdf } p. \end{cases} \tag{10.3}$$

Besides $\text{Var}(X)$, another common notation for the variance of a scalar RV is $\sigma^2(X)$. Like with expectation, the variance is only defined if the sum (or integral) is finite.

The square root of $\sigma^2(X)$, that is $\sigma(X) = \sqrt{\sigma^2(X)}$, is called the *standard deviation* of X. A common abbreviation is “stddev”.

A (real-valued, one-dimensional) RV is transformed to a standardized version, called the *standardized* RV, by centering and dividing by the standard deviation, getting a version that is often denoted by the letter Z:

$$Z = \frac{X - \mu}{\sigma}.$$

While the expectation is immediately defined by (10.1), (10.2) also for vector-valued RVs, the variance of vector-valued quantities needs a separate treatment. The counterpart of σ^2 for a RV that takes values in $\mathbb{R}^n$ is the $n \times n$ *correlation matrix* Σ defined by

$$\Sigma = E[(X - \mu)\,(X - \mu)^{\mathsf{T}}], \tag{10.4}$$

that is, $\Sigma(i,j) = E[(X_i - \mu_i)(X_j - \mu_j)]$, where X_i, μ_i denote the i-the component RV of X and its expectation (I use $\cdot^{\mathsf{T}}$ to denote vector/matrix transpose). Somewhat counter-intuitively, one uses the symbol Σ, not Σ^2, for this matrix. One does not commonly define or use a standard deviation for vector RVs.

Beyond the expectation $E[X]$ and the variance $E[(X - \mu)^2]$ one also considers $E([(X - \mu)^3], E([(X - \mu)^4], \ldots$, which are know as the third, fourth, etc *moments* of a (scalar) numerical RV.

The third moment of the standardized version of X, $E([((X - \mu)/\sigma)^3]$ is also known as the *skewness* of X; it is a measure of the asymmetry of the distribution. Centered RVs with a symmetric distribution have zero skewness.

The fourth standardized moment $E([((X - \mu)/\sigma)^4]$ is known as the *kurtosis* of X. To briefly explain the information conveyed by the kurtosis, I quote from `https://en.wikipedia.org/wiki/Kurtosis`: “The kurtosis of any univariate normal distribution is 3. It is common to compare the kurtosis of a distribution to this value. Distributions with kurtosis less than 3 are said to be *platykurtic*, although this does not imply the distribution is ”flat-topped” as sometimes reported. Rather, it means the distribution produces fewer and less extreme outliers than does the normal distribution. An example of a platykurtic distribution is the uniform distribution, which does not produce outliers. Distributions with kurtosis greater than 3 are said to be *leptokurtic*. An example of a leptokurtic distribution is the Laplace distribution, which has tails that asymptotically approach zero more slowly than a Gaussian, and therefore produces more outliers than the normal distribution.”

The *covariance* between two numerical RVs $X, Y : \Omega \to \mathbb{R}$ is defined by

$$\mathrm{Cov}[X, Y] = E[(X - E[X])\,(Y - E[Y])]. \tag{10.5}$$

This quantity measures the extent to which X and Y “co-vary”, that is, the extent to which large values of X coincide with large values of Y (and small values with small values). It is easy to see (see it!) that

$$\mathrm{Cov}[X, Y] = E[X\,Y] - E[X]\,E[Y]. \tag{10.6}$$

The covariance scales with scalings of X and Y:

$$\mathrm{Cov}[a\,X, b\,Y] = ab\,\mathrm{Cov}[X, Y].$$

Normalizing the covariance by the stdevs, one obtains a measure of co-variation which is scaling-independent. It is called the *correlation* of X and Y:

$$\mathrm{Corr}[X, Y] = \frac{E[(X - E[X])\,(Y - E[Y])]}{\sigma(X)\,\sigma(Y)}. \tag{10.7}$$

This quantity ranges between -1 and $+1$ and remains the same when X or Y are scaled:

$$\mathrm{Corr}[a\,X, b\,Y] = \mathrm{Corr}[X, Y].$$

The case $\mathrm{Corr}[X, Y] = 1$ is obtained when X and Y are identical up to positive scaling and shift:

$$\mathrm{Corr}[X, Y] = 1 \;\Leftrightarrow\; X = a\,Y + b \text{ (for } a > 0).$$

The case $\mathrm{Corr}[X, Y] = -1$ is obtained when X and Y are identical up to negative scaling and shift:

$$\mathrm{Corr}[X, Y] = -1 \;\Leftrightarrow\; X = -a\,Y + b \text{ (for } a > 0).$$

In this case one says that X and Y are *anti-correlated.*

The squared correlation $\mathrm{Corr}[X, Y]^2$ ranges between 0 and 1 and is called the *correlation coefficient* of X and Y.

When $\mathrm{Cov}[X, Y] = 0$, which is equivalent to $\mathrm{Corr}[X, Y] = 0$, X and Y are called *uncorrelated.*

Uncorrelatedness of X and Y is easy to check (approximately) from a sample $(X(\omega_i), Y(\omega_i))_{i=1,\ldots,N} = (x_i, y_i)_{i=1,\ldots,N}$, as follows. First subtract the sample means $\hat\mu_X = (1/N) \sum_i x_i, \hat\mu_Y = (1/N) \sum_i y_i$, obtaining *centered* samples $(\tilde{x}_i, \tilde{y}_i) = (x_i - \hat\mu_X, y_i - \hat\mu_Y)$. Then normalize the centered samples to unit variance, by computing the sample stddevs $\hat\sigma_{\tilde{X}} = (\sum_i \tilde{x}_i{}^2)^{1/2}$ and $\hat\sigma_{\tilde{Y}} = (\sum_i \tilde{y}_i{}^2)^{1/2}$, and dividing the centered samples by these, obtaining centered normalized samples $(\tilde{\tilde{x}}_i, \tilde{\tilde{y}}_i) = (\tilde{x}_i/\hat\sigma_{\tilde{X}}, \tilde{y}_i/\hat\sigma_{\tilde{Y}})$. Then, if

$$\frac{1}{N} \sum_{i=1,\ldots,N} \tilde{\tilde{x}}_i \cdot \tilde{\tilde{y}}_i \approx 0,$$

we have a numerical indication that X and Y are uncorrelated (compare (10.6)).

A note on terminological confusion. The terminology that I just introduced is used by mathematicians, statisticians, machine learners, and the Matlab software. However, in the field of signal processing and control engineering, the word "correlation" is often used to denote just the expected product $E[X\,Y]$ of two RVs which do not have to have zero mean. In the signal processing literature (and in the papers that I write!) the word "correlation matrix" refers not to $E[(X-\mu)\,(X-\mu)^\mathsf{T}]$ but to $E[X\,X^\mathsf{T}]$ and is commonly denoted by the symbol R, not by Σ. In that literature, $\Sigma = E[(X-\mu)\,(X-\mu)^\mathsf{T}]$ is called *covariance matrix.*

I conclude this section with a list of elementary facts concerning expectation and variance:

- $$E[a\,X + b\,Y] = a\,E[X] + b\,E[Y], \tag{10.8}$$
  that is, the expectation is a *linear* operator on numerical RVs.

- For a ($\mathfrak{B}$-$\mathfrak{B}$-measurable) function $f : \mathbb{R} \to \mathbb{R}$, and a RV $X : \Omega \to \mathbb{R}$ with a pdf p,
  $$E[g(X)] = \int_{\mathbb{R}} g(x)\,p(x)\,dx, \tag{10.9}$$
  that is, we do not need to calculate the pdf of $g(X)$ to obtain $E[g(X)]$.

- $$\sigma^2(X) = E[X^2] - E[X]^2. \tag{10.10}$$

- $$\sigma^2(a\,X + b\,Y) = a^2\,\sigma^2(X) + b^2\,\sigma^2(Y) + 2ab\,\mathrm{Cov}[X, Y]. \tag{10.11}$$

Chapter 11

More on the independence of RVs

In Section 8.2 I gave the definition of when some RVs are called independent. For convenience I repeat Eqn. 8.2, a standard textbook version of that definition: $(X_n)_{n=1,\ldots,N}$ are independently distributed if for all $A_1 \in \mathcal{F}_1, \ldots, A_N \in \mathcal{F}_N$:

$$P(X_1 \in A_1, \ldots, X_N \in A_N) = P(X_1 \in A_1) \cdot \ldots \cdot P(X_N \in A_N).$$

Using the terminology that we have introduced in the meantime, we can express this condition (8.2) in words: $(X_n)_{n=1,\ldots,N}$ are independently distributed if their joint distribution is the product of their marginal distributions.

Consider the following two (made-up) joint distribution tables of two RVs $X : \Omega \to \{g_1, g_2, g_3\}, Y : \Omega \to \{a_1, \ldots, a_4\}$ (the first is repeated from (9.7)):

Table 1:

	a_1	a_2	a_3	a_4	
g_1	0.005	0.3	0.2	0.04	**0.545**
g_2	0.005	0.15	0.15	0.04	**0.345**
g_3	0.0	0.05	0.05	0.01	**0.110**
	0.01	**0.5**	**0.4**	**0.09**	

Table 2:

	a_1	a_2	a_3	a_4	
g_1	0.05	0.05	0.1	0.3	**0.5**
g_2	0.02	0.02	0.04	0.12	**0.2**
g_3	0.03	0.03	0.06	0.18	**0.3**
	0.1	**0.1**	**0.2**	**0.6**	

In these tables, the boldface columns/rows give the marginals of X and Y, respectively. Table 1 shows the joint distribution of two dependent RVs, while the joint distribution shown in Table 2 reveals independence of X and Y. In Table 2, each column is a copy of the X marginal distribution, scaled by the Y-marginal of that column; and each row is a copy of the Y marginal distribution, scaled by the X-marginal of that row.

When we are dealing with continuous-valued numerical RVs which have pdf's, independence reaveals itself in the fact that the pdf $p_{X_1,\ldots,X_N}$ of the joint distribution can be written as a product of the pdf's of the individual RVs:

$$p_{X_1,\ldots,X_N}(x_1, \ldots, x_N) = p_{X_1}(x_1) \cdot \ldots \cdot p_{X_N}(x_N).$$

Often one wishes to model situations when the effects of different numerical random quantities *sum up* to some final effect. For instance, the effective speed of a ship may be modeled as the sum of speed components yielded by wind, water currents, and the ship's engine. If the component quantities are independent, their summed-up effect can be computed by an operation called *convolution.* In the simplest case, let $X, Y : \Omega \to \mathbb{R}$ be two independent RVs, with pdf's p_X, p_Y. Then the sum RV $Z = X + Y$ has the pdf given by

$$p_Z(z) = p_{X+Y}(z) = \int_{\mathbb{R}} p_X(z - y)\, p_Y(y)\, dy = \int_{\mathbb{R}} p_Y(z - x)\, p_X(x)\, dx. \tag{11.1}$$

A generalization of independence which is of monumental importance for machine learning is *conditional independence.* Two discrete RVs X, Y are said to be *conditionally independent given* Z if for all values x, y, z of these three RVs, one has

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)\; P(Y = y \mid Z = z).$$

In the continuous-valued case, where X, Y, Z have pdfs p_X, p_Y, p_Z, this spells out to

$$\forall z : \quad p_{X,Y|Z=z} = p_{X|Z=z}\; p_{Y|Z=z}.$$

Statisticians and even more so, machine learners are extremely eager to hunt for independently distributed RVs. There are several reasons for this. A simple reason is economy. In order to specify the joint probability shown in Table 1, one has to specify $3 \cdot 4 - 1 = 11$ parameters, namely all the entries in the table (why the -1?). In contrast, in order to specify the joint distribution in Table 2, only $(3 - 1) + (4 - 1) = 5$ parameters have to specified, namely the marginals. This saving in required storage capacity, from 11 to 5, may not seem striking. But – consider the joint distribution of $N = 100$ RVs, each of which can take values in the same minature sample space $S = \{0, 1\}$. In case that the $X_1, \ldots, X_{100}$ are not independent, in order to fully specify the joint distribution table (a 100-dimensional array in that case) one would have to compute and store $2^{100} - 1$ parameters – impossible. But if these 100 RVs would be independent, only $100 \cdot (2 - 1) = 100$ parameters would be necessary! Dealing with the joint distribution of very many RVs — up to thousands or even millions — is commonplace in machine learning (and, by the way, in computer vision, where each pixel turns into a RV). This becomes technically possible (storage capacity is limited!) *only* if many of the concerned RVs are conditionally independent of each other in some way or other. In the terminology of statistics and machine learning, one says that the joint distribution can be *factorized.* Significant portions of the theory of machine learning deal with methods to find ways to (approximately) factorize complex joint distributions.

Factorizing joint distributions is desirable not just for storage capacity economy. Another, equally crucial reason is that if fewer parameters have to *estimated* from training data, one needs fewer training data. A blunt rule of thumb says that for a halfway reliable parameter estimation, one should have ten times as many training data points as one has parameters to estimate. In the above example this would mean that in order to estimate the joint distribution of 100 binary RVs, one would need $10 \cdot (2^{100} - 1)$ training data points, whereas if their distribution would fully factorize, a training data set of size $10 \cdot 100$ would suffice.

Independence of two RVs X, Y is a *much* stronger condition than uncorrelatedness of X, Y, for two reasons:

- Independence is defined for any X, Y regardless of what their sample spaces are. In contrast, uncorrelatedness is defined only for RVs whose sample space is $\mathbb{R}$.
- For two real-valued RVs X, Y, independence implies uncorrelatedness (exercise) but not vice versa (another exercise!).

I mentioned in the previous section that uncorrelatedness of X, Y is easy to check (approximately) from a sample. In contrast, verifying independence cannot be easily checked from samples; it requires an insight into the joint distribution of X and Y which cannot easily be gleaned from a sample. An extensive, nontrivial literature deals with approximate numerical checks for independence from samples.

Chapter 12

Markov chains

Markov chains are the most simple non-trivial kind of stochastic process; according to a core dogma of physics, the world is a Markov chain (even when described by quantum mechanics); Markov chains are at the core of all modern *sampling algorithms*, and Markov chains provide a nice demo case for the powers of statistical independence. Four good reasons to become acquainted with them!

In its basic version, a Markov chain is a discrete-time stochastic process $(X_n)_{n \in \mathbb{N}}$ where each X_n takes values in the same set S. Recall from Chapter 4 that a sequence $X_0(\omega), X_1(\omega), X_2(\omega), \ldots$ is called a realization, or path, or trajectory of the process. I will use the word "realization". For different ω we obtain different realizations in general – this is what makes this a *stochastic* process. Here is defining criterion that qualifies a discrete stochastic process as a Markov chain (where we use the shorthand notation $P(x_i)$ for $P(X_i = x_i)$:

Definition 12.0.1 *A discrete-time stochastic process* $(X_n)_{n \in \mathbb{N}}$ *with values in a discrete set* S *is a* Markov chain *(MC) if for all* $n \geq 2, 0 < m < n, (x_0, \ldots, x_n) \in S^{n+1}$

$$P(x_0, x_1, \ldots, x_{m-1},\ x_{m+1}, \ldots, x_n \mid x_m) = P(x_0, \ldots, x_{m-1} \mid x_m) \cdot P(x_{m+1}, \ldots, x_n \mid x_m). \tag{12.1}$$

In plain English: a discrete-time stochastic process is a MC if for all times m, what happens next $(x_{m+1}, \ldots, x_n)$ is conditionally independent from what happened before $(x_0, \ldots, x_{m-1})$, given knowledge of what happens now (x_m). An equivalent definition, the one which one finds in most textbooks, is

Definition 12.0.2 [Alternative, equivalent definition] *A discrete-time stochastic process* $(X_n)_{n \in \mathbb{N}}$ *with values in* S *is a* Markov chain *(MC) if for all* $n > 0$

$$P(x_{n+1} \mid x_0, x_1, \ldots, x_n) = P(x_{n+1} \mid x_n). \tag{12.2}$$

Proving the equivalence of both definitions is a good exercise. The alternative definition could be paraphrased as "a Markov chain is a stochastic process where the distribution of future observations $(= X_{n+1}(\omega))$ depends only on the current observation $(= X_n(\omega))$ and not on the previous history $(= X_{n-1}(\omega), X_{n-2}(\omega), \ldots, X_0(\omega))$". This is why Markov chains are also called *memoryless* processes.

I insert a note why physicists see the world as a Markov chain (I ignore the fact that they use continuous time, which technically then would lead to the continuous-time twin of Markov chains, called *Markov processes*). The big central unshakeable foundation of all physics models and formalisms is the idea that one can describe any physical system by its *states*. A state $x(t)$ of a physical system is a "snapshot" of *everything* in that system at a given point in time. For instance, a state $x(t)$ of our earth's atmosphere would be a full specification of the current position, orientation, velocities, angular velocities, and oscillatory modes of each and every molecule in the atmosphere. Physicists believe that if the state $x(t)$ is given, then the laws of nature – deterministic in classical physics and stochastic in quantum mechanics – take care of the future evolution of the system state for times $t' > t$. Any information from the past which is relevant for the future must be encoded in the current state $x(t)$. This commitment is implicit in the physicists' way of describing physical systems by state-based formalisms, like differential equation or partial differential equations.

As an echo of this physicists' view, it is customary in mathematics to call the sample space S the *state space* of a Markov chain, and the elements $s \in S$ its *states*.

For the rest of this section I will take a closer look at the case where the state space is finite, that is $S = \{s_1, \ldots, s_N\}$. Furthermore, I will only consider MCs that are *homogeneous*:

Definition 12.0.3 *A Markov chain $(X_n)_{n\in\mathbb{N}}$ is called* homogeneous *if for all time points $n, n' \in \mathbb{N}$ and for all states $s, s' \in S$*

$$P(X_{n+1} = s' \mid X_n = s) = P(X_{n'+1} = s' \mid X_{n'} = s). \tag{12.3}$$

In words: (X_n) is homogenous if the probabilities to jump from state s to state s' are the same at all times. The probabilities $P(X_{n+1} = s' \mid X_n = s)$ are also called the *transition probabilities* of the MC. We will use shorthand $p(s, s')$ for the transition probability $P(X_{n+1} = s' \mid X_n = s)$. I will only consider homogeneous MCs in the remainder of this section.

Any finite-valued stochastic process with discrete time $T = \mathbb{N}$ is fully characterized by the probabilities of its *initial realizations*

$$P(X_0 = s_{i_0}, \ldots, X_n = s_{i_n}),$$

where $n \geq 0$ and $s_{i_0}, \ldots, s_{i_n} \in S$. The shortest initial realization is given by $X_0(\omega)$ – we just observe the initial state of a trajectory. The distribution of X_0 is an N-dimensional probability mass function, which can be written as an N-dimensional probability vector, often denoted by π_0. We write $\pi_0(s_i)$ for the i-th element $P(X_0 = s_i)$ of this vector.

The distribution of length-2 initial realizations, that is the distribution of $X_0 \otimes X_1$, i.e. the probabilities $P(X_0 = s_{i_0}, X_1 = s_{i_1})$, can be computed by

$$P(X_0 = s_{i_0}, X_1 = s_{i_1}) = P(X_1 = s_{i_1} \mid X_0 = s_{i_0})\, P(X_0 = s_{i_0}) = \pi_0(s_{i_0})\, p(s_{i_0}, s_{i_1}), \tag{12.4}$$

that is by the product of the initial probability of s_{i_0} with the probability of transiting from s_{i_0} to s_{i_1}.

Similarly, the probability of a length-n realization is given by

$$P(X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_n = s_{i_n}) = \pi_0(s_{i_0})\, p(s_{i_0}, s_{i_1})\, p(s_{i_1}, s_{i_2}) \cdots p(s_{i_{n-1}}, s_{i_n}). \tag{12.5}$$

The transition probabilities $p(s_i, s_j)$ can be arranged in a $N \times N$ sized *transition matrix* $M = (p(s_i, s_j))_{1 \leq i,j \leq N}$. If you think about it you will see that each row of this matrix sums to 1. Nonnegative matrices with unit row sums are called *Markov matrices* or *stochastic matrices*. Such matrices occur ubiquitously in the study of stochastic systems. Be alerted that some authors use these terms for nonnegative matrices whose *columns* sum to 1 – be careful when reading texts about Markov matrices.

The probability that a trajectory of a MC hits state s_i at time $n = 1$ is the sum of all probabilities for possible ways how the process may find itself in state s_i a time 1:

$$P(X_1 = s_i) = \sum_{k=1,\ldots,N} P(X_0 = s_k, X_1 = s_i),$$

in which formula we meet our old acquaintance, the marginal. Re-writing this according to (12.4) gives

$$P(X_1 = s_i) = \sum_k \pi_0(s_k)\, p(s_k, s_i).$$

The entire distribution P_{X_1} of a MC at time 1 can be written as an N-dimensional probability vector π_1 which collects all the probabilities $P(X_1 = s_i)$. If you think about it (I like this phrase) you will find that

$$\pi_1 = M^{\mathsf{T}}\, \pi_0, \tag{12.6}$$

and more generally, the state distribution π_n at time n is given by

$$\pi_n = (M^{\mathsf{T}})^n\, \pi_0. \tag{12.7}$$

This is a very convenient and powerful formula. It opens the study of stochastic processes for tools of linear algebra.

So much for a glimpse on Markov chains. From the seeds that I could show here, an extensive theory branches out which fills textbooks. Physicists, mathematicians, machine learners and all other natural scientists and economists are likewise strongly interested in this theory, albeit for different reasons. The case is clear for physicists because they regard the timeline of reality as Markovian. Mathematicians like Markov chains and Markov processes because they are easy to study on the one hand, and on the other provide entry points for the study of more involved stochastic processes. Machine learners use Markov chains as baseline models when they are asked to deliver time series prediction algorithms. Finally, scientists at large, including economists, need Markov chains whenever they want to solve complex optimization problems – most state of the art approaches to find optimal values in complex cost landscapes make use of so-called Markov Chain Monte Carlo (MCMC) methods. Check out `https://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo`!

Chapter 13

Null sets

Given a sample space S and a distribution P_X on it, some subsets $A \subset S$ may occur with zero probability, that is,

$$P(X \in A) = 0.$$

Such sets are called *null sets* (with respect to a given distribution). For instance, given the uniform distribution on $[0, 2]$ shown in Figure 9.2, the set $A = [-1, 0]$ is a null set. This is not very interesting, as it seems obvious that the probability of this set is zero because the pdf is zero over this set. But also the set $A = \{1\}$ is a null set, although here the pdf has a positive value. This can be seen by applying (9.3):

$$P(X = 1) = P_X([1, 1]) = \int_1^1 p(x)\, dx = 0.$$

Of course it is *possible* to observe an outcome $X = 1$ when the distribution is the uniform distribution on $[0, 2]$. Being possible, and having a nonzero probability is not the same thing!

Considering the second condition in the definition of a probability measure (Def. 7.1.1), we find that the union of any countable collection of null sets is again a null set. You know that the set $\mathbb{Z}$ of rational numbers is countable. Let $\mathcal{N} = [0, 2] \cap \mathbb{Z}$ be the set of all rational numbers between 0 and 2. Since $\mathcal{N}$ is a countable union of null sets, its probability is zero! In a continuous sample space, the probability of hitting a rational-valued observation outcome is zero.

When you read probability theory texts, you may see – quite often, actually – the expression "almost surely", or "P-almost-surely", or "almost all", or the abbreviation "w.p.1" which spells out to "with probability 1". They all mean the same thing: some fact about some probability space $(\Omega, \mathfrak{A}, P)$ holds true for all $\omega \in \Omega \setminus \mathcal{N}$, where $\mathcal{N}$ is a null set.

An example: in Definition 9.2.2 we said that some function p is *a* (!) pdf for P_X if it satisfies the condition (9.3). This formulation leaves open the possibility that a distribution P_X may have more than one pdf. And in fact, this is the case. The pdf's (plural!) for P_X are defined only almost surely: for any two (!) pdf's $p, p' : \mathbb{R}^n \to \mathbb{R}$ of P_X there exists a null set $\mathcal{N} \subset \mathbb{R}^n$ (that is, $P(X \in \mathcal{N}) = 0$), such that $p(x) = p'(x)$ only for $x \in \mathbb{R}^n \setminus \mathcal{N}$.

Chapter 14

A note on what you find in textbooks

Warning. Don't read this section, it will only confuse you.

Novices to probability, statistics and machine learning have a hard time to get a firm footing in these fields for a number of reasons.

1. Different textbooks use different notations for probability formulas.

2. Authors writing textbooks for non-mathematicians (students of engineering, computer science, social and natural sciences, economy) want to make it easy for their readers. Specifically, they avoid talking about σ-fields and events, which prevents them from explaining that random variables are functions. For instance, in a highly respected handbook/textbook on pattern recognition and machine learning [Duda et al., 2001], the synopsis of probability theory given in the Appendix starts like this: *"Let x be a discrete random variable that can assume any of the finite number m of different values in the set $X = \{v_1, v_2, \ldots, v_m\}$. We denote by p_i the probability that x assumes the value v_i: $p_i = Pr[x = v_i]$, ..."* That is all what is given by way of explaining "probability". Here a "random variable" really looks like what this name suggests: a "variable" that can "assume" different values. It even gets worse. A few lines later the text reads: *"Sometimes it is more convenient to express the set of probabilities $\{p_1, p_2, \ldots, p_m\}$ in terms of the* probability mass function $P(x)$, *which must satisfy the following conditions: $P(x) \geq 0$, and $\sum_{x \in X} P(x) = 1$."* This is inconsistent/incomplete in more than one way. It is not stated what is the domain of this P, the notation $\sum_{x \in X}$ suggests that the "random variable" is identified with the values of its value set, it is unclear why/how $\Pr[x = v_i]$ is different from $P(x)$, etc. Such inconsistencies in notation abound in textbooks that have been designed to make reading (too) simple. It can't be otherwise: without the kind of full story told in this tutorial, there is no way to explain a consistent (and understandable) use of notation. Something by necessity must be left dangling in mid-air.

3. A non-negligible fraction of statistics tutorials and lecture notes for non-mathematicians omits random variables altogether. Probability distributions are then directly described as probability measures on observation value spaces, and instead of the notation $P(X \in A)$ or $P_X(A)$ you find notations like $P(A)$ or $P(a)$ or similar.

4. Textbooks of probability theory written by mathematicians for math students have a consistent terminology and notation, which will be a variant of what I presented

in this tutorial. That's good. But these textbooks concentrate on the intricacies of σ-fields and the theory of measures that comes along with them. It is not or almost not explained how the mathematical structure $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ can be mapped to reality. The interpretation of Ω, X, S in terms of reality segments of interest, observation opportunities, observation procedures, observation acts, and data value spaces is pointed out in passing at best. Non-math students don't read such textbooks because of the difficulty of the mathematical theory and the lack of contact with the real world.

The deepest-buried obstacle to understanding probability however is the scintillating interpretation of what, exactly, is modeled by the fundamental probability space $(\Omega, \mathfrak{A}, P)$. Of course this obstacle only raises its head in texts and textbooks that give the full picture $X : (\Omega, \mathfrak{A}, P) \to (S, \mathcal{F})$ in the first place. There are two very different interpretations. The first is the one that I presented in this tutorial: $(\Omega, \mathfrak{A}, P)$ as a model of reality segments of interest. It is advocated by your two course instructors Adalbert F. Wilhelm and Herbert Jaeger and a few other authors, for instance Keel [2004a]. In volume 3 of his (very extensive and online) lecture notes for students of the social sciences and economy he writes (my translation from German; I don't translate the fundamental term *Grundgesamtheit*):

*"**Definition.** The* Grundgesamtheit Ω *with respect to a goal of statistical analysis is the set of possible units of investigation. Examples:*

1. *employees of a company*
2. *daily production output of a machine*
3. *time points of a day*
4. *carriers of a certain disease*
5. *drawings of a lottery*

A complete definition of Ω must include a precise prescription that determines which units of investigation, precisely, belong to Ω."

This definition of Ω carries the same intuitions as in our interpretation of Ω as a model of a reality segment of interest.

The second interpretation identifies Ω with what we called the sample space. For instance, Bauer [1978] explains (my translation from the German original edition of this textbook):

"The elements of Ω are called elementary events. *Intuitively these are the possible random outcomes of an experiment or an observation."*

Bauer does not spend effort on exploring the connection of Ω to reality — the above quote appears on page 129 of a 400-page textbook which otherwise is concerned with the pure-mathematical apparatus of σ-fields and measures. Bauer notes that this interpretation goes back to Kolmogorov [1956]. This venerable source endows this interpretation with the greatest possible authority. Kolmogorov in his foundational work writes (footnote also from Kolmogorov):

*"**2. The Relation to Experimental Data**[4] We apply the theory of probability to the actual world of experiments in the following manner:*

1. *There is assumed a complex of conditions, $\mathfrak{S}$, which allows of any number of repetitions.*
2. *We study a definite set of events which could take place as a result of the establishment of the conditions $\mathfrak{S}$. In individual cases where the conditions are realized, the events occur, generally, in different ways. Let E be the set of all possible variants $\xi_1, \xi_2, \ldots$ of the outcome of the given events. [...] We include in set E all the variants which we regard* a priori *as possible.*

...

----

[4] *The reader who is interested in the purely mathematical development of the theory only, need not read this section, since the work following it ... makes no use of the present discussion. Here we ... disregard the deep philosophical dissertations on the concept of probability in the experimental world. [...]"*

The "complex of conditions $\mathfrak{S}$" Kolmogorov refers to is what we called a reality segment of interest, and his "events" corresponds to what we called elementary events. However, Kolmogorov does not give a formal model of $\mathfrak{S}$ — this object remains in the extra-mathematical, un-modelled realm of external reality where plain English is the only description tool. The set E of "all possible variants $\xi_1, \xi_2, \ldots$ of the outcome of the given events" is what we called the sample space S. Later in his work (Chapter III) Kolmogorov proceeds to introduce random variables as maps from E to some other set E'.

The two interpretations are sometimes confounded within the writing of a single author. For instance, in volume II of his introduction to statistics Keel [2004b] writes, in contrast to his description of Ω in volume III (my translation from German):

> *"**Def. 2.1** Let a random experiment $\mathcal{E}$ have possible outcomes $\omega_1, \omega_2, \ldots, \omega_n$. The event space Ω is the set of all possible outcomes $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}; \omega_i, i = 1, \ldots, n$: elementary event."*

This is the second interpretation. Similarly, the online lecture notes of Kosuke Imai, Dpt. of Politics, Princeton University, subscribe to the second interpretation in handout 1 (`http://imai.princeton.edu/teaching/files/probability.pdf`):

> *"**Definition 1** The set of all possible outcomes of an experiment is called the sample space of the experiment and denoted by Ω. Any subset of Ω is called an event. [...] **Definition 5** A random variable is a function $X : \Omega \to \mathbb{R}$ ..."* ,

but subscribe to the first interpretation in handout 4 (`http://imai.princeton.edu/teaching/files/Convergence.pdf`) of the same course:

> *"So far we have learned about various random variables and their distributions. These concepts are, of course, all mathematical models rather than the real world itself. [...] Even if there is such a thing as "the true probability model," we can never observe it! Therefore, we must connect what we can observe with our theoretical models. The key idea here is*

that we use the probability model (i.e., a random variable and its distribution) to describe the data generating process*. What we observe, then, is a particular realization (or a set of realizations) of this random variable."*

I found the most striking co-habitation of both interpretations in a single author's writing on slide 3 of a handout for J. Utts' statistics course at the University of California at Irvinve (`http://www.ics.uci.edu/~jutts/8/Lecture14Compact.pdf`):

*"**What is a Random Variable? – Random Variable:** assigns a number to each outcome of a random circumstance* [= interpretation 2]*, or, equivalently, to each unit in a population* [= interpretation 1]*."*

To add to the confusion, the usage of the word "sample space" by different authors reflects the existence of two different interpretations of Ω. Specifically, some authors follow the first interpretation outlined above by declaring random variables as maps from "population" elementary events $\omega \in \Omega$ to observation value spaces, but they call Ω the "sample space" (for instance, `http://www.math.ku.edu/~mandal/math365/newMath365/les4.html`).

Chapter 15

A note on subjectivist approaches to probability theory

This tutorial adopted the objectivist (empiricist, realist) conception of probability as a real-world property of real-world systems, which can be quantified and measured by relative frequencies of outcomes of repeated experiments. This conception dominates mathematical textbooks and statistics courses. However, subjectivist conceptions of probability also have led to mathematical formalisms that can be used in statistical modeling. A hurdle for the learner here is that a number of different formalisms exist which reflect different modeling goals and approaches, and tutorial texts are rare.

The common starting point for subjectivist theories of probability is to cast "probability" as a subjective degree of belief, of certainty of judgement, or plausibility of assertions, or similar — instead of as an objective property of real-world systems. Subjectivist theories of probability do not develop analytical tools to describe randomness in the world. Instead they provide formalisms that code how rational beings (should) *think about* the world, in the face of various kinds of uncertainty in their knowledge and judgements. The formalisms developed by subjectivists can by and large be seen as generalizations of classical *logic*. Classical logic only knows two truth values: true or false. In subjectivist versions of logic formalisms, a proposition can be assigned graded degrees of "belief", "plausibility", etc. For a very first impression, contrast a classical-logic syllogism like

if A is true, then B is true
A is true

---

therefore, B is true

with a "plausibility reasoning rule" like

if A is true, then B becomes more plausible
B is true

---

therefore, A becomes more plausible

This example is taken from Jaynes [2003, first partial online editions in the late 1990ies], where a situation is described in which a policeman sees a masked man running away from a juweler's shop whose window was just smashed in. The plausibility rule captures the policeman's inference that the runner is a thief (A) because if a person is a thief, it is more likely that the person will run away from a smashed-in shop window (B) than when the person isn't a thief. From starting points like this, a number of logic formalisms have been devised which enrich/modify classical two-valued logic in various ways. If you want to explore these areas a little further, the Wikipedia articles probabilistic logic, Dempster-Shafer theory, fuzzy logic, or Bayesian probability are good entry points. In some of these formalisms the Kolmogorov axioms of "classical" probability re-appear as part of the respective mathematical apparatus. Applications of such formalisms arise in artificial intelligence (modeling reasoning under uncertainty), human-machine interfaces (supporting discourse generation), game theory and elsewhere.

The discipline of statistics is almost entirely developed in an objectivist spirit, firmly rooted in the frequentist interpretation of probability. Machine learning also in large parts roots in this view. However, a certain subset of machine learning models and computational procedures have a subjectivist component. These techniques are referred to as *Bayesian model estimation* methods. Bayesian modeling is particulary effective and important when training datasets are small. I will explain the principle of Bayesian model estimation with a super-simple synthetic example. For a more realistic but still didactic example from the domain of protein modeling see Section 3.4 in my machine lecture notes `http://minds.jacobs-university.de/uploads/teaching/lectureNotes/LN_ML_Fall11.pdf`.

Consider the following statistical modeling task. A measurement procedure yields real-valued outcomes (like in the SPEED-OF-LIGHT example). It is repeated N times, a condition that is modeled by samples $X(\omega) = (X_1(\omega), \ldots, X_N(\omega)) \in \mathbb{R}^N$. The RVs $X_i : \Omega \to \mathbb{R}$ which model the individual measurements are i.i.d.. We assume that their distribution can be represented by a pdf $p_{X_i} : \mathbb{R} \to \mathbb{R}^{\geq 0}$. The i.i.d. property of the family $(X_i)_{i=1,\ldots,N}$ implies that the N-dimensional pdf $p_X : \mathbb{R}^N \to \mathbb{R}^{\geq 0}$ for the distribution of the product RV $\bigotimes X$ can be written as

$$p_{\bigotimes X}((x_1, \ldots, x_N)) = p_{X_1}(x_1) \cdot \ldots \cdot p_{X_N}(x_N). \tag{15.1}$$

or in another notation (observing that all distributions P_{X_i} Are identical, so we introduce a generic RV X with $P_X = P_{X_i}$ for all i), as

$$p_{\bigotimes X}((x_1, \ldots, x_N)) = \prod_i p_X(x_i). \tag{15.2}$$

For concreteness let us consider a case where $N = 2$, that is two observations (*only* two!) have been collected, forming a sample $(X_1(\omega), X_2(\omega)) = (0.9, 1.0)$. In machine learning it is customary to call such a sample "training data", or "data vector", and refer to it by the letter D. We thus have $D = (0.9, 1.0)$ in our simplistic example.

The distribution P_X, represented by the pdf p_X, is unknown. The modeling task is to infer from the observed data D how the distribution P_X looks like, that is, how its pdf p_X is shaped. That is a hugely underdetermined problem: one cannot learn much about a potentially complicated pdf from knowing only $N = 2$ data points. So one resorts to a more modest subtask: infer from the sample what is the *expectation* (mean value) $E[X]$ of P_X:

$$E[X] =: \mu = \int_{-\infty}^{\infty} x \, p_X(x) dx. \tag{15.3}$$

The classical frequentist answer to this question is to estimate the true expectation μ_X by the sample mean, that is to compute the estimator

$$\hat{\mu}(\omega) = \frac{1}{N} \sum_{i=1}^{N} X_i(\omega), \tag{15.4}$$

which in our example gives $\hat{\mu}(\omega) = (0.9 + 1.0)/2 = 0.95$.

This is the best a classical-frequentist modeler can do. In a certain well-defined sense which we will not investigate, the sample mean is the optimal estimator for the true mean of a real-valued distribution. But "best" is not "good": with only two data points, this estimate is quite shaky. It has a high variance: if one would repeat the observation experiment with a new elementary event ω', very likely one would obtain a quite different sample and thus an estimate $\hat{\mu}(\omega')$ that is quite different from $\hat{\mu}(\omega)$.

Bayesian model estimation shows a way how to do better. It is a systematic method to exploit *prior knowledge* that the modeler may have beforehand. This prior knowledge takes the form of assumptions concerning the nature of the true distribution P_{X_i}. In our super-simple example let us assume that the modeler knows or believes that

1. the true distribution P_X is a normal distribution with unit standard deviation, that is, the pdf has the form $p_X(x) = 1/\sqrt{2\pi}\exp(-(x-\mu)^2/2)$, and

2. the true expectation μ can't be negative and it can't be greater than 1.

This kind of prior knowledge is often available. It includes knowledge or assumptions concerning the *functional type* of the true distribution (here: it is assumed to be normal distributed with unit standard deviation). Abstracting from our example, this kind of knowledge means to fix a *parametric family* of distributions as candidate solutions to the modeling task. In our example, the family is given by the familiy of pdfs $(1/\sqrt{2\pi}\exp(-(x-\mu)^2/2))_{\mu\in\mathbb{R}}$, which happens to be parametrized by a single parameter μ only – it is a one-dimensional family. In most real-world tasks one faces multi-parameter families, where each candidate distribution from the family is specified by several parameters $\theta_1, \ldots, \theta_K$ which for notational convenience are lumped together in a parameter vector $\theta = (\theta_1, \ldots, \theta_K) \in \mathbb{R}^K$. One then writes $p_X(\theta)$ to highlight the fact that the pdf p_X comes from a K-parametric family with parameters θ.

Another component of the prior knowledge is given by assumptions concerning the parameters θ of the K-parametric family $(p_{X_i}(\theta))_{\theta\in\mathbb{R}^K}$. In our mini-example the modeler felt confident to restrict the possible range of the single parameter $\theta_1 = \mu$ to the interval $[0, 1]$.

To start the Bayesian model estimation machinery, available prior knowledge about parameters θ must be cast into the form of a distribution over parameter space. For a K-parametric pdf p_X, the parameter space is $\mathbb{R}^K$. In our little example, $K = 1$ and we need to rephrase the priorly known constraint $0 \leq \mu \leq 1$ into a distribution over $\mathbb{R}$. Not knowing anything more detailed than $0 \leq \mu \leq 1$, the natural, most non-committing distribution is the uniform distribution over the interval $[0, 1]$. This distribution has the rectangular pdf $h(\mu)$ shown in Figure 15.1. Two comments:

- The distribution for model parameters θ is not a distribution in the sense that was explained in Chapter 9. It is not connected to a random variable and does not model a real-world outcome of observations. Instead it captures subjective beliefs that the modeler has about how the true distribution P_{X_i} of data points *should* look like. It is here that subjectivistic aspects of "probability" intrude into an otherwise classical-frequentist picture.

- Each parameter vector $\theta \in \mathbb{R}^K$ corresponds to one specific pdf $p_X(\theta)$, which in turn represents one possible candidate distribution $\hat{P}_X$ for empirical observation values $X_i(\omega)$. A distribution over parameter vectors $\theta \in \mathbb{R}^K$ is thus a distribution over distributions. It is called a *hyperdistribution*.

According to (15.2), the pdf $p_X(\theta)$ gives rise to a sample distribution with pdf $p_{\bigotimes X}(\theta) : \mathbb{R}^N \to \mathbb{R}^{\geq 0}$ on the N-dimensional value space of data samples by simply multiplying the N values $p_X(\theta)(x_i)$. We write $p_{\bigotimes X}(D \,|\, \theta)$ to denote the pdf value $p_{\bigotimes X}(\theta)((x_1, \ldots, x_N)) = p_{\bigotimes X}(\theta)(D)$ of $p_{\bigotimes X}(\theta)$ on a particular training data sample D.

Summarizing:

- The pdf $h(\theta)$ encodes the modeler's prior beliefs about how the parametrized distribution $p_X(\theta)$ should look like. Parameters θ where $h(\theta)$ is large correspond to data distributions that the modeler a priori finds more plausible. The distribution represented by $h(\theta)$ is a hyperdistribution and it is often called the (*Bayesian*) *prior*.

- If θ is fixed, $p_{\bigotimes X}(D \,|\, \theta)$ can be seen as a function of data vectors D. This function is a pdf over the training sample data space. For each possible training sample $D = (x_1, \ldots, x_N)$ it describes how probable this particular outcome is, assuming the true distribution of X is $p_X(\theta)$.

- If, conversely, D is fixed, then $p_{\bigotimes X}(D \,|\, \theta)$ can be seen as a function of θ. Seen as a function of θ, $p_{\bigotimes X}(D \,|\, \theta)$ is *not* something like a pdf over θ-space. Seen as a function of θ, $p_{\bigotimes X}(D \,|\, \theta)$ is called a *likelihood function* — given data D, it reveals certain models θ as being more likely than others. A model (characterized by θ) "explains" given data D better if $p_{\bigotimes X}(D \,|\, \theta)$ is higher.

We thus have two sources of information about the sought-after, unknown true distribution $p_X(\theta)$: the *likelihood* $p_{\bigotimes X}(D \,|\, \theta)$ of θ given data, and the prior plausibility encoded in $h(\theta)$. These two sources of information are independent of each other: the prior plausibility is settled by the modeler *before* data have been observed, and should not be informed by data. Because the two sources of information come from "independent" sources of information (belief and data), it makes sense to combine them by multiplication and consider the product

$$p_X(D \,|\, \theta)\; h(\theta).$$

This product combines the two available sources of information about the sought-after true distribution $p_X(\theta)$. When data D are given, this product is a function of model candidates θ. High values of this product mean that a candidate model θ is a good estimate, low values mean it's bad — in the light of both observed data and prior assumptions.

With fixed D, the product $p_{\bigotimes X}(D \,|\, \theta)\; h(\theta)$ is a non-negative function on the N-dimensional parameter space $\theta \in \mathbb{R}^N$. It will not in general integrate to unity and thus is not a pdf. Dividing this product by its integral however gives a pdf, which we denote by $h(\theta \,|\, D)$:

$$h(\theta \,|\, D) = \frac{p_{\bigotimes X}(D \,|\, \theta)\; h(\theta)}{\int_{\mathbb{R}^K} p_{\bigotimes X}(D \,|\, \theta)\; h(\theta)\; d\theta} \tag{15.5}$$

The distribution on model parameter space represented by the pdf $h(\theta|D)$ is called the *posterior distribution* or simply the *posterior*. The formula (15.5) shows how Bayesians combine the subjective prior $h(\theta)$ with empirical information $p_{\bigotimes X}(D|\theta)$ to get a posterior distribution over candidate models. Comments:

- The posterior distribution $h(\theta \,|\, D)$ is the final result of a Bayesian model estimation procedure. It is a *probability distribution over candidate models*, which is a richer and often a more useful thing than the single model that is the result of a classical frequentist model estimation (like the sample mean from Equation 15.4).

- Here I have considered real-valued distributions that can be represented by pdfs throughout. If some of the concerned distributions are discrete or cannot be represented by pdfs for some reason, one gets different versions of (15.5).

- If one wishes to obtain a single, definite model estimate from a Bayesian modeling exercise, a typical procedure is to compute the mean value of the posterior. The resulting model $\hat{\theta}$ is called the *posterior mean estimate*

$$\hat{\theta} = \int_{\mathbb{R}^K} \theta \, h(\theta \,|\, D) \, d\theta.$$

- Compared to classical-frequentist model estimation, generally Bayesian model estimation procedures are computationally more expensive and also more difficult to design properly, because one has to invest some thinking into good priors. With diligently chosen priors, Bayesian model estimates may give far better models than classical-frequentist ones, especially when sample sizes are small.

- If one abbreviates the normalization term $\int_{\mathbb{R}^K} p_{\bigotimes X}(D \,|\, \theta) \, h(\theta) \, d\theta$ in (15.5) by $P(D)$ ("probability of seeing data D"), the formula (15.5) looks like a version of Bayes' rule that we know from Equation 7.6:

$$h(\theta \,|\, D) = \frac{p_{\bigotimes X}(D \,|\, \theta) \, h(\theta)}{P(D)}, \tag{15.6}$$

which is why the this entire approach is called "Bayesian". Note that while (7.6) is a *theorem* that can be proven from the axioms of classical probability theory, (15.6) is a *definition* (of $h(\theta \,|\, D)$).

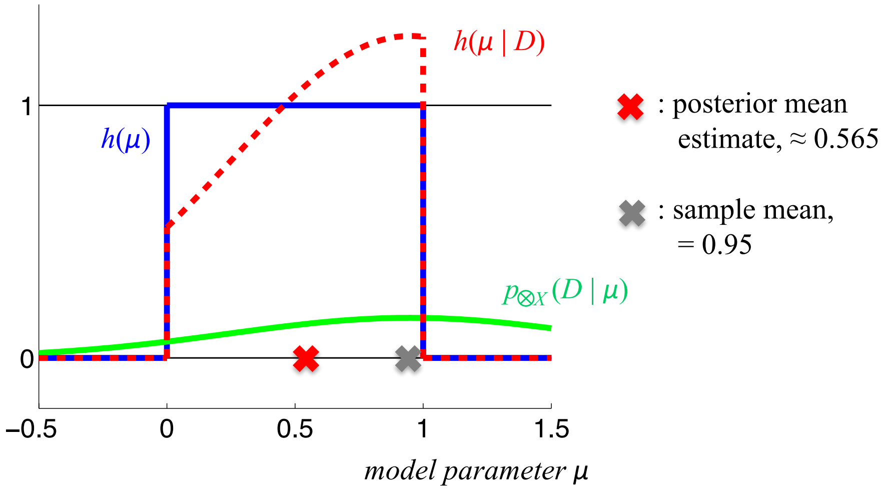


Figure 15.1: Bayesian model estimation. Compare text.

Let us conclude this section with a workout of our simple demo example. For the given sample $D = (0.9, 1.0)$, the likelihood function becomes

$$\begin{aligned} p_X(D \,|\, \mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(0.9-\mu)^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(1.0-\mu)^2}{2}\right) \\ &= \frac{1}{2\pi} \exp\left(-0.905 + 1.9\,\mu - \mu^2\right) \end{aligned}$$

The green line in Figure 15.1 gives a plot of $p_X(D \,|\, \mu)$, and the red broken line a plot of the posterior distribution $h(\mu \,|\, D)$. A numerical integration of $\int_{\mathbb{R}} \mu \; h(\mu \,|\, D) \; d\mu$ yields a posterior mean estimate of $\hat{\theta} \approx 0.565$. This is quite different from the sample mean 0.95, revealing the strong influence of the prior distribution on possible models.

Chapter 16

Some practical distributions

This section is adapted from the 2015 and 2016 PSM lecture notes written by Adalbert Wilhelm.

In this section we describe a few distributions which arise commonly in application scenarios. They have standard names and one should just know them, and under which conditions they arise.

16.1 Some standard discrete distributions

16.1.1 Bernoulli distribution

The Bernoulli distribution always arises when one deals with observations that have only two possible outcomes, like

- tail – head
- success – failure
- survival – death
- female – male
- U.S. citizen – no U.S. citizen
- retired – not retired
- correct – incorrect
- pass – fail
- $0 - 1$

Formally, this leads to a two-element sample space $S = \{s_1, s_2\}$ equipped with the power set σ-field, on which a Bernoulli distribution is defined by its pmf:

Definition 16.1.1 *The distribution of the random variable $X : \Omega \to \{s_1, s_2\}$ with probability mass function*

$$p(s_i) = \begin{cases} 1-q & \text{for } i = 1 \\ q & \text{for } i = 2 \end{cases}$$

is called a Bernoulli distribution *with* success parameter q*, where* $0 \geq q \geq 1$*.*

16.1.2 Binomial distribution

The binomial distribution describes the counts of successes if a binary-outcome “Bernoulli Experiment” is repeated N times. For a simple example, consider a gamble where you toss a coin N times, and every time the head comes up, you earn a Dollar (not Euro; gambling is done in Las Vegas, not Vegesack). What is the distribution of Dollar earnings from such N-repetition games, if the coin comes up with head ($= s_2$; outcome tail is s_1) with a success probability of q? Clearly the range of possible earnings goes from 0 to N Dollars. These earnings are distributed according to the *binomial distribution*:

Definition 16.1.2 *Let N be the number of trials of independent Bernoulli experiments with success probability q in each trial. The distribution of the number of successes is called the* binomial distribution *with parameters N and q and its pmf is given by*

$$p(s) = \binom{N}{s} q^s (1-q)^{N-s} = \frac{N!}{s!\,(N-s)!} q^s (1-q)^{N-s}, \quad s = 0, 1, 2, ..., N.$$

We write $Bi(N, q)$ to denote the binomial distribution with parameters N and q.

The factor $\binom{N}{s}$ is called *binomial coefficient*. Figure 16.1 shows the pmf of the binomial distribution $Bi(10, 0.25)$.

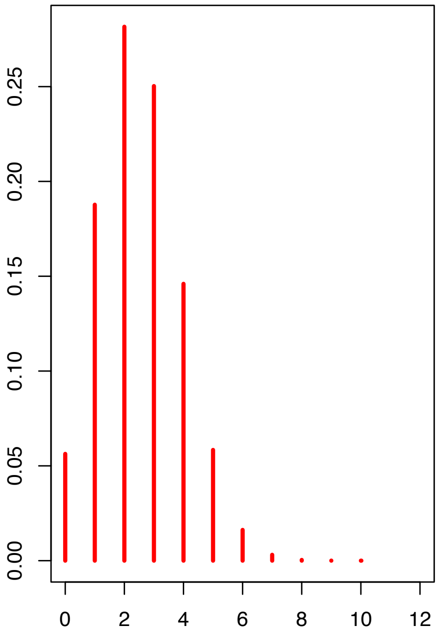


Figure 16.1: The pmf of $Bi(10, 0.25)$. Figure taken from A. Wilhelm’s PSM lecture notes.

A note on notation: it is customary to write

$$X \sim Bi(10, 0.25)$$

as a shorthand for the statement “X is distributed according to $Bi(10, 0.25)$”.

16.1.3 Poisson distribution

This distribution is defined for $S = \{0, 1, 2, \ldots\}$. $P_X(s)$ describes the probability that a particular kind of event occurs s times within a given time interval. Examples (except the last one taken from `https://en.wikipedia.org/wiki/Poisson_distribution`): $P_X(s)$ might be the probability that

- s meteorites impact on the earth within 100 years,
- a call center receives s calls in an hour,
- a block of uranium emits s alpha particles in a second,
- s patients arrive in an emergency ward between 10 and 11 pm,
- a piece of brain tissue sends s neural spike signals within a second.

Similarly, instead of referring to a time interval, s may count spatially or otherwise circumscribed events, for instance

- the number of dust particles found in a milliliter of air,
- the number of diamonds found in a ton of ore,
- the number of adventures a human experiences in his/her lifetime.

The expected number of events $E[X]$ is called the *rate* of the Poisson distribution, and is commonly denoted by λ. The pmf of a Poisson distribution with rate λ is given by

$$p(s) = \frac{\lambda^s \, e^{-s}}{s!}. \tag{16.1}$$

Figure 16.2 depicts the pmf's for three different rates.

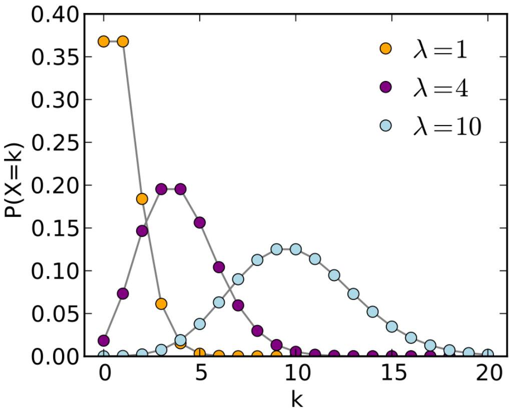


Figure 16.2: The pmf of the Poisson distribution for various values of the parameter λ. The connecting lines bewteen the dots are drawn only for better visual appearance (image source: `https://commons.wikimedia.org/wiki/File:Poisson_pmf.svg`).

16.2 Some standard continuous distributions

16.2.1 The uniform distribution

We don't need to make a big fuzz about this. If $I = [a_1, b_1] \times \ldots \times [a_n, b_n]$ is a n-dimensional interval in $\mathbb{R}^n$, the uniform distribution on I is given by the pdf

$$p(x) = \begin{cases} \frac{1}{(b_1 - a_1) \cdot \ldots \cdot (b_n - a_n)} & \text{if } x \in I \\ 0 & \text{if } x \notin I. \end{cases} \tag{16.2}$$

16.2.2 The exponential distribution

This distribution is defined for $S = [0, \infty)$ and could be paraphrased as "the distribution of waiting times until the next of these things happens". Consider any of the kinds of temporal events listed for the Poisson distribution, for instance the event "meteorite hits earth". The exponential distribution characterizes how long you have to wait for the next impact, given that one impact has just happened. Like in the Poisson distribution, such random event processes have an average *rate* events / unit reference time interval. For instance, meteorites of a certain minimum size hit the earth with a rate of 2.34 per year (just guessing). This rate is again denoted by λ. The pdf of the exponential distribution is given by

$$p(x) = \lambda\, e^{-\lambda x} \quad \text{(note that } x \geq 0). \tag{16.3}$$

It is (almost) self-explaining that the expectation of an exponential distribution is the reciprocal of the rate, $E(X) = 1/\lambda$. Figure 16.3 shows pdf's for some rates λ.

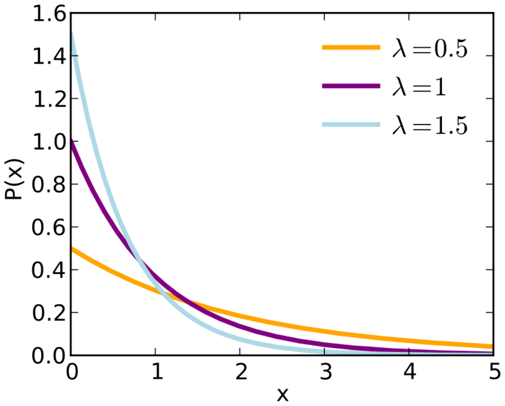


Figure 16.3: The pdf of the exponential distribution for various values of the parameter λ (image source: `https://commons.wikimedia.org/wiki/File:Exponential_pdf.svg`).

16.2.3 The normal distribution

Enter the queen of distributions! Bow in reverence! I am sure you know her from the media and facebook... For royal garments, as everybody knows, she wears the pdf

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{16.4}$$

which is fully specified by its mean μ and standard deviation (square root of variance) σ, has the famous bell shape with μ being the location of the maximum and $\mu \pm \sigma$ being the locations of the zeros of the second derivative (Fig. 16.4).

Make sure you are aware of the difference between the normal distribution, which is a function from the Borel σ-field $\mathcal{B}$ to $[0, 1]$ and denoted by $\mathcal{N}(\mu, \sigma^2)$, versus its pdf, which is a function from the reals to the non-negative reals and denoted by p. The normal distribution with zero mean and unit variance, $\mathcal{N}(0, 1)$, is called the *standard normal distribution*. The normal distribution is also called *Gaussian distribution* or simply *Gaussian*.

The normal distribution has a number of nice properties that very much facilitate calculations and theory. Specifically, the sum of normal distributed, independent RVs is again normal distributed, and the variance of the sum is the sum of variances:

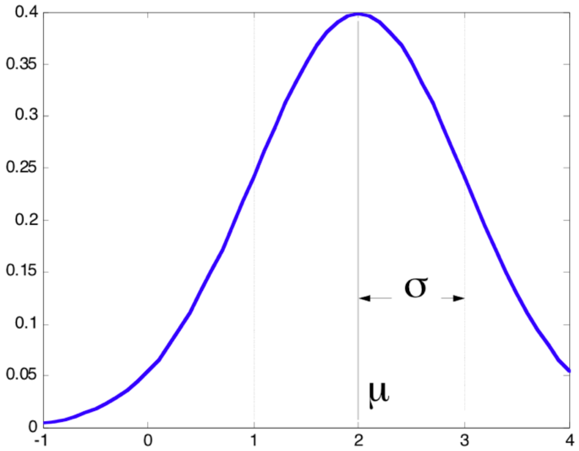


Figure 16.4: pdf of a normal distribution with mean 2 and standard deviation 1.

Proposition 16.2.1 *Let $X, Y : \Omega \to \mathbb{R}$ be two independent, normally distributed RVs with means μ and ν and variances σ^2 and τ^2. Then the sum $X + Y$ is normally distributed with mean $\mu + \nu$ and variance $\sigma^2 + \tau^2$.*

The majestic power of the normal distribution, which makes her reign almost universally over almost all natural phenomena, comes from one of the most central theorems of probability theory, the *central limit theorem.* It is stated in textbooks in a variety of (not always exactly equivalent) versions. It says, in brief, that one gets the normal distribution whenever random effects of many independent small-sized causes sum up to large-scale observable effects. The following definition makes this precise:

Definition 16.2.1 *Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of independent, real-valued, square integrable random variables with nonzero variances $Var(X_i) = E[(X_i - E[X_i])^2]$. Then we say that* the central limit theorem holds for $(X_i)_{i \in \mathbb{N}}$ *if the distributions P_{S_n} of the standardized sum variables*

$$S_n = \frac{\sum_{i=1}^n (X_i - E[X_i])}{\sigma\left(\sum_{i=1}^n X_i\right)} \tag{16.5}$$

converges weakly to $\mathcal{N}(0, 1)$.

Explanations:

- A real-valued random variable with pdf p is *square integrable* if its second moment, that is the integral $E[X^2] = \int_{\mathbb{R}} x^2\, p(x) dx$ is finite.
- If $(P_n)_{n \in \mathbb{N}}$ is a sequence of distributions over $\mathbb{R}$, and P a distribution (all over the same measure space $(\mathbb{R}, \mathcal{B})$), then $(P_n)_{n \in \mathbb{N}}$ is said to *converge weakly* to P if

$$\lim_{n \to \infty} \int f(x)\, P_n(dx) = \int f(x) P(dx) \tag{16.6}$$

for all continuous, bounded functions $f : \mathbb{R} \to \mathbb{R}$. You will find the notation of these integrals unfamiliar, and indeed you see here cases of *Lebesgue integrals* – a far-reaching generalization of the *Riemann integrals* that you know. Lebesgue integrals can deal with a far greater range of functions than the Riemann integral. Mathematical probability theory is formulated exclusively with the Lebesgue integral. We cannot give an introduction to Lebesgue integration theory in this course.

> Therefore, simply ignore the precise meaning of "weak convergence" and take home that sequences of distributions are required to converge to a target distribution in some subtly defined way.

A sequence $(X_i)_{i\in\mathbb{N}}$ of random variables (or, equivalently, its associated sequence of distribution (P_{X_i})) obeys the central limit theorem under rather weak conditions – or in other words, for many such sequences the central limit theorem holds.

A simple, important class of (X_i) for which the central limit theorem holds is obtained when the X_i are identically distributed (and, of course, are independent, square integrable and have nonzero variance). Notice that regardless of the shape of the distribution of each X_i, the distribution of the normalized sums converge to $\mathcal{N}(0,1)$!

The classical demonstration of the central limit theorem is the *Galton board*, named after Sir Francis Galton (1822–1911), an English multi-scientist. The idea is to let little balls (or beans, hence this device is sometimes called "bean machine") trickle down a grid of obstacles which randomly deflect the ball left or right (Figure 16.5). It does not matter how, exactly, these deflections act — in the simplest case, the ball is just kicked right or left by one space grid unit with equal probability. The deeper the trickling grid, the closer will the resulting distribution be to a normal distribution. A nice video can be watched at `https://www.youtube.com/watch?v=PM7z_03o_kk`.

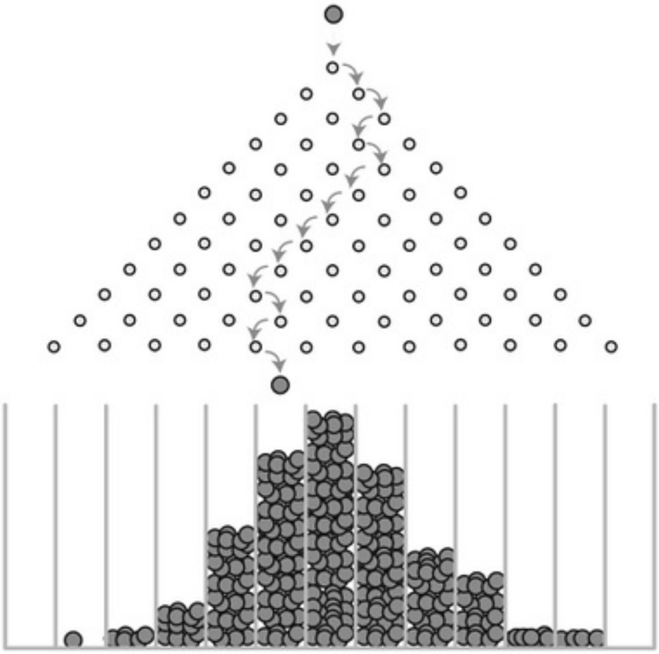


Figure 16.5: The Dalton board. Compare text for explanation. Figure taken from `https://janav.wordpress.com/2013/09/26/power-law/`.

However, this simple case does not explain the far-reaching, general importance of the central limit theorem (rather, property). In textbooks one often finds statements like, "if the outcomes of some measurement procedure can be conceived to be the combined effect of many independent causal effects, then the outcomes will be approximately normal distributed". The "many independent causal effects" that are here referred to are the random variables (X_i); they will typically not be identically distributed at all. Still the central limit theorem holds under mild assumptions. Intuitively, all that one has to require is that none of the individual random variables X_i dominates all the others – the effects of any single X_i must asymptotically be "washed out" if an increasing number of other $X_{i'}$ is entered into the sum variable S_n. In mathematical textbooks on probability you may find numerous mathematical conditions which amount to this "washing out". A special case that captures many real-life cases is the condition that the X_i are uniformly bounded, that is, there exists some $b > 0$ such that $|X_i(\omega)| < b$ for all i and ω. However, there exist much more general (nontrivial to state) conditions that likewise imply the central limit theorem. For our purposes, a good enough take-home message is

> *if (X_i) is a halfway reasonably behaved sequence of numerical RV's, then the normalized sums converge to the standard normal distribution.*

As we will see in Part 2 of this lecture, the normal distribution plays an overwhelming role in applied statistics. One often has to actually compute integrals of the pdf (16.4):

$$\textbf{Task:}\text{ compute the numerical value of } \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

There is no closed-form solution formula for this task. Instead, the solution is found in a two-step procedure:

1. Transform the problem from its original version $\mathcal{N}(\mu, \sigma^2)$ to the standard normal distribution $\mathcal{N}(0,1)$, by using

$$\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} dx. \qquad (16.7)$$

   In terms of probability theory, this means to transform the original, $\mathcal{N}(\mu, \sigma^2)$-distributed RV X to a $\mathcal{N}(0,1)$-distributed variable $Z = (X - \mu)/\sigma$. (The symbol Z is often used in statistics for standard normal distributed RVs.).

2. Compute the numerical value of the r.h.s. in (16.7) by using the cumulative density function of $\mathcal{N}(0,1)$, which is commonly denoted by Φ:

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x)^2}{2}} dx = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma}).$$

   Since there is no closed-form solution for calculating Φ, in former times statisticians found the solution in books where Φ was tabulated. Today, statistics software packages call fast iterative solvers for Φ.

16.3 ... and many more

The few common, named distributions that I displayed in this section are only meant to be illustrative picks from a much, much larger reservoir of well-known, completely analyzed, tabulated, pre-computed, and individually named distributions. The online book "Field Guide to Continuous Probability Distributions" by G. E. Crooks Crooks [2017] attempts a systematic overview. You should take home the following message:

- In 100 years or so of research, statisticians have identified hundreds of basic mechanisms by which nature generates random observations. In this chapter we only looked at only two of them – (i) intermittend rare “impact events” coming from large numbers of independent sources which hit some target system with a mean frequency λ, giving rise to Poisson and exponential distributions; and (ii) stochastic physical measurables that can be understood as the additive effect of a large number of different causes, which leads to the normal distribution.

- One way of approaching a statistical modeling task for a target distribution P_X is to

  1. first analyze and identify the nature of the physical (or psychological or social or economical...) effects that give rise to this distributions,
  2. then do a literature search (e.g. check out what G. E. Crooks says) or ask a statistics expert friend which known and named distribution is available that was tailored to capture exactly these effects, – which will likely give you a distribution formula that is shaped by a small number of parameters θ,
  3. then estimate θ from available observation data, getting a distribution estimate $\hat{\theta}$, and
  4. use the theory that statisticians have developed in order to calculate confidence intervals (or similar accuracy tolerance measures) for $\hat{\theta}$, which
  5. finally allows you to state something like, *“given the observation data, with a probability of 0.95, the true distribution θ^{true} differs from the estimate $\hat{\theta}$ by less than 0.01 percent.”*

- In summary, a typical professional statistical modeling project starts from well-argued assumptions about the type of the true distribution, then estimates the parameters of the distribtution, then reports the estimate together with a quantification of the error bound or confidence level (or the like) of the estimate.

- Professionally documented statistical analyses will *always* state not only the estimated model, but also in addition quantify how accurate the model estimate is. This can take many forms, like error bars drawn around estimated parameters or stating “significance levels”.

- If you read a report that only reports a model estimate, without any such quantification of accuracy levels, then this report has not been written by a professional statistician. There are two kinds of such reports. Either the author was ignorant about how to carry out a statistical analysis – then trash the report and forget it. Or the report was written by a machine learner in a task context where accuracy levels are not important and/or cannot be technically obtained because it is not possible to identify the distribution kind of the data generating system.

Chapter 17

The very big picture – what probability theory is good for

Probability theory is what the mathematicians give us. It is a universal mathematical modeling language for phenomena that have some aspects of randomness. This powerful tool is used for different purposes in different user communities. As I see it, there are four major user groups that have distinctly different modeling objectives, which in turn have led to distinctly different branching extensions of probability theory:

Natural scientists use the formalism of probability theory "just to model" physical, chemical or biological systems. These models are *analytical* models — which means that the subformulas appearing in a probabilistic model of a real-world system should correspond 1–1 to physical substructures / forces / quantities of the modeled system. An analytical model should provide a truthful and revealing account of the inner workings of the system that is modeled — just like the plan of an architect, or the blueprint of an airplane design should give a correct mirror image of the real object.

Natural scientists love to describe their target systems with ordinary differential equations (ODEs) or partial differential equations (PDEs). ODEs and PDEs are deterministic. When the target system exhibits some randomness, the DEs must be augmented by randomness too. This has led to *stochastic differential equations*, whose solutions are stochastic processes.

Stochastic differential equations describe the "ground truth" of stochastic physical systems in all detail, but they are difficult to work with and difficult to interpret and analyze. In many cases, a more summary model of the evolution of a stochastic system is more insightful. It is often possible to describe a stochastic system as an evolution of probability distributions. When at time t_0 the system's state x is known to be distributed according to a distribution $P_X(t_0)$, the distributions at future times $t > t_0$ can often be determined as a evolution starting from $P_X(t_0)$ ruled by a *deterministic* evolution operator. Such deterministic evolutions of probability distributions can be described by ODEs acting on the parameters of $P_X(t)$ if these distributions are parametric; if they are non-parametric but admit a pdf $p_X(t)$, the evolution of $p_X(t)$ can often be captured by a PDE. The heat equation, which describes the propagation of temperature gradients in physical media, is an example of this latter kind. See the Wikipedia articles on "Master equation" and "Fokker-Planck equation" for more. The formalism of quantum mechanics follows related ideas.

The entire discipline of *thermodynamics* and its generalization and abstraction, *statistical physics* is concerned with properties of systems composed of very large numbers of interacting subsystems (like a flask of air containing zillions of gas molecules) whose outwardly observable properties (like temperature or pressure) result from averaging over zillions of micro-interactions. Such systems can dramatically change their properties when some external *control parameter* crosses a critical value. For example, a bottle of water freezes when the temperature falls below zero; or a surface patch on a hard drive can become permanently magnetic if it is exposed to an external magnetic field of a certain superthreshold strength. The study of such *phase transitions* — which greatly shape the face of physical reality! — has spurred a rich body of original mathematical tools developed in statistical physics.

Ever since powerful digital computers became available, natural scientists have begun to *simulate* natural systems on computers, in order to understand and predict emergent properties. When the system under consideration is determined by the interaction of many small subsystems, a detailed simulation on the microlevel is infeasible. We owe to physics a number of *Monte Carlo* simulation methods which allow one to derive approximate conclusions from partial, computationally feasible simulations of such systems. These methods have been adopted outside the natural sciences and are particularly useful in numerically solving complex optimization problems.

Many of my colleagues from the natural sciences at Jacobs University are particularly strong in stochastic modeling. If you are interested in the sciences, you will find ample opportunities for theses projects when you knock at their doors. In this course we will not study natural science use-cases and methods. In my lecture notes of the bygone graduate course "Algorithmical and Statistical Modeling" you will find an accessible presentation of some Monte Carlo methods (`http://minds.jacobs-university.de/uploads/teaching/lectureNotes/LN_AlgMod.pdf`, Sections 4 and 5).

Signal processing and control engineers see randomness mostly as an enemy and call it *noise*. Being engineers they wish to be the rulers of their systems, and being mathematicians they know that the most biting laws for ruling reality are provided by linear algebra and its infinite-dimensional sister, *functional analysis*. For more than a century they have been growing and securing their kingdom in a world haunted by noise, building an eminent arsenal of linear methods to quantify, measure, and eliminiate noise. With their *spectral methods* they can identify and cancel many components of noise in signals (and images, by the way) and with *Kalman filters* they can gain access to a system's clean state that is hidden under a cover of noise. In my lecture notes of a legacy graduate course on ML you can find a section on *adaptive filtering* (`http://minds.jacobs-university.de/uploads/teaching/lectureNotes/LN_ML_Fall11.pdf`, Section 6).

Furthermore, in a quite different and thoroughly nonlinear mathematical spirit, we owe to signal processing engineers the inception of *information theory*. This mathematical theory has many links to the physical theory of thermodynamics and statistical physics (discussion in `https://en.wikipedia.org/wiki/Entropy_in_thermodynamics_and_information_theory` but develops a specific perspective on and tools for analyzing information processing dynamics. Today, information theory paired with methods from statistical physics are used throughout all sciences.

Machine learning in its modern form aims at modeling real-world systems, just like in the natural sciences. But the motivation is different. Machine learners are application oriented, they want to *exploit* their models. ML models of pieces of reality must *function* well in their respective application context. It is not necessary that they are veridical (from latin, "saying the truth"). The inner workings of an ML model need not mirror the inner workings of the modeled system. Machine learning models are thus almost always *blackbox* models. A blackbox model of some real-world system just captures the externally observable input-output behavior of the modeled system, but it may use any internal math or algorithm tricks that its designer can think of. The quality of a ML model is assessed with regards how precisely it can capture the externally observable properties of the modeled system – that is, the distribution of RVs. The quality of an (analytical) science model is assessed in terms of how correctly it captures the internal workings of the target system. One could summarize this by saying that a ML model aims at modeling *data*, whereas a natural science model wants to model *data-generating mechanisms.*

This gives rise to an interesting pair of opposite findings.

Case 1: When a target system is simple and can be well isolated from its environment (that is, from impacts of "noise" which the natural scientists call "perturbations"), and if the analytical model indeed captures the underlying natural laws and structures, analytical models may yield close to perfect matches with reality. A superior example of this is the Newton model of gravitational forces which can yield super accurate predictions of planetary motion decades ahead. No blackbox ML model that would be trained numerically on tons of stellar observation data could match that.

Case 2: When the target system is nonlinear and complex, and cannot easily be isolated from its environment, analytical models can turn out to be much less accurate in their prediction of outwardly observable system behavior than ML models. The reason is that analytical models will typically incorporate many simplifying assumptions, which lead to a mismatch with reality that can become crippling when the target system is very nonlinear. A point in case are language models. The (analytical) neuroscience models of the language-generating circuits of the human brain cannot (at present) predict any serious portions of the structure of language, whereas ML language models have recently been refined to generate entire Wikipedia texts – at the html source code level [Graves, 2014]!

Figure 17.1 sketches how blackbox models (should) work. They are derived ("learnt", "estimated") from data emitted by the source system, and they should generate synthetic data that "fits" (has a similar distribution as) the real-world data. And that's it. The structure of the blackbox model need not correspond in any way with the source system – in the figure, this is a robot, while the model is a neural network whose structure has no correspondence whatsoever in the robot design.

Unconstrained by the veracity demands of analytical modeling, ML research has spun off a large number of formalisms to represent probability distributions, and learning algorithms to estimate them from data. At the present time, *neural networks* dominate the public perception of ML, but there are very many others. The diversity of formalisms makes it difficult for students to "learn ML", and ML textbooks (of which there are many good ones) focus on different selections of methods, according to the preferences of the author.

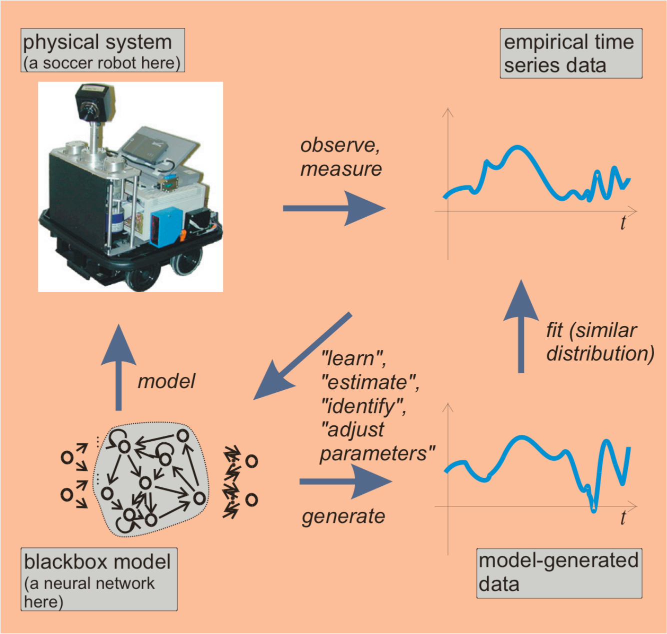


Figure 17.1: How blackbox models work. For explanation see text.

Statistics. The last group of probability theory end-users that I want to highlight are statisticians. The general objective of statistical analyses can be stated as, *find out from empirical data how optimal decisions can be made in a world full of uncertainty.* Decision making scenarios occur virtually everywhere where humans act, for instance, in economics ("buy or sell?"), human resources ("hire or fire?"), military ("attack or run?"), spacecraft start countdown ("ignite or abort?"), medicine ("surgery or pills?"), you name it. Besides finding out which decision to take, a secondary – and often primary – goal is to assess the degree of certainty of whether the suggested decision is truly the better one. Statistics thus has developed an arsenal of methods not only to estimate distributions, but also to estimate how reliable is the estimate. Students of the natural sciences get a tiny glimpse of this when they are taught to always draw *confidence intervals* into their diagrams. Students of the social sciences, economics and psychology get a fuller treatment. They have to take one or two full statistical methods courses, based on a heavyweight textbook, where they learn about the entire workflow: from acquiring data in the first place (how to avoid "biased" data sources or cope with biasedness in case it is unavoidable), to *planning* a data-based study (for example, how many questionnaires have to be filed in order to secure a given minimal level of confidence in the final recommendations for action arising from the study), to select adequate statistical models (what assumptions can be made about the data distribution), and finally, to the actual "mathworks". Statisticians have become more painfully aware than other probability end-users that there is not a unique, correct, best method for analysing data. This community thus has been the main driving force for the development of *estimation theory*, a branch in the great tree of probability which attempts to characterize the pros and cons of different methods for statistical model estimation – a meta-theory of probabilistic modeling if you wish.

Summarizing: natural scientists want to analyze and understand data generating systems; signal processing engineers want to cleanse information in data from noise on data;

machine learners want to capture the structure of the data generated by data-generating systems; and statisticians want to find the safest possible decision-making grounds in an uncertain world. All of these communities of thinkers and makers have created specific extensions to the core probability theory delivered by mathematicians. Several of these extensions have grown into scientific disciplines of their own standing. There are, of course, many mathematical crosslinks. Yet, the four communities that I listed pursue their researches in an amazingly strict mutual isolation and with much less cross-fertilization than one would expect. This is a historical and sociological fact which would deserve its own investigation.

Pure and basic probability theory is the common point of departure for all of these endeavours. The sweat (and swearings I might guess) that you have been shedding in this course up to this point will pay off if you venture into any of these four domains in your future professional life.

In the remaining two parts of this course, I will try to give you an insightful introduction to the basic ways of thinking in two of those fields, namely statistics and machine learning.

Chapter 18

Summary of Part I

We have arrived at the end of Part I. The purpose of this part was to give you a safe anchoring in the basic conceptualizations of probability theory. Here is a list of what I would consider the essential take-home messages. If you *understand* each of the following items and can confidently write down the associated basic formal expressions, you will have secured a true magic box of math keys that will open all modeling doors in the natural sciences, engineering, the social sciences, economics and machine learning.

Here is the checklist:

- The "universe" or "population" Ω is the mathematical abstraction of a piece of reality. Since the only aspect of reality which is relevant for empirical modeling is its capacity to provide opportunities for making observations, the mathematical model of a piece of reality is just a plain set of what I called "observation opportunities" $\omega \in \Omega$, and what educated modelers call *elementary events*.
- Random variables are the mathematical abstraction of observation procedures.
- Arbitrarily many RVs can be added to a growing model when a modeler starts modeling a piece of reality. This reflects that in real life, the same situation can be observed and described in arbitrarily many ways.
- Each RV X_i comes with its own sample space S_i.
- Any family $(X_i)_{i \in I}$ of RVs can be grouped into a single new RV $\bigotimes X_i : \Omega \to \prod_i S_i$ whose sample space is the product of the individual sample spaces S_i.
- Measurement values in S delivered by a RV $X : \Omega \to S$ can be further processed, for instance by "feature extraction" operations $f : S \to S'$, obtaining new RVs $f \circ X : \Omega \to S'$.
- There is no escape from σ-fields if you want to understand probability in any serious way. σ-Fields arise automatically and immediately when you want to combine information by NOT, OR and AND, like in: "the probability that it will NOT rain is...", or "the probability that this patient has a brain lesion AND a vitamin B12 deficit is...". This leads to a structuring of a sample space S by declaring a σ-field $\mathcal{F}$ on it. Sample spaces *always* come in the structured form $(S, \mathcal{F})$. Pairs $(S, \mathcal{F})$ where S is any kind of non-empty set and $\mathcal{F}$ is a σ-field on S are called *measurable spaces.* Make sure you are very familiar with the defining properties of a σ-field.
- Two special cases of measurable spaces $(S, \mathcal{F})$, which cover almost all practical applications that you are likely to meet, are obtained when

- S is *discrete* (that is, finite or countably infinite) — then one standardly uses the powerset σ-field $\mathcal{F} = \text{Pot}(S)$,
- or when $S \subseteq \mathbb{R}^n$ is a continuously connected subset of the n-dimensional Euclidean space $\mathbb{R}^n$, typically $\mathbb{R}^n$ itself, or an n-dimensional interval within $\mathbb{R}^n$, or a linear subspace, or a manifold in $\mathbb{R}^n$. Then one uses the *Borel* σ-field $\mathfrak{B}$ on S, which is *generated* by the intervals within S (make sure you understand what it means to "generate" a σ-field).

- A RV $X : \Omega \to (S, \mathcal{F})$ induces a σ-field $X^{-1}(\mathcal{F}) = \{X^{-1}(A) \mid A \in \mathcal{F}\}$ on Ω. If a probabilistic model includes a family $(X_i)_{i \in I}$ of RVs, all of them together induce the σ-field $\{\bigotimes X_i^{-1}(A) \mid A \in \bigotimes \mathcal{F}_i\}$ on Ω.

- A complete probabilistic model of a piece of reality looks like this:

$$\bigotimes X_i : (\Omega, \mathfrak{A}, P) \to (\prod S_i, \bigotimes \mathcal{F}_i),$$

or, equivalently, for every used RV X_i, specify

$$X_i : (\Omega, \mathfrak{A}, P) \to (S_i, \mathcal{F}_i).$$

The σ-field $\mathfrak{A}$ must include $\{\bigotimes X_i^{-1}(A) \mid A \in \bigotimes \mathcal{F}_i\}$ in order to make all concerned RVs *measurable*. For most purposes it is good enough to identify $\mathfrak{A}$ with $\{\bigotimes X_i^{-1}(A) \mid A \in \bigotimes \mathcal{F}_i\}$.

- In the abstract mathematical theory of probability, any triple $(\Omega, \mathfrak{A}, P)$, where $(\Omega, \mathfrak{A})$ is a measurable space and P is a probability measure on $(\Omega, \mathfrak{A})$, is called a *probability space*.

- The one and only and fundamental way to make probability statements is to write, think and say

$$P(E) = a,$$

where $E \in \mathfrak{A}$ and $a \in [0, 1]$. Since E will always be specified through observation values, this fundamental form appears in practice mostly in the special format

$$P(X^{-1}(A)),$$

or in equivalent notations, $P(X \in A)$ or $P_X(A)$.

- The probability measure P_X on the sample space $(S, \mathcal{F})$ of X, defined by $P_X(A) = P(X \in A)$, is called the *distribution* of X.

- Many continuous distributions P_X on $\mathbb{R}^n$ can be represented by a pdf $p_X : \mathbb{R}^n \to \mathbb{R}^{\geq 0}$, which is a function satisfying $P_X(A) = \int_A p_x(x)\, dx$ for all n-dimensional intervals $A \subseteq \mathbb{R}^n$.

- In a probability model $\bigotimes X_i : (\Omega, \mathfrak{A}, P) \to (\prod S_i, \bigotimes \mathcal{F}_i)$, the distribution $P_{\bigotimes X_i}$ is the *joint distribution* of all used RVs X_i. This joint distribution is the major target object in all probabilistic modeling.

- Given the joint distribution $P_{X_1 \otimes \cdots \otimes X_k}$ of k RVs X_i, *marginal distributions* $P_{X_{i_1} \otimes \cdots \otimes X_{i_l}}$ (where $J := \{i_1, \ldots, i_l\} \subset \{1, \ldots, k\}$) can be defined/calculated by "summing away" (or "integrating away" in the case of continuous distributions) the other RVs. For notational convenience let us assume that $J = \{1, \ldots, l\}$. Then in the case that all RVs are discrete,

$$P_{X_1 \otimes \cdots \otimes X_l}(A_1 \times \cdots \times A_l) = \sum_{s_{l+1} \in S_{l+1}, \ldots, s_k \in S_k} P_{X_1 \otimes \cdots \otimes X_k}(A_1 \times \cdots \times A_l \times \{s_{l+1}\} \times \ldots \times \{s_k\}),$$

and in the case where all RVs are real-valued and the joint distribution has a pdf $p_{X_1 \otimes \cdots \otimes X_k}$,

$$P_{X_1 \otimes \cdots \otimes X_l}(A_1 \times \cdots \times A_l) = \int_{x_1 \in A_1, \ldots, x_l \in A_l, x_{l+1} \in \mathbb{R}, \ldots, x_k \in \mathbb{R}} p_{X_1 \otimes \cdots \otimes X_k}(\mathbf{x}) \, d\mathbf{x}.$$

This can be extended to marginals in joint distributions of infinitely many RVs, but we did not cover that in our lecture.

- The *conditional distribution* $P_{X | Y \in B}$ Is defined by

  $$P_{X | Y \in B}(A) = \frac{P(X \in A, Y \in B)}{P(Y \in B)}$$

  in case that $P(Y \in B) > 0$. The conditional probability $P_{X | Y \in B}(A)$ is commonly written as $P(X \in A \,|\, Y \in B)$. In continuous joint distributions of X and Y with a joint pdf $p_{X,Y}$, conditional distributions can also be defined for conditions $Y = y$ although $P(Y = y)$ is zero. Provided that the pdf of the conditional distribution $P_{X|Y=y}$ satisfies certain minimal requirements (for instance, it suffices that it is continuous), then the conditional distribution of X given $Y = y$ has the pdf

  $$p_{X|Y=y}(x) = \frac{p_{X,Y}(x, y)}{\int p_{x,y}(x, y) \, dx}$$

  .

- A discrete distribution can always be specified by its pmf. A continuous distribution may or may not have a pdf. There are many more ways to represent distributions. For instance, the distribution $P_{\bigotimes X_n}$ of a homogeneous Markov Chain $(X_n)_{n \in \mathbb{N}}$ is represented by the starting probability vector π_0 and the Markov transition matrix M. In a sense, every machine learning formalism comes with its own format for representing distributions (often only implicitly).

- A collection of k RVs $X_1, \ldots, X_k$ with values in $(S_i, \mathcal{F}_i)$ (where $i = 1, \ldots, k$) is *independent* if the joint distribution of the RVs is the product of their marginal distributions, that is,

  $$P(X_1 \in A_1, \ldots, X_k \in A_k) = P(X_1 \in A_1) \cdot P(X_2 \in A_2) \cdot \ldots \cdot P(X_k \in A_k)$$

  for all $A_i \in \mathcal{F}_i$. More generally, a family $(X_i)_{i \in I}$ of RVs is independent if all finite subfamilies are independent.

- A collection of k RVs $X_1, \ldots, X_k$ with values in $(S_i, \mathcal{F}_i)$ is *conditionally independent* given $Y_1 = B_1, \ldots, Y_l = B_l$ if

  $$\begin{aligned} & P(X_1 \in A_1, \ldots, X_k \in A_k \,|\, Y_1 = B_1, \ldots, Y_l = B_l) \\ = \; & P(X_1 \in A_1 \,|\, Y_1 = B_1, \ldots, Y_l = B_l) \cdot \ldots \cdot P(X_k \in A_k \,|\, Y_1 = B_1, \ldots, Y_l = B_l). \end{aligned}$$

Detecting conditional independencies in complex joint distributions is a key enabling factor for entire branches of machine learning. This is because computing probabilities in multi-RV distributions in general requires summation or integration over exponentially blowing-up sized summation or integration spaces, which is infeasible. If concerned RVs are (conditionally) independent, these summations/integrations reduce to simple products of easy-to-compute marginal probabilities. One also says that joint distributions can be *factorized.* The branches of ML that use Markov Chains, hidden Markov models, Bayesian networks, or other *graphical networks* are practically useful only because they can exploit conditional independencies.

- Make sure you understand the differences between *uncorrelatedness* and *independence.*

- A stochastic process is a family $(X_t)_{t \in T}$ of RVs, all of which take values in the same space S, and where the set T is ordered and considered a set of time indices. A *path* (or *trajectory,* or *realization*) of a stochastic process is a sequence $(X_t(\omega))_{t \in T}$ of values in S. A (deep) result from probability theory, which we did not cover in the lecture, states that the distribution of a stochastic process is characterized by all the finite marginal distributions of the process, that is, by all distributions $P_{X_{t_1},\ldots,X_{t_k}}$ where $k > 0$.

- The *expectation* of a numerical (that is, vector-valued) RV X whose distribution has a pdf p is
$$E[X] = \int x \, p(x) \, dx.$$
It is often written as μ. This is very different from the *sample mean*
$$\frac{1}{N} \sum_{i=1,\ldots,N} X_i(\omega).$$
The expectation is a characteristic of an entire distribution - it is a single, fixed number (or vector). The sample mean, by contrast, is a random variable! The sample mean gives an *estimate* of the expectation, which is why one may write $\hat{\mu} = \frac{1}{N} \sum_{i=1,\ldots,N} X_i(\omega)$.

- Beyond the expectation one is interested in *higher-order moments* of the distribution of a real-valued RV $X : \Omega \to \mathbb{R}$. For $n > 1$ the n-th moment of X is $E[(X - E[X])^n]$. In statistics, where one often analyses distributions whose pdfs have nice analytical properties, a distribution can often be characterized up to negligible residual errors by giving the first few moments.

- There are two fundamentally different interpretations of the nature of *randomness*:
  - The *objectivist* (or *frequentist*) view of probability regards randomness as a measurable physical property of real-world systems. Probability is measured by averaging measurement outcome counts over (infinitely) repeated observations. The frequentist view is the most widespread one in mathematical textbooks on probability theory, and it has been worked out to an enormous depth.
  - From the beginnings of philosophical and scientific thinking about randomness, other interpretations of “probability” than the frequentist one have been explored and worked out. These *subjectivist* accounts have the aim to establish mathematically and conceptually sound systems of *rational reasoning* about

  uncertainty. Subjectivist accounts of probability can be considered a subfield of mathematical logic.

  - An intermediate case is found in what is called *Bayesian modeling*, a computational strategy of growing importance in machine learning. Bayesian model estimation allows one to merge subjective insight about the nature of the to-be-modeled distribution with objective information contained in samples.

- For both objectivist and subjectivists accounts of probability, the axioms of a σ-field plus the *Kolmogorov axioms*, which describe how probabilities are consistently assigned to the elements of a σ-field, yield the framework for the mathematical study of probability.

- The basic concepts and formalism of probability theory, which we introduced in Part I of this lecture, are extended and specialized (and, unfortunately, often written in other notations) in different ways in the natural sciences, the social sciences and psychology, economics, machine learning, or signal processing and control.

Part II

A short helicopter ride over the lands of statistics

In this part of our lecture notes I follow the leads of two sources. The first is the textbook *Introduction to Statistical Inference* by J. C. Kiefer Kiefer [1987] (I will call it "the Kiefer book"). This book, although it first appeared 30 years ago, is still in print — a classic reference that keeps being used as the main reference in statistics courses all over the world. In the case that you will in your future professional life be called to carry out statistical analyses, you will be well advised to buy a copy of the Kiefer book. The other source is the lecture notes of my JacobscColleague Adalbert Wilhelm. He wrote them for the 2015 and 2016 versions of this course, and he followed the Kiefer book.

The Kiefer book uses a notation that is in some places confusingly different from the notation that I use in these lecture notes. I will continue to use "my" symbols — they adhere to notational conventions in probability theory and machine learning. In Appendix B I give a contrastive summary of the disconsonant notation in the Kiefer book vs. these lecture notes.

The purpose of this short Part II is to afford you with an understanding of the statistics way of thinking, the core terminology in this field, and how it builds upon the concepts from the mathematical theory of probability. This condensed treatment cannot provide you with a working-level familiarity with the methods of statistics (to rise to that level one would have to invest an entire semester and a solid textbook), but it can help you to decide whether you will need or want to employ statistics methods in your professional future, and it will ease your access to that field in case you want to become seriously engaged in it.

Chapter 19

The mindset and basic terminology of statistics

Statistics is a decidedly application-oriented field. The core question of statistical analyses can be paraphrased as follows:

> *How can the information gleaned from empirical data be used to optimally guide decision-making, such that the possible benefits incurred by the decision are maximized (or equivalently, such that the possible negative effects incurred by the decision are minimized)?*

Some examples of statistical analysis scenarios will put flesh on this abstract characterization.

- A federal ministery of health must decide whether a new pharmaceutic drug will be approved for distribution. Possible benefits of the decision are a better health status of the country's population, possible negative effects may result from risky side-effects, ineffectiveness of the drug, or the additional financial load on public healthcare. Data come from clinical studies.

- A physicist has secured a research grant which buys him 48 hrs use of the CERN particle accelerator. Within this time he carries out a measurement suite that results in 500 TB of particle shattering data. He has to decide whether some effects that appear to shine up in the data warrant the announcement of a new elementary particle. Possible positive effects: fame and follow-up funding in case he announces that a new particle has been detected, and other researchers can subsequently confirm the finding. Possible negative effects: shame and drying-out funding in case he announces the new particle but this is later found erroneous. Or, alternatively, lifelong frustration in case he does not announce the finding of a new particle, but later others do find it and reap in the reputation.

- An insurance company wants to spare the expenses for human salespersons and designs an online questionnaire for potential customers, together with an evaluation algorithm which uses the customer's questionnaire response to issue a tailored insurance contract. The data for calibrating the questionnaire and the evaluation algorithm come from the company's previous customer history data. The decision made by the company here is very complex: it consists in the specific formulation of the questionnaire and the design of the algorithm. Possible benefits are future average financial gains, possible damage are future financial losses.

- No way to avoid this classical textbook example... here goes. A gambler wants to know whether the coin she is using is fair within an error tolerance of 0.1 percent. The decision is binary: juding the coin to be fair and therefore using it, versus judging it to be not fair and throwing it into the trashcan. Possible negative effects of misjudging the fairness of the coin are gambling losses (in case the gambler thinks the coin is fair but it isn't) or wasting the price for one coin (in case the gambler decides the coin is unfair and trashes it though in fact it is fair). Possible positive effects of correct decisions are earning gambling money with a fair coin, or avoiding gambling losses with an unfair coin. Data come from throwing the coin many times.

19.1 Informal overview of the core coordinates of statistics

The statistical way of dealing with such scenarios can be abstractly summarized in the following considerations:

- The general situation consists of a rational agent (helped by a statistician) who has to make a decision to take one of several possible actions.

- The action has consequences which can be measured on a scalar scale that one could interpret, depending on the specific scenario, as a scale of bad – good, punishment – reward, loss – gain, pain – pleasure, frustration – satisfaction, failure – success or the like. This quantification of the value of action consequences is called *loss*. By a general convention this loss is scaled to start at 0 (best, most desirable), with larger values indicating less desirable consequences. The loss might even reach positive infinity (like death, must be avoided by all means).

- How large the effectively incurred loss becomes after a decision has been made depends on how the world reacts to the decision. But, the world cannot be completely known or perfectly predicted, and it may even be inherently stochastic. The best one can do is to describe the relevant parts of reality by the distribution P_X of a random variable X. (Recall from Part I that this RV X can be a product of a large number of component variables, admitting arbitrarily rich descriptions of reality).

- Unfortunately, the true distribution P_X^{true} is unknown and unknowable. At this point, statistical modeling introduces a fundamental assumption. One assumes that the true distribution must be one of a *predetermined* collection of candidate distributions. Statistical modeling always starts from declaring which set of candidate distributions is taken into consideration for all subsequent arguments. Following Adi's lecture notes, I will use the symbol $\mathcal{P}$ for this candidate set. One hopes or believes or assumes that $P_X^{\text{true}} \in \mathcal{P}$.

- The loss that will become reality after a decision has been made therefore depends on the unknown true distribution P_X^{true}. For making a good decision (which will minimize somehow the expected loss) it is vital to get as much information about P_X^{true} as one can. This information is extracted from a *sample* $X(\omega)$.

- The final outcome of a statistical analysis is to derive a decision on the basis of a sample. This "derivation" takes the form of an algorithm which *computes* the decision from the sample. Once the algorithm is chosen, the decision follows mechanically from the available data. The art and science of statistics is to choose that algorithm (from a possibly wide range of candidate algorithms) which minimizes the expected loss.

- Thus, one can summarize the essence of statistical thinking like this: *be aware that many decision-making algorithms are potential candidates in a given decision situation, thus try to understand the consequences of each candidate algorithm for the ultimate loss, and pick the best one.*
- Statistical research has revealed that this programme is not as straightforward as it might at first seem. Among other, the following difficulties arise:
  - How can one procure something like a "collection of candidate decision-making algorithms" (needed to pick the best?) It turns out that the options to design such algorithms are as unbounded and ill-defined as the options that a music composer has to compose a symphony. As a consequence, the field of statistics has produced a large "zoo" of decision-making algorithms, together with attempts to get a systematic overview of fundamental types and subtypes of such algorithms – really, quite similar to the attempts of zoologist to order animals in a system of classes, tribes, and species. Textbooks of statistics thus will often have a number of main chapters each of which is devoted to a special class of decision making algorithms.
  - It is not clear what it means to "pick the best" decision algorithm from a collection of candidates. While "the best" algorithm should clearly be the one that minimizes the loss, it is unclear in what way the loss should be "minimized". Should it be minimal on average? or should decisions that lead to an especially high loss (like death) be prevented by all means, at the cost of maybe a worse average loss outcome? Are we allowed to factor in subjective beliefs about the true distribution, or should one strictly adhere to using only the information that is contained in the data? These questions are connected to the problem mentioned above of establishing an overview of possible candidate decision algorithms, because specific kinds of such algorithms will be most suitable for specific ways of how "minimizing loss" is understood.
  - It is clear that the quality of the final decision hinges on the quality of the sample data. Statistics is distinguished from other fields of applied probability in that it devotes a serious study to the quality of samples. Statistical research has developed techniques for assessing the quality of a sample, and for experimental data-generating designs that generate useful samples in the first place. This is different from, for example, machine learning where a sample is just "given". In statistics, samples can be *made.* Students of psychology, for instance, have to take a statistical methods course (or two) where they are trained to design an empirical experiment in a way that the sample information coming out of the experiment will (hopefully) be good enough to enable a decision which satisfies predefined quality criteria.

19.2 Core terminology and formalism

In this section I supply the essential formalism that is used in statistics to substantiate the grand picture outlined in the previous section. The basic formalism will look simple to you, but be aware that it reflects 100 years of struggling of great minds, and that it sets the the foundations as well as the limiting fencing walls for a mathematical discipline that has an enormous impact on decision-making in the sciences, economy, medicine and politics.

A cautionary note. Statisticians use words like "decision", "loss" and "risk" in ways that are related, but not identical, to the way how these words are used in machine learning and how I introduced them in the Machine Learning lecture (the recommended companion lecture to this PSM course). The machine learners have imported these words from the much more ancient fields of statistics, but placed them in the new context of (supervised) machine learning and this changed their formal definitions. I recommend that you forget (for the next few sections) how "decision", "loss" and "risk" were introduced in the ML lecture, and digest the following definitions with a blank mind. In my comments after Definition 19.3.2 below I explain the differences in terminology usage between ML and statistics.

19.2.1 Data and samples

We start with the data. It is formalized as a finite sample $X(\omega) = (X_1(\omega), \ldots, X_N(\omega))$, also written as $(x_1, \ldots, x_N)$ for convenience. All data points x_i come from the same sample space S_0, so the sample space for $X = \bigotimes_{i=1,\ldots,N} X_i$ is $S := S_0^N$. In other words, we are dealing with what we called a mathematical sample in Chapter 8. Again I want to emphasize that each X_i can itself be a richly structured source of information (for instance, embodied by a questionnaire or a company's customer profile), with S_0 being a complex product space.

The question of how to *design an experiment* such that the data generated by the experiment are useful for the decision task is an important issue in the professional practice of a statistician. However, I will not touch upon this issue in these lecture notes and instead, like the Kiefer book, simply assume X and S are given. If you are interested in getting a first impression of how this questions is approached, I can recommend the Wikipedia article `https://en.wikipedia.org/wiki/Design_of_experiments`.

19.2.2 Decision spaces

The next component we consider are the decisions. One assumes that a pre-defined collection of possible decisions is given to the statistician. We denote decisions by d and the collection of all possible decisions by $\mathcal{D}$. This *decision space* $\mathcal{D}$ can be finite or infinite. How the elements $d \in \mathcal{D}$ concretely look like depends on the specific situation – anything is admissible. From the mathematical and algorithmic perspective, $\mathcal{D}$ is just any set (equipped with a σ-field, but we skip that). Here are some examples for illustration:

1. For a coin-throwing gambler investigating her coin, relevant decision spaces might be $\mathcal{D} = \{$fair, unfair$\}$ or $\mathcal{D} = \{$fair with an error tolerance of 0.01%, fair with an error tolerance of 0.05%, unknown $\}$ or $\mathcal{D} = \{$will use this coin for at least 100 throws then see further, will not use this coin $\}$.

2. A physicist measuring the speed of light might want to use decision spaces like $\mathcal{D} = \mathbb{R}$ or $\mathcal{D} = \{[c - 0.1, c + 0.1] \mid 290,000,000 \leq c \leq 310,000,000\}$ or $\mathcal{D} = \{$true SoL deviates from 299,792,458 by at least 0.00000001%, true SoL is equal to 299,792,458 with a precision of 0.00000001%$\}$.

3. An insurance salesman might be faced with $\mathcal{D} = \{$sell contract to customer, don't sell$\}$.

It becomes clear from these examples that the word "decision" is understood in a wide sense – it may mean decisions-to-act (example 3) or decisions-to-judge (example 2) or decision-to-postpone-decision (last case in example 1) or, in fact, anything else.

19.2.3 The distribution space

One does not know what the true data distribution P_X^{true} is, but one knows that the potential good or bad effects of a decision depend on what the true distribution is. A statistician thus has to acknowledge that the world might offer different true distributions, and s/he must explore the consequences implied by different true distributions. But, one cannot scientifically explore something like "all possible true distributions" because this is an ill-defined concept. Therefore, statistical analysis begin by defining a mathematically circumscribed set of candidate distributions. We denote this set by $\mathcal{P}$.

Ideally the candidate set $\mathcal{P}$ should contain P_X^{true}. However, reality is more cunning than mathematics will ever be, thus in effect the assumption $P_X^{\text{true}} \in \mathcal{P}$ will almost certainly be wrong. The art of specifying $\mathcal{P}$ is thus to find a candidate set that has members which come very close to P_X^{true}. On the other hand, the mathematical format of $\mathcal{P}$ should be as simple as one can afford, in order to facilitate the derivation of subsequent formulas and algorithms. This is a delicate tradeoff situation. For example, one frequently follows the temptation of simplicity and establishes $\mathcal{P}$ as a set of normal distributions. This can be entirely appropriate, or it can be the road to hell ("statistics lie").

At this point it is interesting to comment on a difference in the ways of thinking in statistics vs. machine learning. In ML, modelers take pride in coming close to reality, which leads them to using very complex sets of candidate distributions – for instance, $\mathcal{P}$ in ML might be represented by a collection of neural networks with hundreds of thousands of parameters. Using such ultra-complex models of distributions bars the way to mathematical analysis and to understanding the nature and ultimate effects of what may go wrong (see, for instance, the vivid discussions on fooling trained neural networks [Nguyen et al., 2014], — google "attack neural network"). In statistics, the emphasis is laid on a formal analysis of the consequences of decisions, which leads to a preference for analytically tractable distribution models assembled in $\mathcal{P}$. — Another instructive example: classical control theory vs. neural network models of controllers. Classical control theory is obsessed with the goal of mathematically proving the dynamical stability of a controller design. For instance, it should be mathematically proven that the autopilot controller of an aircraft will not "go berserk" and make the aircraft unexpectedly dive into a deadly spin. This has led control engineers to very much prefer linear models over nonlinear ones, because the mathematics of linear controllers is very well understood, while the maths of nonlinear controllers is a hairy business. However, controllers implemented in the format of trained neural networks typically perform more accurately, and may be easier to obtain, than linear controllers. Still, in many industrial applications (robotics, aircraft) linear controllers continue to be used because of the mathematical stability assurances, which can hardly be derived for neural network based controllers. This is one of the reasons why methods of "deep learning" are only hesitantly becoming accepted in engineering.

Back to statistics. There are many ways how one can formally specify a set $\mathcal{P}$ of candidate distributions. The easiest is to establish $\mathcal{P}$ as a set of *parametrized* distributions:

Definition 19.2.1 *A collection $\mathcal{P}$ is a* parametrized family *of distributions, if there is a mathematical formula φ with real-valued parameters $\theta_1, \ldots, \theta_k$ such that every element $P_X \in \mathcal{P}$ can be represented by $\varphi(\theta_1, \ldots, \theta_k)$, where the parameter vector $(\theta_1, \ldots, \theta_k)' =: \theta$*

is chosen from a set $\Theta \subseteq \mathbb{R}^k$ of admissible parameters*. For a distribution P_X represented by $\varphi(\theta)$ one also writes P_θ.*

Four examples of parametrized families of distributions:

1. Consider the Bernoulli distributions given by the pmf's $P_X(s_1) = 1 - q,\ P_X(s_2) = q$. This is a one-parametric family with parameter $\theta = \theta_1 = q$ and admissible parameters $\Theta = [0, 1]$.

2. The family of one-dimensional normal distributions is a 2-parametric family whose members are usually characterized by the parameter vectors $\theta = (\mu, \sigma^2)'$, where μ is the mean and σ^2 the variance of the distribution. The set of admissible θ can be further constrained depending on the modeling situation, for instance one may choose $\Theta = \{(\mu, \sigma^2) \mid \mu \geq 0, \sigma^2 \leq 5\}$.

3. Our physicist measuring the speed of light probably believes that the true speed of light is a single real number, because according to a dogma of physics, reality does not "use" randomly distributed speeds of light. Therefore, the only natural kind of distribution for the speed of light is a *point measure*. Such a point measure over the measurable space $(\mathbb{R}, \mathcal{B})$ assigns a probability of 1 to every interval containing a specific point $c \in \mathbb{R}$, and assigns a zero probability to every interval that does not contain c (note: by this requirement, one uniquely specifies a probability measure on $(\mathbb{R}, \mathcal{B})$ according to Definition 7.1.1). The point measure in point c is standardly denoted by δ_c. Intuitively, this measure states that the specific value c must be observed with probability 1. Since the physicist does not know the exact value c of the speed of light, but safely may assume that it lies between 290,000,000 and 310,000,000, he might reasonably use $\Theta = [290,000,000, 310,000,000]$, where a parameter $\theta \in \Theta$ stands for the point measure δ_θ.

4. One can fix a neural network structure with m real-valued inputs and one output such that the output is always nonnegative. Such a neural network is parametrized by a (large) number k of so-called *synaptic weights*, which are lumped together in a *weight matrix* W. Different choices of W lead to networks whose structure is the same but whose computational properties vary. Assume that the inputs are constrained to a finite hypercube $H \subset \mathbb{R}^m$. A network $\mathcal{N}_W$ parametrized by weights W implements a function $\mathcal{N}_W : H \to \mathbb{R}^{\geq 0}$. From this function one can mathematicall define a pdf p_W over H by setting

$$p_W(x) = \frac{\mathcal{N}_W(x)}{\int_H \mathcal{N}_W(x)\, dx}. \tag{19.1}$$

   In this way one obtains a k-parametric family of distributions on H with parameter vectors $\theta = (w_1, \ldots, w_k)'$. Five notes are in place here. First, such neural network models of distributions are the bread and butter of today's machine learning. Second, the parameters W are here "trained" from training data. Third, k is typically large and can easily reach many thousands or even hundreds of thousands. Fourth, the denominator in (19.1) – called the *partition function* by theorists – cannot be computed analytically, and numerical estimates are very expensive to calculate. Therefore, in ML one does not make use of (19.1) directly, but one trains and uses the un-normalized variant $e_W(x) = \mathcal{N}_W(x)$. While this stripped-down version does not allow one to make proper probability statements, it can still be used to *compare* two probabilities by calculating *probability ratios*

$$\frac{p_W(x_1)}{p_W(x_2)} = \frac{\mathcal{N}_W(x_1)}{\mathcal{N}_W(x_2)},$$

which is good enough for most tasks, in particular, classification tasks. Fifth, because of these intractability issues, neural network based models of distributions are only reluctantly becoming used in statistics (as far as I know; see Paliwal and Kumar [2009] for a very enlightening comparison and survey of the performance of neural network versus "classical" statistical methods).

In these lecture notes I will only consider parametrized families of distributions. In such cases one can identify $\mathcal{P}$ with the set Θ of admissible parameter vectors θ.

19.2.4 The loss function

The *loss* is a penalty measure (higher = worse) which is inflicted on the decision maker as a consequence of his/her decision. Since reality acts its game through P_X^{true}, the true loss will depend both on P_X^{true} and the decision d. But the true distribution is unknown. In order to get an overview of what *may* happen, the statistician has to take all candidate distributions $P_X^i \in$ into consideration (or equivalently, all parametrizations $\theta \in \Theta$ in the case of a parametrized family of candidate distributions), and s/he also has to account for what may happen after any of the possible decisions $d \in \mathcal{D}$. This makes it necessary to define a *loss function* W from arguments in $\mathcal{P} \times \mathcal{D}$:

$$W : \mathcal{P} \times \mathcal{D} \to \mathbb{R}^{\geq 0} \quad (\text{or } W : \Theta \times \mathcal{D} \to \mathbb{R}^{\geq 0})$$

This loss function needs to be established by the statistician before a statistical analysis of a decision-making scenario can be started. I illustrate this concept with some simple demo examples:

- Our coin-throwing gambler, having opted for $\mathcal{D} = \{\text{fair, unfair}\}$ and $\Theta = [0, 1]$ (where $q \in \Theta$ is the probability for the coin to come up with heads), might give herself a loss function

  $$\begin{aligned} W(q, \text{fair}) &= (4\,(q - 1/2))^{10} \\ W(q, \text{unfair}) &= \begin{cases} 1 & \text{if } |q - 1/2| \leq 0.01, \\ 0 & \text{if } |q - 1/2| > 0.01. \end{cases} \end{aligned}$$

  This loss function strongly penalizes unfair coins in the case of a (mistaken) "fair" decision; and in the case of an "unfair" decision, it softly penalizes the cases where the coin is fair up to a tolerance of 0.01.

- The speed-of-light measuring physicist with decision space $\mathcal{D} = \mathbb{R}$ and parameter space $\Theta = [290,000,000, \;\; 310,000,000]$ for point distributions δ_θ may employ the *quadratic loss*

  $$W(\theta, d) = (\theta - d)^2.$$

- A federal safety engineer has to decide whether a design for a planned nuclear powerplant can be approved. Depending on numerous design elements, the engineer's model estimates the average time λ (in years) to a catastrophic reactor containment rupture and core meltdown, using for $\mathcal{P}$ the set of exponential distributions with parameter λ. By law, a nuclear reactor must be designed such that catastrophic failure events are expected to occur at a rate of less than once in 10 Mio years (I made this up, don't know about the laws and rules in reality). The decision space is $\mathcal{D} = \{\text{approve, reject}\}$. Here is a halfway reasonably looking loss function:

$$\begin{aligned} W(\lambda, \text{accept}) &= \begin{cases} \infty & \text{if } \lambda \leq 10\text{Mio}, \\ \frac{1}{\lambda - 10\text{Mio}} & \text{if } \lambda > 10\text{Mio}. \end{cases} \\ W(\lambda, \text{reject}) &= \begin{cases} 0 & \text{if } \lambda \leq 10\text{Mio}, \\ 10 & \text{if } \lambda > 10\text{Mio}. \end{cases} \end{aligned}$$

This loss should balance economic considerations against nuclear accident consequences. You may question whether the loss that I suggested is well thought out. This is a good example to demonstrate that designing loss functions is a hairy business.

19.2.5 The statistical problem

The specification of

- the format of available data, given by the sample space S,
- the choice $\mathcal{D}$ of possible decisions,
- a commitment to a set $\mathcal{P}$ of candidate distributions describing "reality",
- and a quantification of the possible consequences of decisions in the form of a loss function $W : \mathcal{P} \times \mathcal{D} \to \mathbb{R}^{\geq 0}$,

the stage is prepared for carrying out a statistical analysis aiming at finding a good decision. The four elements $S, \mathcal{D}, \mathcal{P}, W$ taken together constitute what is known as a *statistical problem.*

19.2.6 Some basic types of statistical problems

In principle, *any* kind of decisions are amenable to a statistical analysis. However, very often one faces a kind of decision that has already a long research history, an established name, and a compendium of specialized analysis techniques and algorithms. Here I give a brief listing of such standard decision scenarios. This is essentially a brief summary of Chapter 3 in the Kiefer book.

Point estimation. In point estimation problems, the objective is to estimate the values of some characteristics of the true distribution P_X^{true}. For instance, when the distribution space $\mathcal{P}$ is the set of one-dimensional normal distributions $\mathcal{N}(\mu, \sigma^2)$, one may be interested in obtaining an estimate $\hat{\mu}$ of the true expectation μ. In such cases, when one uses a parametric distribution space $\mathcal{P} \equiv \Theta$, the decision space is a subset $\mathcal{D} = \Theta_{\mathcal{D}} \subseteq \Theta$. But other characteristics besides the original parameters may also serve as the decision target. For example, when $\mathcal{P}$ is the set of exponential distributions with $\Theta = \{\lambda\}$, one may be interested to estimate the variance σ^2 of the true distribution, a parameter that is not typically used to characterize an exponential distribution.

In more generality, when $\varphi : \Theta \to \mathbb{R}^n$ is some real-valued function of candidate distributions — that is, a quantitative property of the distributions — the decision space consists of all points (vectors) in $\mathcal{D} = \varphi(\Theta)$.

For a nontrivial real-life example, consider the following statistical problem that might arise in an bank company setting. Data are obtained by a sample-generating RV X that returns filled-in online customer questionnaires. The sample used by the company's chief statistician is $X(\omega) = (X_1(\omega), \ldots, X_N(\omega)) = (x_1, \ldots, x_N)$. What kind of thing is P_X? P_X is a model of the distribution of X. Since X is a product of i.i.d. RVs X_i, a distribution P_X of X is fully specified by the individual distribution P_{X_i} of any of the X_i. So, what kind of object is P_{X_i}? Each X_i yields the answer vector of a questionnaire. If the questionnaire has m items, which take values in item-specific sample spaces $S_1, \ldots, S_m$, an answer vector x_i is an element of $S_1 \times \ldots \times S_m$. The distribution P_{X_i} of X_i is thus a distribution over the product space $S_1 \times \ldots \times S_m$ (we leave out the specification of the corresponding σ-fields for simplicity). Thus, the company statistician must design a computationally tractable representation that can describe such distributions P_{X_i}. She might, for instance, employ neural network formalisms or a formalism called a *Bayesian network* to achieve this goal. We don't bother here to make this more specific but assume that the statistician has found a way to represent P_{X_i} by a parametric model, where each candidate distribution P_{X_i} is specified by a parameter vector θ (neural networks and Bayesian networks are parametric models). The set $\mathcal{P}$ of candidate distributions thus can be identified with a parameter space Θ. — However, the company is very likely not primarily interested in modeling the joint distribution of questionnaire item answers. Instead, they will be interested in specific questions like, *"what is the probability Q that a customer whose age is between 40 and 45 and who is married and who has a regular income between 2300 and 2500 Euros will pay back a loan of 10,000 Euros within 2 years?"*. This interesting probability Q will not be directly expressed in the distribution P_{X_i}. The company statistician, if she is worth her high salary, will however be able to mathematically infer Q from P_{X_i}, that is, from θ. This mathematical inference procedure is a function $\varphi : \Theta \to [0, 1]$. The decision space becomes $\varphi(\Theta) = [0, 1]$, and the whole affair is revealed as a point estimation problem.

It is interesting to notice that machine learning problems can often be regarded as point estimation problems. For example, when one trains a neural network, one first fixes the "architecture" of the neural network, that is the number and connectivity topology of the neurons. If the outputs of the neural network are interpreted as a probability vector ("hypothesis vector" in networks trained for classification tasks, the most common usage of neural networks), such a neural network architecture can be seen as a parametric model of the conditional distribution of outputs given the inputs to the network. The parameter space Θ is the space of all possible *synaptic weights* inside the network, which are usually collected in a *weight matrix* $\Theta = \mathcal{W}$. Neural network training algorithms are designed to estimate from a sample (called *training data* in ML) a weight matrix $\hat{\mathcal{W}}$ which should in some way come close to $\mathcal{W}^{\text{true}}$, the neural network describing the true conditional output-given-input distribution. Training a neural network thus would be considered, from a statistics perspective, as a statistical point estimation problem where $\mathcal{D} = \mathcal{P} = \Theta$.

Interval and region estimation. It may be wise for a decision maker to make cautious decisions. A point estimate decision is a bold statement. It amounts to saying, "I declare the correct value $\varphi(\theta^{\text{true}})$ to be exactly this number y". Very likely such a statement is wrong. In many situations it is more advisable to acknowledge one's limits in assessing reality and instead decide to make a statement like, "I declare that the true speed of light to lie in the interval 299792458 ± 2", or "The meteor will hit earth in the ocean, at a distance at least 100 miles from the closest landmass".

The decision space is then a set of intervals or non-rectangular regions in $\varphi(\Theta)$.

A special case of interval/region estimation is obtained when the decision statements are augmented by a probability that the statement is correct, as in "I declare that the true speed of light to lie in the interval 299792458 ± 2, with a confidence of 95%". When all decision regions $D \in \mathcal{D}$ are qualified with the same confidence level, one speaks of *confidence interval / region estimation.* This is in widespread use in reporting experimental findings throughout the sciences. In fact this can be regarded the de facto minimal standard for good scientific practice. You can recognize it when result graphs show plot points that are braced by *error bars*, which come in various graphical appearances (Figure 19.1).

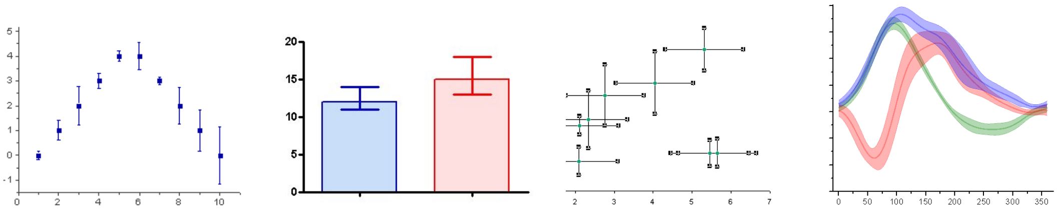


Figure 19.1: Some random webpicks showing result plots with error bars (from www.originlab.com, s3.amazonaws.com/cdn.graphpad.com, peltiertech.com, www.adeptscience.co.uk)

Hypothesis testing. This is the most classical kind of statistical decisions. The treatment of hypothesis testing fills half or more of standard textbooks for statistics in applied fields. It concerns situations where a scientific yes/no hypothesis is confronted with empirical data.

The archetypical example is to demonstrate that a newly found pharmaceutical substance C has a curing effect, using data from a clinical survey where some patients have been treated with the new substance C and others with a placebo. The hypothesis for which evidential support is sought is "substance C has a positive effect". This target hypothesis is often called (somewhat counter-intuitively) the *alternative hypothesis*, abbreviated H1. It is contrasted with the opposite hypothesis, "substance C has no positive effect", called the *null hypothesis*, abbreviated H0. A *level of significance* (typically $\alpha = 1\%$ or $\alpha = 5\%$) is fixed, and the decision set contains two judgments $\mathcal{D} = \{C$ has a positive effect at the level α of significance, C has no positive effect at the level α of significance$\}$.

Students in the natural sciences, and students in psychology and medicine in particular, will (or should) take a "statistical methods" course in their first year of studies. The main course contents is to give them an intense training in formulating scientific hypotheses in H1 vs. H0 pairs, and selecting and carrying out statistical procedures to decide between them.

Since this type of decision is so fundamental and so extensively covered in "textbooks and cookbooks on statistics" (quote from the Kiefer book), we defer a more in-depth treatment to a separate section.

The Kiefer book describes a few more standard types of statistical problems, like for instance regression or ranking problems, which we skip here.

19.3 Statistical procedures

A *statistical procedure* is an algorithm to determine a decision $d \in \mathcal{D}$ on the basis of a data sample $x = (x_1, \ldots, x_N)$. It is a data-based decision-making algorithm. From a mathematical perspective, such an algorithm implements a function $t : S \to \mathcal{D}$. We will use the symbol t to denote statistical procedures. Formally, any such function qualifies as a statistical procedure:

Definition 19.3.1 *A (any!) function $t : S \to \mathcal{D}$ is called a* statistical procedure*.*

Notice that $t \circ X : \Omega \to \mathcal{D}$ is a random variable. For mathematical analyses, one has to equip the decision space $\mathcal{D}$ with a σ-field $\mathcal{F}_\mathcal{D}$. Often the choice of $\mathcal{F}_\mathcal{D}$ is straightforward: when $\mathcal{D}$ is a finite set, one uses the power set σ-field; when $\mathcal{D}$ is a continuous space (some subset of $\mathbb{R}^n$) one uses the Borel σ-field. We will not further deal with this subject and take decision spaces $\mathcal{D}$ just as plain sets, without thinking about a σ-field structuring of it.

All the art and science of statistics revolves around finding "good" statistical procedures. This necessitates, in the first place, to formalize what "good" means. We will quickly find that this is not an easy question.

19.3.1 Comparing statistical procedures

I begin with a basic example, taken from the Kiefer book and extended by Adi Wilhelm. We consider the following elementary statistical problem:

- The data come from tossing a coin 10 times. This can be formalized by ten identically distributed, independent RVs $X_1, \ldots, X_{10}$, each X_i with sample space $S_i = \{0, 1\}$. Combining these ten items into a single RV $X = X_1 \bigotimes \ldots \bigotimes X_{10}$ with sample space $S = \{0, 1\}^{10}$ yields our problem's sample space S. The samples $X(\omega)$ are 10-tuples of zeros and ones, written for convenience as$X(\omega) = (x_1, \ldots, x_{10})$.

- The targetted decision is to claim an estimate $\hat{\mu}$ of the expectation $E[X_1]$ (which is the same for all X_i because the X_i are identically distributed). The coin is perfectly fair if $E[X_1] =: \mu = 1/2$. The decision space is $\mathcal{D} = [0,\, 1]$.

- The distribution of each X_i is an Bernoulli distribution. A Bernoulli distribution is specified (recall Section 16.1.1) by a single parameter θ (called q in Section 16.1.1) which is defined by $\theta = P_{X_i}(1)$. The natural choice for the set of candidate distributions $\mathcal{P}$ for X is thus given by the family of all pmf's p_θ defined by

$$p_\theta(x_1, \ldots, x_{10}) = \theta^{\sum_i x_i} \, (1 - \theta)^{10 - \sum_i x_i},$$

  where $\theta \in [0,\, 1]$.

- For a loss function, we opt for the quadratic loss,

$$W(\theta, \hat{\mu}) = (\theta - \hat{\mu})^2.$$

Let us now inspect and compare a choice of four statistical procedures $t_j : S \to [0,\, 1]$. Recall that, roughly speaking (we will refine that idea very soon), the purpose of a statistical procedure is to give a "good" decision, which in this simple statistical problem means to produce from a sample $(x_1, \ldots, x_{10})$ a "good" estimate $\hat{\mu}$ of θ. Here is the list:

$$\begin{aligned}
t_1(x_1, \ldots, x_{10}) &= \frac{1}{10} \sum_{i=1}^{10} x_i \\
t_2(x_1, \ldots, x_{10}) &= \begin{cases} \frac{\pi}{5} & \text{if } \sum_{i=1}^{10} x_i \text{ is odd} \\ 0 & \text{if } \sum_{i=1}^{10} x_i \text{ is even} \end{cases} \\
t_3(x_1, \ldots, x_{10}) &= \frac{1}{2} \\
t_4(x_1, \ldots, x_{10}) &= \frac{1}{6} \sum_{i=1}^{6} x_i
\end{aligned}$$

As you can see, not all procedures are appear to be equally "good". The first one is the sample mean which appears quite reasonable. The second procedure appears to be plain crazy. The third procedure completely ignores the experimental results and firmly beliefs that the coin is fair. The fourth one only considers the first six tosses of the coin, ignoring the results of the last four ones.

The clear winner seems to be t_1. Or, ... is it not?

In order to make it precise what is a "good" statistical procedure t, we recall that we quantified that the "goodness" of a decision by the loss function W. The loss function assigns a penalty value $W(P_X, d)$ to every assumed data distribution $P_X \in \mathcal{P}$ and decision candidate $d \in \mathcal{D}$. A decision is actually made on the basis of a sample $X(\omega) = (x_1, \ldots, x_{10})$. But, the sample data $(x_1, \ldots, x_{10})$ are the outcome of a random observation. If we would repeat the decision-making again on the basis of another sample $X'(\omega) = (x'_1, \ldots, x'_{10})$, using the same procedure t, we may be led to *another* decision d'. The crucial quality measure is thus not the loss $W(P_X, d)$ which we compute for a specific decision d, but the *expected* loss that we would earn on average if we could repeat the decision-making on the basis of freshly drawn samples. This leads to the notion of the *risk* of a statistical procedure t:

Definition 19.3.2 *Let $W : \Theta \times \mathcal{D} \to \mathbb{R}^+$ be a loss function, X the sample-generating RV, and $t : S \to \mathcal{D}$ a statistical procedure. Then the function*

$$R_t : \mathcal{P} \to \mathbb{R}^+, \quad R_t(P_X) = E_{P_X}[W(P_X, t(X))],$$

providing an a-priori measure of the performance of the statistical procedure t, is called the risk function of the procedure *t.*

Here E_{P_X} denotes the expectation taken over the sample distribution P_X. When the family $\mathcal{P}$ of distributions P_X is a parametric family, we can identify a distribution P_X with a parameter vector θ and $\mathcal{P}$ with a set Θ of such parameter vectors, and we can also write

$$R_t : \Theta \to \mathbb{R}^+, \quad R_t(\theta) = E_\theta[W(\theta, t(X))].$$

A note on terminology: different uses of the words "loss" and "risk". In the machine learning literature, even in the mathematical-theoretical branches of that literature, the words *loss* and *risk* are used for concepts that are somewhat simplified and restricted versions of how we just defined them, following the Kiefer book. Namely, in that literature, these notions arise solely in the supervised learning contexts, where the typical objective is to learn a regression function D which upon pattern inputs $X(\omega) = x$ returns a regression output $D(x)$ which can be compared to a "correct" output $Y(\omega) = y$. The quality of that output is quantified by a "loss" function which compares the result of applying D on x with the correct output y given by reality. Formally, if outputs $D(x)$ and true results y reside in a set S_Y, a loss function in that literature is a function $L_D : S_Y \times S_Y \to \mathbb{R}^+$ and is designed in a way that a small loss $L(D(x), y)$ means that $D(x)$ is in some way good and a high loss is bad. A much-used loss is the *quadratic loss* which can be defined when $S_Y = \mathbb{R}^k$ and is given by $L(D(x), y) = \|(D(x) - y\|^2$. In the machine learning literature, the "risk" $R(D)$ is defined to be a characteristic of D, namely, it is the expected loss $D(E) = E[L(D(X), Y)]$, where the expectation is taken over *the* joint distribution of X and Y. I wrote "*the* expectation" because in machine learning one starts from "the" distribution of real-world data, period — different from how statisticians think who start their theories from the notion of a set $\mathcal{P}$ of *possible* candidate distributions that the real world data *might* embody. In summary, in machine learning the risk is construed as a characteristic of a fixed regression function (or classification function as a special case) D, whereas in the statistical way of thinking, the risk is a characteristic of a statistical procedure. Furthermore, while in machine learning the risk of a regression function D is a single *number* from $\mathbb{R}^+$, in Kiefer's world the risk of a statistical procedure is a *function* from a space of candidate distributions to $\mathbb{R}^+$. Confusing!

For the coin tossing example with $\Theta = [0, 1]$, the quadratic loss and the above defined four statistical procedures we obtain the following risk functions:

$$\begin{aligned}
R_{t_1}(\theta) &= E_\theta[(t_1(X) - \theta)^2] \\
&= E_\theta\left[\left(\frac{1}{10}\sum_{i=1}^{10} X_i - \theta\right)^2\right] \\
&= \left(\frac{1}{10}\right)^2 E_\theta\left[\left(\left(\sum_{i=1}^{10} X_i\right) - 10\,\theta\right)^2\right] \\
&= \left(\frac{1}{10}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{10} X_i\right) \\
&= \left(\frac{1}{10}\right)^2 10 \cdot \theta(1-\theta) \\
&= \frac{1}{10}\theta(1-\theta)
\end{aligned}$$

$$\begin{aligned}
R_{t_2}(\theta) &= E_\theta[(t_2(X) - \theta)^2] \\
&= \frac{1}{2}(\frac{\pi}{5} - \theta)^2 + \frac{1}{2}(0 - \theta)^2 \\
&= \frac{\pi^2}{50} - \frac{\pi}{5}\theta + \theta^2
\end{aligned}$$

$$\begin{aligned}
R_{t_3}(\theta) &= E_\theta[(t_3(X) - \theta)^2] \\
&= E_\theta\left[(\frac{1}{2} - \theta)^2\right] \\
&= (\frac{1}{2} - \theta)^2
\end{aligned}$$

$$\begin{aligned}
R_{t_4}(\theta) &= E_\theta[(t_4(X) - \theta)^2] \\
&= E_\theta\left[\left(\frac{1}{6}\sum_{i=1}^{6} X_i - \theta\right)^2\right] \\
&= \frac{1}{6}\theta(1-\theta)
\end{aligned}$$

The risk functions of procedures 1, 2, 3 and 4 are visualised in Fig. 19.2.

If you think about what you see in Figure 19.2 you will find that procedure t_1 is not necessarily the "best". If one would have some previous knowledge that the coin is at most only *a little* unfair — say, one believes that the true value of θ deviates from 1/2 by no more than 0.1 in either direction, so one would use $\Theta = [0.4, 0.6]$ — then the procedure t_3 would always give a lower expected loss than the natural-looking procedure t_1, because the risk curve of t_3 is always below the curve for t_1 in the intervale [0.4, 0.6].

If a statistician searches for a low-risk procedure, he/she must carry out this search in a set of candidate procedures (in mathematics, when one searches for some "optimal" item, one must *always* first define the "search space" in which one carries out the search — a search without a specified candidate space is ill-defined). We will use the symbol $\mathcal{T}$ for the collection of candidate procedures t_i that a statistician considers and compares.

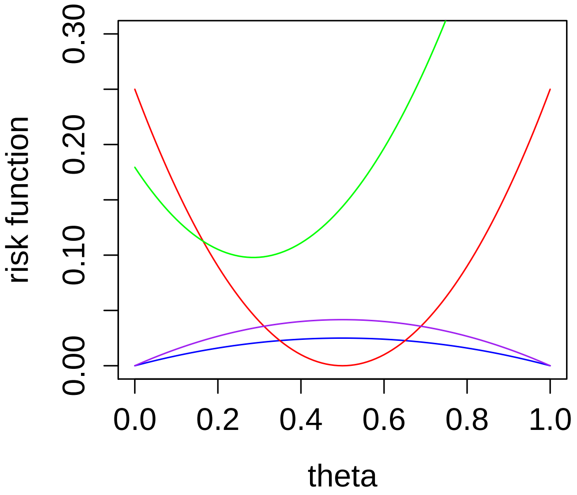


Figure 19.2: Some risk functions for the coin tossing example. The risk functions of procedure t_1 is given in blue, the one of t_2 in green, t_3 in red, and t_4 in purple.

In principle, one would love to find a statistical procedure that is best in the sense that its risk is always smaller than the risk for any competing procedure, across all candidate distributions $\theta \in \Theta$, and across all competing procedures $t \in \mathcal{T}$. If such a globally optimal procedure exists, it is called the *uniformly best* procedure in $\mathcal{T}$.

The sad truth is that typically such uniformly best procedures do not exist (unless Θ and $\mathcal{T}$ are very peculiar). However, one can always exclude from considerations all statistical procedures in $\mathcal{T}$ which are everywhere outperformed by a competing procedure. If t is a decision procedure such that $R_t(\theta) \leq R_{t'}(\theta) \quad \forall \theta \in \Theta$, and for some θ_0 we have $R_t(\theta_0) < R_{t'}(\theta_0)$, we say that t *dominates* the procedure t'. Procedures t' which are dominated by some other t can be excluded from further consideration. Such certainly non-optimal procedures are called *inadmissible*. All procedures which are not dominated by some other one are called *admissible*. Weeding out from $\mathcal{T}$ all inadmissible procedures leaves the statistician with a search space wherein each candidate procedure is better (lower risk) than another one for some but not all candidate distributions θ.

In our coin-tossing example (see Figure 19.2) we find that procedures t_2 and t_4 are dominated by t_1. Thus t_2 and t_4 are inadmissible and can be ignored. Neither t_1 nor t_3 dominate the respective other. After purging our original collection $\mathcal{T}_0 = \{t_1, t_2, t_3, t_4\}$ from t_2 and t_4 we are left with $\mathcal{T} = \{t_1, t_3\}$. We need additional tools and ideas to decide which of these two admissible procedures we should use.

19.4 Criteria for choosing a statistical procedure

The question of how to pick a "best" procedure from a collection $\mathcal{T}$ of admissible procedures has no simple universal answer. Large textbook sections (and important subtraditions within statistics) are devoted to specific approaches to characterize "best" procedures in an admissible candidate set. Depending on what approach (or tradition) one subscribes to, one will ultimately opt for different candidate procedures as the "best" one. We proceed to take a look at some of the major strategies.

19.4.1 Minimax criterion

Definition 19.4.1 *A statistical procedure $t^* : S \to \mathcal{D}$ is said to be* ***minimax****, if*

$$t^* = \operatorname*{argmin}_{t \in \mathcal{T}} \max\{R_t(\theta) \mid \theta \in \Theta\}.$$

The minimax strategy selects the one statistical procedure that minimizes the worst-case error. In very simple cases, we might be able to directly compute the minimax procedure by comparing the risk functions for all statistical procedures under consideration. This typically only works for small parameter spaces and limited classes of statistical procedures. For any reasonable statistical problem, the minimax procedure needs to be computed indirectly. There are various approaches for computing minimax procedures. One of them is closely related to the next approach.

19.4.2 Bayes criterion

The Bayesian approach assumes that the parameter space Θ can be turned into a probability space. We specify some distribution $\tilde{P}$ on the parameter space Θ (using some appropriate σ-field on it). Since our parameter space is mostly taken to be a subset of $\mathbb{R}^n$, the default choice is to use the corresponding Borel σ-field over this subset. In the classical Bayesian context, the probability distribution on the parameter space selected is supposed to quantify the knowledge we have about the likelihood of the different parameters *prior* to observing data.

As a measure of risk the Bayes criterion now asks to choose a statistical procedure t that minimizes the *expected* error according to our prior beliefs about θ. This quantity is called *Bayes risk* of t and is defined by

$$r(t) = E_{\tilde{P}}[R_t(\theta)] = E_{\tilde{P}}[E_\theta[W(\theta, t(X))]] = \int_\Theta \left(\int_S W(\theta, t(x))\, p_\theta(x)\, dx \right) \tilde{p}(\theta)\, d\theta.$$

Here, $\tilde{p}$ is the probability density function for the prior distribution $\tilde{P}$ over Θ and $p_\theta(x)$ is the pdf on S for the distribution $P_X = \theta$ (we assume for simplicity that both distributions have pdf's).

19.4.3 Unbiasedness and statistical efficiency

The criterion of unbiasedness is widely used when the statistical problem is of the point estimation kind. Let us assume that we want to estimate some characteristic $\varphi(P_X) = \varphi(\theta)$ of the (parametrized) sample distribution P_X. The decision space then is equal to $\mathcal{D} = \varphi(\Theta)$.

Definition 19.4.2 *A statistical procedure $t : S \to \mathcal{D}$ is called an* unbiased estimator *of $\varphi(\theta)$ if*

$$E_{P_X}[t(X)] = \varphi(P_X), \qquad \forall P_X \in \mathcal{P},$$

or more specifically, when we are dealing with parametric models θ of P_X, if

$$E_\theta[t(X)] = \varphi(\theta), \qquad \forall \theta \in \Theta.$$

An unbiased estimator for $\varphi(\theta)$ thus will yield estimates $\hat{\varphi}(\theta)$ that *on average across different samples $X(\omega)$* returns the correct $\varphi(\theta)$, for any of the candidate distributions $\theta \in \Theta$. Since unbiasedness is so often desired, we take a closer look at this concept and investigate what it means for an estimator to be *not* unbiased.

Definition 19.4.3 *For any estimator t whose expectation exists for all P_X (or for all θ), the function*

$$b_t : \mathcal{P} \to \mathcal{D}, \qquad b_t(P_X) = E_{P_X}[t(X)] - \varphi(P_X)$$

or

$$b_t : \Theta \to \varphi(\Theta), \qquad b_t(\theta) = E_\theta[t(X)] - \varphi(\theta)$$

is called the bias function *of t for estimating $\varphi(\theta)$.*

Thus, a statistical procedure t is unbiased if and only if the bias function $b_t(P_X)$ equals zero for all candidate distributions $P_X \in \mathcal{P}$.

Assume furthermore that our loss function is the quadratic loss. Then our risk function turns into

$$R_t(P_X) = E_{P_X}[\|t(X) - \varphi(P_X)\|^2]$$

and is called the *mean squared error*, denoted by MSE$[t(X)]$.

When we are considering a scalar $\varphi(\theta) \in \mathbb{R}$, a straightforward calculation shows that the MSE results from the added effects of the variance of the statistical procedure and the bias of our statistical procedure:

Proposition 19.4.1

$$\mathrm{MSE}[t(X)] = b_t(P_X)^2 + \mathrm{Var}[t(X)].$$

Some comments that help to appreciate the practical importance of this proposition:

- Point estimation problems where a squared error loss is employed are encountered very often. For instance, many machine learning problems are of this kind.

- Proposition 19.4.1 is intimately connected to a fundamental difficulty encountered in machine learning, the *overfitting* problem. This difficulty arises in scenarios where φ is the identity function, that is, one wants to directly estimate the parameters θ of the distribution P_X — in simple words, one wants to estimate a good model of the data distribution from a sample. In such scenarios an estimator (that is, a machine learning algorithm) is prone to have high variance when its model estimates $\hat{\theta}$ are strongly influenced by variations across different training samples. This happens when the estimator is capable of fitting the particular, random detail of a training sample $X(\omega)$. While this leads to small *training error*, the obtained model $\hat{\theta}$ will poorly *generalize*. Conversely, when the estimator is very inflexible and its output $\hat{\theta}$ is only weakly shaped by information in the training sample $X(\omega)$, its variance will be small but its bias will usually be large — *underfitting* happens. The close connection between underfitting and bias on the one hand, overfitting and variance on the other have led to the terminology to refer to the overfitting difficulty as the *bias-variance dilemma*.

- It is often quite feasible to design a candidate set $\mathcal{T}$ of estimators that are all unbiased. The MSE loss then is determined solely by the variance Var$[t(X)]$. This leads to a clear selection criterion: choose an estimator $t \in \mathcal{T}$ whose variance is minimal (if possible, for all $P_X \in \mathcal{P}$). The variance of an estimator is also referred to as the *statistical efficiency* of an estimator (usage of this terminology: lower variance = higher statistical efficiency). Investing effort to trim down the variance of an estimator is often key for designing useful machine learning algorithms.

19.4.4 Maximum likelihood

The last approach to determining a "good" statistical procedure which we will inspect is the method of *maximum likelihood* model estimation. The acronym ML is commonly used for "maximum likelihood". The set-up goes like this:

- The candidate distributions $\mathcal{P}$ are a paremetrized family, so we can identify candidate distributions with their parameter vectors θ.
- The task is to obtain an estimate $\hat{\theta}$ of the true distribution θ^{true}. That is, we are facing a point estimation problem with $\varphi = \mathbf{id}$, the identity function.

For ease of discussion, we only consider the case where the distributions $P_X \in \mathcal{P}$ are discrete and are described by a pmf p_θ, which is a function $p_\theta : S \to [0, 1]$; this function is parametrized by θ.

The ML approach starts from the concept of the *likelihood* of a model θ:

Definition 19.4.4 *Let $\theta \in \Theta$ be a model of a distribution, and $x = X(\omega)$ a sample. Then the* likelihood *of θ is the value $p_\theta(x)$ of the model pmf at the sample point x.*

Notice that for discrete distributions (only which we consider), $p_\theta(x) = P_\theta(X = x)$. The likelihood of the model θ at x is thus the same as the probability of the sample data x under the model θ. While "Likelihood of model θ" and "probability of data under distribution θ" refer to the same quantity, namely $P_\theta(X = x)$, well-educated statisticians cleanly distinguish between the two concepts: likelihood is a property of a model; probability is a property of a sample.

The maximum-likelihood estimator $t^{\text{ML}} \in \mathcal{T}$ is the estimator which is distinguished among all candidate estimators by the condition

$$t^{\text{ML}} = \underset{t \in \mathcal{T}}{\operatorname{argmax}}\; p_{t(x)}(x). \tag{19.2}$$

In plain English, the ML estimator t^{ML} is the one whose outcome, the distribution model $t^{\text{ML}}(x) = \hat{\theta}$, assigns the highest probability to the observed data x. The "ML principle" in a nutshell: find a model that best explains the data.

ML (maximum likelihood) model estimation is the bread and butter in ML (machine learning)! Specifically, the currently popular neural network models ("deep learning") are trained by learning algorithms which attempt to maximise the probability of the training data x under the network model θ. But ML approaches for finding "good" model estimates are also popular in statistics proper. There are two reasons for their popularity:

- The ML principle appears intuitively plausible.
- Actually finding t^{ML} by solving the optimization problem (19.2) can be done by well-established computational procedures. Specifically, *gradient-descent* methods can be directly used to solve (19.2), as well as likewise well-established algorithms known as *Expectation-Maximization* (EM-) algorithms. For both types of computational procedures, a long tradition, a rich literature, and comprehensive software tool support exists.

However, directly heading for ML model estimates is a slippery road which directly leads to overfitting. Consult my machine learning lecture notes to learn more about all of this.

Appendix A

Elementary mathematical structure-forming operations

A.1 Pairs, tuples and indexed families

If two mathematical objects $\mathcal{O}_1, \mathcal{O}_2$ are given, they can be grouped together in a single new mathematical structure called the *ordered pair* (or just *pair*) of $\mathcal{O}_1, \mathcal{O}_2$. It is written as

$$(\mathcal{O}_1, \mathcal{O}_2).$$

In many cases, $\mathcal{O}_1, \mathcal{O}_2$ will be of the same kind, for instance both are integers. But the two objects need not be of the same kind. For instance, it is perfectly possible to group integer $\mathcal{O}_1 = 3$ together with a random variable (a function!) $\mathcal{O}_2 = X_7$ in a pair, getting $(3, X_7)$.

The crucial property of a pair $(\mathcal{O}_1, \mathcal{O}_2)$ which distinguishes it from the *set* $\{\mathcal{O}_1, \mathcal{O}_2\}$ is that the two members of a pair are *ordered*, that is, it makes sense to speak of the "first" and the "second" member of a pair. In contrast, it makes not sense to speak of the "first" or "second" element of the set $\{\mathcal{O}_1, \mathcal{O}_2\}$. Related to this is the fact that the two members of a pair can be the same, for instance $(2, 2)$ is a valid pair. In contrast, $\{2, 2\}$ makes no sense.

A generalization of pairs is *N-tuples*. For an integer $N > 0$, an N-tuple of N objects $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_N$ is written as

$$(\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_N).$$

1-tuples are just individual objects; 2-tuples are pairs, and for $N > 2$, N-tuples are also called *lists* (by computer scientists that is; mathematicians rather don't use that term). Again, the crucial property of N-tuples is that one can identify its i-th member by its position in the tuple, or in more technical terminology, by its *index*. That is, in an N-tuple, every index $1 \leq i \leq N$ "picks" one member from the tuple.

The infinite generalization of N-tuples is provided by *indexed families*. For any nonempty set I, called an *index set* in this context,

$$(\mathcal{O}_i)_{i \in I}$$

denotes a compound object assembled from as many mathematical objects as there are index elements $i \in I$, and within this compound object, every individual member $\mathcal{O}_i$ can be "addressed" by its index i. One simply writes

$$\mathcal{O}_i$$

to denote the ith "component" of $(\mathcal{O}_i)_{i\in I}$. Writing $\mathcal{O}_i$ is a shorthand for applying the ith projection function on $(\mathcal{O}_i)_{i\in I}$, that is, $\mathcal{O}_i = \pi_i((\mathcal{O}_i)_{i\in I})$.

A.2 Products of sets

We first treat the case of products of a finite number of sets. Let $S_1, \ldots, S_N$ be (any) sets. Then the product $S_1 \times \ldots \times S_N$ is the set of all N-tuples of elements from the corresponding sets, that is,

$$S_1 \times \ldots \times S_N = \{(s_1, \ldots, s_N) \mid s_i \in S_i\}.$$

This generalizes to infinite products as follows. Let I be any set — we call it an *index* set in this context. For every $i \in I$, let S_i be some set. Then the *product set indexed by I* is the set of functions

$$\prod_{i\in I} S_i = \{\varphi : I \to \bigcup_{i\in I} S_i \mid \forall i \in I : \varphi(i) \in S_i\}.$$

Using the notation of indexed families, this could equivalently be written as

$$\prod_{i\in I} S_i = \{(s_i)_{i\in I} \mid \forall i \in I : s_i \in S_i\}.$$

If all the sets S_i are the same, say S, then the product $\prod_{i\in I} S_i = \prod_{i\in I} S$ is also written as S^I.

An important special case of infinite products is obtained when $I = \mathbb{N}$. This situation occurs universally in modeling stochastic processes with discrete time. The elements $n \in \mathbb{N}$ are the points in time when the amplitude of some signal is measured. The amplitude is a real number, so at any time $n \in \mathbb{N}$, one records an amplitude value $a_n \in S_n = \mathbb{R}$. The product set

$$\prod_{n\in\mathbb{N}} S_n = \{\varphi : \mathbb{N} \to \bigcup_{n\in\mathbb{N}} S_n \mid \forall n \in I : \varphi(n) \in S_n\} = \{\varphi : \mathbb{N} \to \mathbb{R}\}$$

is the set of all right-infinite real-valued timeseries (with discrete time points starting at time $n = 0$).

A.3 Products of functions

First, again, the case of finite products: let $f_1, \ldots, f_N$ be functions, all sharing the same domain D, with image sets S_i. Then the product $f_1 \otimes \ldots \otimes f_N$ of these functions is the function with domain D and image set $S_1 \times \ldots \times S_N$ given by

$$\begin{aligned} f_1 \otimes \ldots \otimes f_N : D &\to S_1 \times \ldots \times S_N \\ d &\mapsto (f_1(d), \ldots, f_N(d)). \end{aligned}$$

Again this generalizes to arbitrary products. Let $(f_i : D \to S_i)_{i\in I}$ be an indexed family of functions, all of them sharing the same domain D, and where the image set of f_i is S_i. The product $\bigotimes_{i\in I} f_i$ of this set of functions is defined by

$$\begin{aligned} \bigotimes_{i\in I} f_i : D &\to \prod_{i\in I} S_i \\ d &\mapsto \varphi : I \to \bigcup_{i\in I} S_i \quad \text{given by } \varphi(i) = f_i(d). \end{aligned}$$

Appendix B

Kiefer book notation

Symbol	Meaning in Kiefer	Meaning in these LN
Ω	Set of candidate distributions	universe
W	Loss function. The standard symbol for loss functions that is used almost universally in the literature other than Kiefer's book is L.	Weight matrix (in linear regression or neural networks); I follow however Kiefer in the Statistics part and use W for the loss function there

Mathematical object	Notation in Kiefer	Notation in these LN
probability of an event A	$P_{F_0}\{A\}$, where F_0 is one of the candidate distributions; or $P_{\theta_0}\{A\}$, where θ_0 is a parametrization of a candidate distribution F_0	$P(X \in A)$ or $P_X(A)$, where X is the RV generating *the* (single!) distribution
distribution	F	P_X

Bibliography

H. Bauer. *Wahrscheinlichkeitstheorie und Grundzüge der Maßtheorie.* de Gruyter, Berlin/New York, 3 edition, 1978. English translation: Probability Theory and Elements of Measure Theory, New York: Holt, Rinehart & Winston, 1972.

G. E. Crooks. Field guide to continuous probability distributions, v 0.11 beta. online manuscript, retrieved april 2017, 2017. URL `http://threeplusone.com/fieldguide`.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (second edition).* Wiley Interscience, 2001.

A. Graves. Generating sequences with recurrent neural networks. Unpublished manuscript, Univ. of Toronto, 2014. arXiv:1308.0850.

E. T. Jaynes. *Probability Theory: the Logic of Science.* Cambridge University Press, 2003, first partial online editions in the late 1990ies. First three chapters online at `http://bayes.wustl.edu/etj/prob/book.pdf`.

A. Keel. *Statistik III: Induktive Statistik, 16-th edition (in German).* Verlag Wilhelm Surbir Wittenbach/St. Gallen, 2004a.

A. Keel. *Statistik II: Wahrscheinlichkeit, 15-th edition (in German).* Verlag Wilhelm Surbir Wittenbach/St. Gallen, 2004b.

J. C. Kiefer. *Introduction to Statistical Inference.* Springer Verlag, 1987.

A. N. Kolmogorov. *Foundations of the Theory of Probability (Second English Edition).* Chelsea Publishing Company, 1956. English translation of Kolmogorov's original German book *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Ergebnisse der Mathematik 2 Nr 3, Berlin 1933.

B. Mau, M.A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, 55:1–12, 1999.

P. H. Müller, editor. *Wahrscheinlichkeitsrechnung und Mathematische Statistik, 3rd edition.* Akademie Verlag Berlin, 1983.

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. http://arxiv.org/abs/1412.1897, 2014.

M. Paliwal and U. A. Kumar. Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36:2–17, 2009.

J. A. Sanchez, A. H. Toselli, V. Romero, and E. Vidal. ICDAR 2015 competition HTRtS: Handwritten text recognition on the transcriptorium dataset. In *Proc. ICDAR 2015*, 2015.