Probability and Statistics

Jossy Sayir
js851@cam.ac.uk

Contents

Chapter 1

Probability Fundamentals

1.1 Tribute and acknowledgments

I wrote the first version of these lecture notes in 2016 while teaching the course as an emergency stand-in for Professor Sir David J.C. MacKay, FRS, who had become ill and unable to teach the course. Working with David on these lecture notes remains one of the most memorable experiences of my professional life: I spent hours with David, sometimes in his house when he was not well enough to come to work, drinking smoked tea and discussing the details of these notes. David did not think it necessary for me to write these notes and hence I would not feel comfortable listing him as a co-author, despite the fact that there is so much in these notes that was inspired by him, changed by him and corrected by him. David told me "Probability theory is very easy: there is a sum rule and a product rule and that's all you need to know. But it takes the average student about two years until they learn to apply these rules." I am afraid I myself must have been a below average student because it took me much longer than two years to understand this, much though I agree with him. You can get a flavour of David's simpler approach in 1.5.1 at the end of Section 1.5 and in his wonderfully engaging book [Mac03], which is recommended reading for this course and for many of the information engineering courses in Part II (3F7, 3F8, 3F4, 3F1, 3F3, 4F5, 4F7, 4F10 and 4F13). David passed away in April 2016 at the age of 49, one month after the last chapter of these lecture notes was completed.

I also want to thank many students of the 2016 class, as well as some in later years and some colleagues and supervisors, for their help and corrections, in particular Martina Cheadle, Ben Clayton, Stephen Dimmock, Andrew Foong, Jia Min Gan, James Gard, Dr Adam Greig, Clinton Igwegbu, Jamil Jami, Christos Kakoutas, Florian Kreyssig, Nati Neggatu, Sebastian Ober, Benedict Poh, Andreas Theodosiou, Jon Arne Toft, Dr Ramji Venkataramanan, Dr Rich Wareham, Syed Akhasss Adnan Wasti, Nicholas Wong, and Ji Xun Yeoh.

1.2 What will be covered in this course?

You've possibly had some probability and statistics at secondary school[1]. In Part IA you had a "Teach yourself probability" examples paper. What more is there to know about probability and statistics? We will start by discussing:

- where you stand (what you may already know about probability and statistics) ;

- what I plan to teach you about probability and statistics;

- what you still won't know about probability and statistics after this course.

---

[1]much of what is taught in school as probability and statistics is in fact combinatorics, a mathematical discipline concerned with counting things. There will be very little combinatorics in this course.

Rather than provide my own elaborate map of the fields of probability and statistics, it's easiest to borrow the contents pages of two books. The first of these books is the Riley, Hobson and Bence [RHB06], a very heavy book covering all the mathematics that would ever be of use to engineers (and some more.) Chapters 30 and 31 in this book cover probability and statistics, respectively. Their contents are shown in `http://assets.cambridge.org/97805216/79718/toc/9780521679718_toc.pdf`. The level you achieved in the IA "Teach yourself probabilty" examples paper covers roughly 30.1-3 and 31.1-2. (Venn Diagrams, basic probability, permutations and combinations, averages, variances and standard deviation.) In this chapter, I plan to go back over what you learned in some more depth, then take you onwards to cover a cross-section of chapters 30 and 31 in the [RHB06]. By the end of the module, you should have a fair understanding of most sections in these two chapters, or understand enough to be able to read up on anything in these chapters not covered in the lectures. The [RHB06] is recommended reading for those of you who want a different perspective or want to go beyond the material in the lectures.

The next book I want you to take a look at is the Billingsley [Bil95] "Probability and Measure", which is a standard textbook in probability for mathematicians. Its content is shown in `http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118122372.html`. There is no need to dwell on the detail of all 38 chapters in this book at this stage. The main point of this short excursion into mathematics is to note the differences between our treatment of this subject for engineering and the treatment for mathematicians. Notice that the topics of "Random Variables" occur in Chapter 5, "Expected values" and "Moment-generating function" in Chapter 21. These are all topics that we will cover in our lectures. Does this mean that we are jumping straight into very advanced and intricate mathematical concepts, bypassing 20 preliminary chapters that mathematicians require before delving into these subjects? Not really: our treatment will be based on a simplification, and this simplified approach luckily covers most applications of probability of interest to engineers and physicists. I would love to be able to tell you that you will never need a more advanced treatment of probabilities, but that's not necessarily true. There exist applications of engineering, physics and economics where our intuitive understanding no longer applies. My purpose in showing you the contents of [Bil95] is for you to get a feeling that probability theory is a rich and intricate mathematical theory. Knowing the limits of your understanding may help you identify cases later in life when it may become necessary to knock on the door of a friendly mathematician, or to learn the full theory yourself.

1.3   Probability or statistics?

The title of this part of the course is "Probability and Statistics". In fact, we will talk a lot more about probability than about statistics. A simplistic classification of the two fields could read something like

---

Probability theory is a branch of mathematics that deals with uncertain events

Statistics is the analysis and interpretation of data

Statistical inference, the process of deriving logical conclusions from premises in the presence of uncertainty, binds the two fields together, and can often blur the distinction between the two.

Traditionally, statistics was seen as an end in itself. Medics, economists, biologists and social scientists were taught very little probability and mostly statistics as it was considered a useful tool for them to analyse data. Once probabilities had been estimated based on data, the range of statements that one could make based on those probabilities tended to be very narrow, following prescribed recipes that give little intuition or understanding. In fact, inferring logical conclusions from data is a lot more delicate than people tended to think. There has been much public concern

about tragic misinterpretations of medical data or miscarriages of justice stemming from erroneous interpretation of data by specialists who had not been adequately trained in probability theory.

Example 1.1. "Sudden death syndrome" (SIDS) is the sudden and unexplained death of an infant. In the UK, SIDS occurs in roughly 4 out of 10,000 live births. Sir Roy Meadow, a reputed paediatrician, repeatedly argued as an expert witness in trials against parents who had lost several infants to SIDS, with what became known as "Meadow's law" *one death is a tragedy, two is suspcious and three is murder*, quoting odds of 1:73,000,000 against two SIDS in the same family in white affluent non-smoking families. Meadow's calculation of this figure from statistics he had available was erroneous and overly simplistic. The Royal Statistical Society took position against Meadow during a historic appeal, with some statisticians calculating adjusted probabilities showing there was no basis for the guilty verdict. The case led to several guilty verdicts being overturned on appeal after parents had spent years in jail falsely convicted of murdering their children on the basis of Meadow's expert opinions. Meadow appeared in front of the General Medical Council accused of serious professional misconduct.[2]

Example 1.2. A hypothetical university "C" bases its admissions decisions on a number of criteria in an elaborate and expensive selection procedure. About 0.44% of the total number of students admitted nationally studies at "C". A sharp admissions officer notices the following: of those admitted to "C", 90% use the word "volunteer" in their personal statements. Among those not admitted to "C", only 20% use the word "volunteer". Is the admissions officer onto something hot? Could "C" abolish its elaborate and expensive selection procedure and take its decisions solely on the basis of the presence of the word "volunteer" in the candidates' personal statements?

This is a typical example of a probability calculation and we will see later on in the course that Bayes' theorem[3] dictates that the probability that a candidate is admitted to "C" given the presence of the word "volunteer" in their personal statement is about 2%, much improved from the 0.44% baseline probability, but far from sufficient for a decision ignoring all other criteria.

The application of probability theory to physics and engineering has not been uncontroversial. Albert Einstein famously said "God doesn't play dice with the world". At the core of this unease, there is a fundamental tension between physics and mathematics. On the face of it, physics simply uses mathematics to establish models of the world. A model is valid for a certain range of parameters. The only test of its validity is whether it allows one to make reliable predictions about the world. On this account, a physicist should have no feelings about mathematical theories. They are neither good nor bad, only useful when models work and irrelevant otherwise. This narrow view however loses sight of intuition. Intuition is the ability of a physicist or engineer to guess the solution of a mathematical problem without doing the maths. Dirac once said[4]: "I understand what an equation means if I have a way of figuring out the characteristics of its solution without actually solving it." In order for intuition to work, there must be interaction between our physical understanding of the world and the mathematical models describing it.

Probability has challenged mathematicians, physicists and many others. Perhaps one reason is that many examples when developing the theory are associated with gambling and games of luck, traditionally frowned upon by religious and moral institutions. Another reason, ironically, is that it deals with notions such as "beliefs", traditionally frowned upon by exact scientists. A third and

---

[2]More details about this case `http://en.wikipedia.org/wiki/Roy_Meadow`
[3]Theorem named after the English theologian and mathematician Rev Thomas Bayes (1701-1761)
[4]As reported by Richard Feynman in the book "The Feynman Lectures on Physics" [FLS64].

final reason is that, as a result of the previous two reasons, probability theory was rarely taught in depth to engineers and physicists. Hence, many influential scientists never felt at ease with it. Luckily, the last reason is in the process of weakening and this IB module aims to erase it, at least as far as Cambridge engineers are concerned.

1.4   Foundations of Probability: the small print

One of the major advances in establishing probability theory is the axiomatic approach developed in the 1930s by the Russian mathematician Andrey Nikloaevich Kolmogorov[5]. This approach adopts a formal treatment that can be useful to some in developing an intuition for the field. If you like formal mathematics, read on, as you will be delighted with this section. If you don't like formal mathematics and prefer to stick to raw intuition, I will describe a shortcut in 1.5.1 at the end of Section 1.5 that will easily cover any use of probability theory you are likely to make.

Definition 1.1. A *sample space* Ω is the set of possible outcomes of a random experiment.

Example 1.3.

- When flipping a coin, $\Omega = \{\text{heads}, \text{tails}\}$.

- When throwing a dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

- When analysing financial data, Ω could be the set of all possible values of all stocks in all of the world's stockmarkets at all dates and times in the past, present and future.

- When designing a communication system, Ω could be the set of all the files a user might ever consider transmitting and all the random behaviours ("noise") the transmission medium ("channel") may possibly exhibit in the past, present or future.

- When studying a physical process, Ω could be every observable quantity in the universe.

A sample space can be a discrete finite set, a discrete countably infinite set such as the set of integers, or a continuous set such as the set of real numbers. Most examples in probability textbooks tend to concentrate on simple random experiments such as flipping a coin or throwing a dice. However, most applications of probability theory typically concern much larger sets Ω like those in our last three examples. In any case, in order for probability theory to make sense, your sample space must include all random quantities you may ever want to examine jointly. If you are going to consider two flips of a coin, then your sample space must include both flips. The concept of a "repeated random experiment" is alien to the axiomatic approach. Probablistic quantities in different sample spaces Ω_1 and Ω_2 can not be compared. The French call a sample space "un univers de probabilités" (a *universe* of probablities) and that is an appropriate description of Ω. It is the universe of anything you may ever want to consider within the same framework or study.

Definition 1.2. An *event* is a subset of Ω.

Example 1.4.

- When throwing a dice, the event $A = \{2, 4, 6\} \subset \Omega$ that the outcome is even.

- The event $B = \{1, 2, 3\}$ that the outcome is 3 or less.

---

[5]https://en.wikipedia.org/wiki/Andrey_Kolmogorov

- The event $C = \{4\}$ that the outcome is 4. This is both an outcome and an event. Some call it an "atomic" event because it contains just one outcome.

- The empty event $D = \emptyset$.

- The certain event $E = \Omega$.

The idea that an event is a subset of possible outcomes is not everyone's cup of tea. Why call it "event" if it's just a subset? Most people would not intuitively equate the notion of event with the idea of a set or subset. However, you've all seen Venn diagrams before, not least in the "Teach Yourself Probability" IA Maths Examples paper 9. Venn diagrams are simply a pictorial representation of events as sets. An example Venn diagram is drawn in Figure 1.1 The diagram
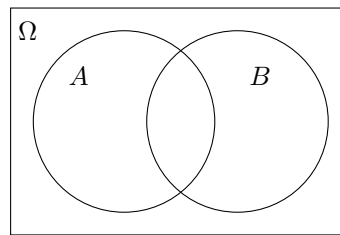

Figure 1.1: A Venn diagram

shows the sample space Ω and two events A and B. The intersection $A \cap B$, the union $A \cup B$, the complements $\bar{A}$ and $\bar{B}$ are further events that can be pictured in the Venn diagram. Suppose the random experiment is a throw of a dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $A = \{2, 4, 6\}$ and $B = \{1, 2, 3\}$. In set notation, it is obvious that $A \cap B = \{2\}$, $A \cup B = \{1, 2, 3, 4, 6\}$, $\bar{A} = \{1, 3, 5\}$ and $\bar{B} = \{4, 5, 6\}$. In "event" terminology, it may be less easy to specify some of these events. A (the event that the outcome is even), B (the event that the outcome is 3 or less), $\bar{A}$ (the event that the outcome is NOT even, i.e., odd) and $\bar{B}$ (the event that the outcome is NOT 3 or less, i.e., 4 or more) are all easy to think about. However, $A \cap B$ (the event that the outcome is even AND 3 or less) or $A \cup B$ (the event that the outcome is even OR 3 or less) are more difficult to consider intuitively. The set analogy and its pictorial representation the Venn diagram are tools that help us refine our intuition about events.

Definition 1.3. A *probability measure p* is a function that assigns numbers in $\mathbb{R}$ to events, such that the following axioms holds.

Axiom 1 For any event $A \subset \Omega$, $p(A) \geq 0$, i.e., the probability of any event is non-negative.

Axiom 2 $p(\Omega) = 1$, i.e., the probability of the certain event is 1.

Axiom 3 For any events A and B with empty intersection $A \cap B = \emptyset$,

$$p(A \cup B) = p(A) + p(B), \tag{1.1}$$

i.e., the probability of the union of disjoint events is the sum of their probabilities.

Based on these three axioms, a number of further properties of probability measures can be deduced, such as

- $p(\emptyset) = 0$.

- Complement rule: $p(\Omega - A) = p(\bar{A}) = 1 - p(A)$.

- If $A \subseteq B$ then $p(A) \leq p(B)$.

- General addition rule: $p(A \cup B) \ = \ p(A) + p(B) - p(A \cap B)$.

You will prove these statements in the Examples Paper 1. For those not familiar with axiomatic constructs, note how the three axioms are the three "truths" that must be accepted without proof as minimal conditions for the theory to hold, whereas the four bullet points above are further "truths" that can be proved as consequences of the axioms.

Example 1.5. The function assigning uniform probabilities of 1/6 to the atomic events containing the 6 outcomes of a throw of the dice is a valid probability measure:

- Since the atomic events have no intersection, the probability of any other event can be calculated using Axiom 3 and corresponds to 1/6 times the number of elements in the event.

- Axiom 1 is satisfied by definition.

- The probability of the sample space is $|\Omega| \times 1/6 = 1$, satisfying Axiom 2 (where the notation $|A|$ denotes the number of elements in set A).

This example should make it easier for you to understand the meaning of the empty event $\emptyset$ and the cetain event Ω. By definition, the random experiment is sure to have an outcome in Ω and hence the probability of the event that no outcome occurs $\emptyset$ must always be zero. The empty event is also called the *impossible* event. For the same reason, the event that any outcome in Ω occurs is certain and must hence have probability 1 by Axiom 2. Note that the probability of an intersection of events $p(A \cap B)$ is sometimes denoted

$$p(A, B)$$

and called the *joint probability* of the events.

Another property that can be deduced from Axiom 3 is called the *sum rule* for probabilities and follows from recognising that, for any two events A and B, $A \cap B$ and $A \cap \bar{B}$ has an emtpy intersection since no element can be in B and in $\bar{B}$. It follows that

$$p(A, B) + p(A, \bar{B}) = p\left((A \cap B) \cup (A \cap \bar{B})\right) = p(A). \tag{1.2}$$

Definition 1.4. The *conditional probabilty* of an event conditioned on an event of non-zero probability is defined as the joint probability divided by the probability of the event, i.e.,

$$p(A \mid B) = \frac{p(A, B)}{p(B)} = \frac{p(A \cap B)}{p(B)}. \tag{1.3}$$

If the conditioning event has probability zero, the conditional probability is undefined. The definition of the conditional probability is sometimes stated alternatively as

$$p(A, B) = p(A|B)p(B) \tag{1.4}$$

and called the *product rule* to match the sum rule above. Intuitively, the conditional probability can be seen as a way to transfer the probability measure on Ω to a re-defined random experiment where the conditional event is the new sample space.

Example 1.6. What is the conditional probability that the outcome of the dice throw will be 4 or more, given that it is even? (assume the uniform probability measure defined previously).

Using the events defined in our previous examples, the conditional probability is

$$p(\bar{B} \,|\, A) = \frac{p(\bar{B} \cap A)}{p(A)} = \frac{p(\{4, 6\})}{p(\{2, 4, 6\})} = \frac{2}{3} \tag{1.5}$$

As illustrated, the conditional probability given the event A is equivalent to re-defining a new random experiment where the dice outcome is always even, i.e., with a new sample space $\Omega' = A$, and transferring the original probability measure on Ω to this new setup. It is easy to verify that the conditional probability measure will always satisfy the axioms of a probability measure with respect to the conditional event interpreted as a sample space.

The notable *Bayes' theorem* follows directly from the definition of conditional probability:

$$p(B \,|\, A) = \frac{p(B \cap A)}{p(A)} = \frac{p(A \,|\, B)p(B)}{p(A)}. \tag{1.6}$$

Example 1.7. We can now use Bayes' theorem and the sum rule to evaluate the probability of a candidate getting admitted at hypthothetical university "C" given the use of the word "volunteer" in the candidate's personal statement.

Let A be the event that a candidate is admitted at "C". Let B be the event that the candidate has the word "volunteer" in their personal statement. We are given the following values:

- $p(A) = 0.0044$, i.e., the probability of being admitted to study at "C" nationally is 0.44%.

- $p(B \,|\, A) = 0.9$, i.e., the probability among those admitted to "C" for the word "volunteer" to have been in their personal statement is 90%.

- $p(B \,|\, \bar{A}) = 0.2$, i.e., the probablilty among those not admitted to "C" for the word "volunteer" to have been in their personal statement is 20%.

Using the sum rule, we deduce that

$$
\begin{align}
p(B) &= p(B, A) + p(B, \bar{A}) \tag{1.7} \\
&= p(B \,|\, A)p(A) + p(B \,|\, \bar{A})(1 - p(A)) \tag{1.8} \\
&= 0.9 \cdot 0.0044 + 0.2 \cdot (1 - 0.0044) = 0.2031 \tag{1.9}
\end{align}
$$

and hence, using Bayes' theorem, that

$$p(A \,|\, B) = \frac{p(A, B)}{p(B)} = \frac{0.9 \cdot 0.0044}{0.2031} = 0.0195, \tag{1.10}$$

i.e., the probability of being admitted to "C" given that a candidate has used the word "volunteer" in their personal statement is 1.95%.

1.5 Random variables

Random variables are the most useful objects of probability theory. Intuitively, random variables are numerical variables whose value is determined by underlying randomness. However, the formal definition of a random variable follows from the axiomatic approach in the previous section as follows.

Definition 1.5. A *random variable* is a scalar-valued function of the outcomes of a random experiment, i.e., a function that assigns elements in Ω to numbers

$$X : \Omega \longrightarrow \text{ numbers}$$

We will use upper case letters, e.g., X, to denote random variables. Some may find this formal definition hard to reconcile with their intuition. If X is a *function*, why in all the world did anyone think of calling it a *variable*? This inconsistency is the result of historical evolution, where random variables existed long before they were formally defined as part of the axiomatic approach. While the function/variable dichotomy may at first be confusing, it does have one major advantage in channeling our intuition towards a view of random variables as effects of an underlying common randomness. In applications of probability theory, we often handle a large number of random variables. The temptation would be to view them each as the outcomes of individual random experiments. However, probablity theory, as stated in the previous section, is unable to deal with distinct random experiments. Viewing all random variables across space and time as functions of an underlying common random experiment has the advantage that all variables remain comparable within the framework of probability theory.

Example 1.8. Consider the throw of a fair dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$.

- let X be a random variable that takes the value 0 if the outcome is even and 1 if the outcome is odd

- let Y be a random variable that takes on the value 0 if the outcome is in $\{1, 4\}$, 1 if the outcome is in $\{2, 5\}$, and 2 if the outcome is in $\{3, 6\}$.

Despite the common underlying randomness, the variables X and Y are "independent" in the sense that knowing one says nothing about the other. They act as if they were the outcome of separate random experiments, but we can only make this statement because they are in fact the outcome a common random experiment and therefore comparable within probability theory. We will formalise what we mean by "independent" random variables shortly.

Events themselves can be viewed as random variables in the sense that they are either true or false for each outcome, and could hence be equated to a random variable taking the value 1 (true) for outcomes in the event, or 0 (false) for outcomes not in the event. Such random variables are called *indicator* random variables for the event. X in the example above is an indicator random variable for the event A in the previous section.

Venn diagrams for events indicate the subset of the sample space Ω corresponding to the event, i.e., the event is "true" for outcomes of the random experiment within the subset and "false" for outcomes outside the subset. Figure 1.2 shows what the equivalent for a random variable would be, though this representation is not comonly called a Venn diagram. The random variable partitions the sample space Ω into subsets corresponding to its possible values. This pictorial representation works well for discrete random variables defined over finite alphabets, and one would need some effort of the imagination to extend the notion to discrete random variables defined over countably infinite alphabets, or continous random variables, but the principle is the same (think e.g. of a heat map over Ω for continuous random variables.)

For any random experiment with sample space Ω, conditions on random variables such as

- is X equal to 1?

- is Y smaller or equal to 1?

define events, in that the condition can be stated as "the set of outcomes for which X is 1" or "the set of outcomes for which Y is 1 or less". Hence, we can rightfully use our probability measure to
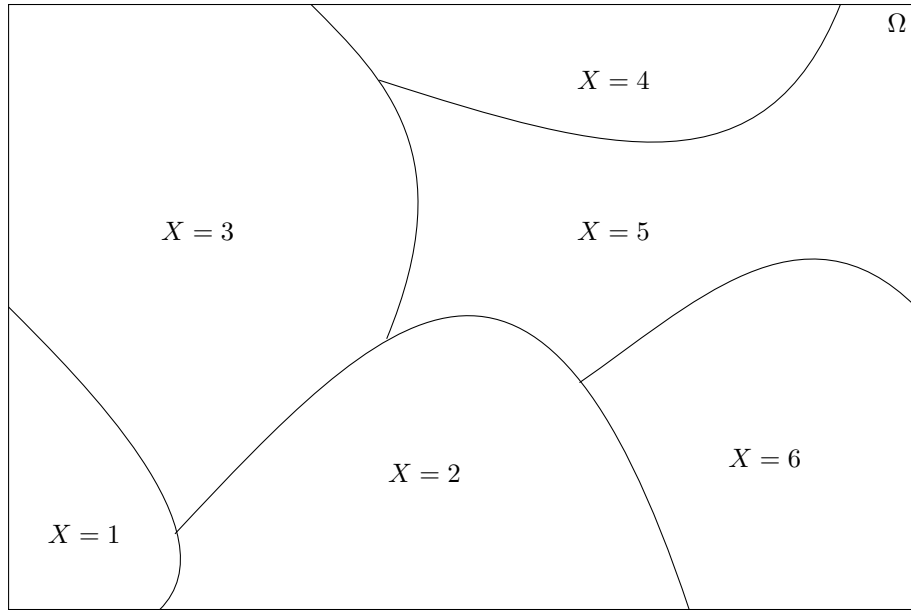
Figure 1.2: The "Venn Diagram" of a Random Variable

denote the probability of such events, e.g., $p(X = 1)$ and $p(Y \leq 1)$. We will use the notation

$$P_X(x) = p(X = x) \tag{1.11}$$

for the probability that the random variable X takes on the value x. P_X is called the *probablity distribution* or the *probability mass function* (PMF) of X. Note that we use upper case letters to denote the random variable and lower case letters to denote possible values.

We will also use the notation

$$F_X(x) = p(X \leq x) \tag{1.12}$$

and call F_X the *cumulative probablity function* of X.

We can also make statements about random variables that define combined events, such as

- is X equal to 1 AND Y equal to 1?

The event corresponding to this statement is the intersection of the two events and we use the notation

$$P_{XY}(x, y) = p(X = x \cap Y = y) \tag{1.13}$$

and call it the *joint probability distribution* of X and Y. Finally, the notion of conditional probability extends naturally to probability distributions as well and we write

$$P_{Y|X}(y \mid x) = p(Y = y \mid X = x) = \frac{p(Y = y \cap X = x)}{p(X = x)} = \frac{P_{XY}(x, y)}{P_X(x)} \tag{1.14}$$

for the *conditional probability distribution* of Y given X. As for events, this can be stated as

$$P_{XY}(x, y) = P_{Y|X}(y|x)P_X(x) \tag{1.15}$$

and is called the *product rule* when stated in this form.

Example 1.9. Consider again a fair dice throw but now let X be a random variable that takes on the value 0 if the outcome is 3 or less and 1 otherwise, and let Y be a random variable that takes on the value 0 if the outcome is 2 or less and 1 otherwise.

We have

$$\begin{cases} P_{XY}(0,0) & = p(\{1,2\}) = 1/3 \\ P_{XY}(0,1) & = p(\{3\}) = 1/6 \\ P_{XY}(1,0) & = 0, \text{ and} \\ P_{XY}(1,1) & = p(\{4,5,6\}) = 1/2. \end{cases} \tag{1.16}$$

Although the sum rule and Bayes' theorem follow immediately from the previous section, there is usually little interest in the complement event $X \neq x$ to the event $X = x$. A more interesting rule follows if we note that, if $\mathcal{Y}$ is the set of all values taken on by the random variable Y, then the events $Y = y$ for each $y \in \mathcal{Y}$ don't intersect and their union must be Ω since every outcome is assigned a value in $\mathcal{Y}$. Hence, we can use Axiom 3 as we did in the sum rule to state

$$\sum_{y \in \mathcal{Y}} P_{XY}(x,y) = p\left(\cup_{y \in \mathcal{Y}}(X = x \cap Y = y)\right) = p(X = x) = P_X(x). \tag{1.17}$$

This sum rule for probability distributions is also known as the *marginalisation* of joint probability distributions and allows us to recover probability distributions of individual variables (also called their *marginal* distributions) from their joint distributions. Bayes' theorem can be applied using marginalisation to give

$$P_{X|Y}(x \mid y) = \frac{P_{Y|X}(y \mid x)P_X(x)}{P_Y(y)} \tag{1.18}$$

$$= \frac{P_{Y|X}(y \mid x)P_X(x)}{\sum_{x' \in \mathcal{X}} P_{Y|X}(y \mid x')P_X(x')} \tag{1.19}$$

where the second step combines the marginalisation with Bayes' theorem. This final expression is the one most commonly used as Bayes' theorem as it allows you to derive the "inverse probabilities" $P_{X|Y}$ from the "forward probabilities" $P_{Y|X}$.

Example 1.10. We can marginalise the joint distribution in the previous example to obtain $P_X(0) = P_{XY}(0,0) + P_{XY}(0,1) = 1/2$ and $P_Y(0) = P_{XY}(0,0) + P_{XY}(1,0) = 1/3$. We can compute conditional probabilities, such as the probability that $Y = 0$ given $X = 0$

$$P_{Y|X}(0 \mid 0) = \frac{P_{XY}(0,0)}{P_X(0)} = \frac{1/3}{1/2} = \frac{2}{3}. \tag{1.20}$$

1.5.1   Shortcut for those who don't like the axiomatic approach

Forget about sample spaces, events, probability measures, and pretty much everything you learned in this chapter so far... Just make a list of all the random variables you will ever need, $X_1, X_2, X_3, \ldots, X_n$. Now define a joint distribution

$$P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) \text{ for all } x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \ldots, x_n \in \mathcal{X}_n, \tag{1.21}$$

that satisfies the following conditions

$$\begin{cases} P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) \geq 0 \text{ for any } x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \ldots, x_n \in \mathcal{X}_n, \\ \sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_n \in \mathcal{X}_n} P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = 1, \end{cases} \quad (1.22)$$

i.e., the joint probability distribution is non-negative everywhere and sums to 1. Once the joint probability distribution of all the random variables is defined, all other joint, conditonal and individual probability distributions can be computed using the sum rule (marginalisation) and the product rule (conditional probabilities).

This is a valid summary of probability theory that doesn't require an axiomatic construction and will accurately cover all applications of interest involving a finite collection of discrete random variables. Once you progress to continuous random variables (which we will do in this module) or to infinite collections of random variables (which you may do in Part IIA if you learn about random processes), this edifice will become a bit shaky and you may want to seek shelter in the safety of the well grounded axiomatic theory.

A more comprehensive treatment based on this approach is is [Mac03, Chapter 2], which is recommended reading for this module.

1.6 Expectation and entropy

Now that we've defined random variables and distributions, there are a number of derived measures of a distribution that are useful. The first is the *expectation* of a random variable, also called the *mean* or *average*, defined as

$$E[X] = \sum_{x \in \mathcal{X}} x P_X(x) \quad (1.23)$$

Example 1.11. Let X be a random variable taking on the value of the outcome of a throw of a fair dice. Then

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5 \quad (1.24)$$

Let Y be the random variable taking the value 0 if the outcome is 2 or less, and 1 if the outcome is 3 or more. Then

$$E[Y] = 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3}. \quad (1.25)$$

Note that the expected value of a binary random variable such as Y is always the probability of 1, i.e., $E[Y] = P_Y(1)$.

We can also take the expectation of a function of a random variable

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P_X(x). \quad (1.26)$$

Example 1.12. For the random variable X above, we can compute

$$E[X - 1] = \sum_{k=1}^{6} (k - 1) P_X(k) = \frac{1}{6}(0 + 1 + 2 + 3 + 4 + 5) = 2.5, \quad (1.27)$$

or
$$E[2X] = \sum_{k=1}^{6} 2kP_X(k) = \frac{1}{6}(2 + 4 + 6 + 8 + 10 + 12) = 7, \tag{1.28}$$

or
$$E[X^2] = \sum_{k=1}^{6} k^2 P_X(k) = \frac{1}{6}(1 + 4 + 9 + 16 + 25 + 36) = 15.167 \tag{1.29}$$

Expectations are linear operators, which means that they will fufil the following two *linearity* properties. For any two random variables X and Y,

$$\begin{aligned}
E[X + Y] &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (x + y)P_{XY}(x, y) \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} xP_{XY}(x, y) + \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} yP_{XY}(x, y) \\
&= \sum_{x \in \mathcal{X}} xP_X(x) + \sum_{y \in \mathcal{Y}} yP_Y(y) \\
&= E[X] + E[Y]
\end{aligned}$$

where we used the marginalisation rule. For any random variable X and constant c,

$$E[cX] = \sum_{x \in \mathcal{X}} cxP_X(x) = c\sum_{x \in \mathcal{X}} xP_X(x) = cE[X].$$

Note that in general, the expectation of a product is not the product of expectations, i.e.,

$$E[XY] \neq E[X]E[Y] \text{ in general,} \tag{1.30}$$

but we will learn about a sufficient condition for random variables in the next section 1.7 where equality holds.

There are a few expectations of particular interest.

$$E[X^2] = \sum_{x \in \mathcal{X}} x^2 P_X(x) \tag{1.31}$$

is called the *second moment* of a distribution. The expectation or mean $E[X]$ is also called the *first moment*. We will learn a lot more about moments in Chapter 4. Another quantity of interest is called the *central second moment* or *variance* and is defined as

$$\text{Var}(X) = E\left[(X - E[X])^2\right].$$

By averaging the squared difference to the mean, it gives an indication of how "spread out" the distribution is. If a distribution is tightly concentrated around its mean, its variance will be small. Using linearity, we can re-write the variance as

$$\text{Var}(X) = E[X^2] - 2E\left[XE[X]\right] + E[X]^2 \tag{1.32}$$
$$= E[X^2] - 2E[X]^2 + E[X]^2 \tag{1.33}$$
$$= E[X^2] - E[X]^2. \tag{1.34}$$

We will now introduce a final expectation-related measure of a random variable that is of particular interest. Probabilities are numbers between 0 and 1. As with many quantities, there is

value in considering the logarithm of the quantity. Since the logarithm of a probability is always negative, we prefer to consider the negative logarithm or

$$\iota(x) = -\log_2 P_X(x) = \log_2 \frac{1}{P_X(x)}. \tag{1.35}$$

This quantity has a nice intuitive interpretation as a measure of our "surprise" at the outcome of a random experiment and some call it the "information content" in the value of the random variable. If the probability distribution assigns a small probability to the value, then its information content is large and we are more surprised if it occurs. If the value x has probability $1/2$, then the information content is $\iota(x) = 1$. If the value x has probability 1, then the information content is $\iota(x) = 0$ which is as low as can be, implying that we are not surprised at all when we observe this value since we believed it would occur with probability 1. The information content is not defined for values that have zero probability, but we could define it to be infinity, implying that we are infinitely surprised if the random variable takes on a value that has zero probability.

The average of the information content is a measure of our uncertainty about a random variable

$$H(X) = E[\iota(X)] = \sum_{x \in \mathcal{X}} P_X(x)\iota(x) = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{1}{P_X(x)}. \tag{1.36}$$

It was introduced by Claude Shannon and is known as Shannon's *entropy*. Its unit is the *bit*[6] when the base of the logarithm is 2. Note the subtlety of the expectation notation in the definition of the entropy, where the expression denotes the expectation of a function of X but the function makes use of the probability distribution of X. For values of X that have probability zero, the entropy expression is undefined. We can avoid this problem by noticing that $\lim_{x \to 0} x \log \frac{1}{x} = 0$ and extend the definition of the entropy by adopting the convention that "$0 \cdot \log \infty = 0$".

Example 1.13. Let X be a binary random variable. Then

$$H(X) = P_X(0) \log_2 \frac{1}{P_X(0)} + P_X(1) \log_2 \frac{1}{P_X(1)}. \tag{1.37}$$

If $P_X(0) = P_X(1) = 1/2$, we have $H(X) = 1/2 + 1/2 = 1$ bit of uncertainty about the random variable, which is the largest uncertainty we can have for a binary random variable. If $P_X(0) = 0$ and $P_X(1) = 1$, we have $H(X) = 0$ bits of uncertainty about the random variable, meaning that we are completely certain about is outcome.

Let Y be the value of a fair dice throw. Then

$$H(Y) = 6 \cdot \frac{1}{6} \log_2 6 = \log_2 6 = \frac{\log 6}{\log 2} = 2.6 \text{ bits} \tag{1.38}$$

One useful entropy value to remember is that a binary random variable with probabilities $P_X(1) = 0.11$ and $P_X(0) = 0.89$ (or vice versa) has entropy $H(X) \approx 1/2$.

Information content and entropy are the object of an application of probability theory called *information theory* that was pioneered by Shannon. It is taught in the Part IIA module 3F7 where you will learn that the entropy of a random variable is an indication of how few binary symbols you can express it in on average, and hence is a measure of central interest when designing *data compression* algorithms.

---

[6]The word *bit* has by now entered common language as describing a binary digit, but it was originally introduced by Shannon as a measure of uncertainty or information, and we will stick to its original meaning here. A binary digit is only a bit if it is equally likely to be 0 and 1.

1.7   Independence

Definition 1.6. Two events A and B are said to be *independent* if

$$p(A, B) = p(A \cap B) = p(A) \cdot p(B), \tag{1.39}$$

i.e., if their joint probability factors into the product of their individual probabilities.

This definition is easily extended to more than two events, e.g., A, B and C are independent if $p(A, B, C) = p(A)p(B)p(C)$.

Example 1.14. When throwing a fair dice[7], let A be the event that the outcome is even, and B the event that the outcome is 2 or less. We have

$$p(A, B) = p(A \cap B) = p(\{2\}) = \frac{1}{6} \tag{1.40}$$

and

$$p(A)p(B) = p(\{2, 4, 6\})p(\{1, 2\}) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \tag{1.41}$$

so the two events are independent. Knowing whether the outcome is even or odd is not helpful in determining whether the outcome is 2 or less.

The last statement in the example above is easier to relate to when considering the definition of conditional probability. If two events A and B are independent and B has non-zero probability, then

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{p(A)p(B)}{p(B)} = p(A) \tag{1.42}$$

and hence the probability of A knowing B is the same as the probability of A without knowing B. This is taken by some as the definition of independence, but has the drawback that it relies on conditional probability which is only defined when the conditioning event has non-zero probability, whereas the definition we gave is more general.

Extending the concept of independence to random variables requires some further thought. Let X and Y be random variables taking on values over the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let x_1 and x_2 be two elements of $\mathcal{X}$ and y_1 and y_2 be two elements of $\mathcal{Y}$. Remember that the probability distribution of a random variable was derived by arguing that $X = x_1$ defines the event (set) of outcomes for which X takes on the value x_1 and hence $P_X(x_1) = p(X = x_1)$. Now, it may well be that $X = x_1$ and $Y = y_1$ define two independent events

$$P_{XY}(x_1, y_1) = p(X = x_1 \cap Y = y_1) = p(X = x_1)p(Y = y_1) = P_X(x_1)P_Y(y_1) \tag{1.43}$$

whereas the two events $X = x_2$ and $Y = y_2$ are not independent, i.e.,

$$P_{XY}(x_2, y_2) = p(X = x_2 \cap Y = y_2) \neq p(X = x_2)p(Y = y_2) = P_X(x_2)P_Y(y_2). \tag{1.44}$$

In this case, although some of the events defined by the random variables are independent, we will not say that the random variables are independent.

Definition 1.7. Two random variables X and Y are independent if all the events corresponding to values of X are independent of all the events corresponding to values of Y, i.e., if

$$P_{XY}(x, y) = p(X = x \cap Y = y) = p(X = x)p(Y = y) = P_X(x)P_Y(y) \tag{1.45}$$

holds for all (x, y) in $\mathcal{X} \times \mathcal{Y}$.

---

[7]Dice or die? According to Oxford Dictionaries, in modern standard English, the singular "die" is uncommon and "dice" is used for both the singular and the plural.

As for independent events, for independent random variables X and Y the relation

$$P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)} = \frac{P_X(x)P_Y(y)}{P_Y(y)} = P_X(x) \tag{1.46}$$

for all x, y such that $P_Y(y) > 0$.

Note that the "for all" statement at the end of the definition implies that, if X has n_X possible values and Y has n_Y values, then the joint probability distribution of X and Y must fulfil $n_X \times n_Y$ conditions in order for X and Y to be independent, whereas events only needed to satisfy a single condition in order to be independent.

Thinking further: an event is essentially a binary (indicator) random variable. The joint distribution of two binary random variables must satisfy $2 \times 2 = 4$ conditions for independence, whereas the associated events need to satisfy only one condition. Clearly, 3 of the 4 conditions for the random variable are redundant. Can you tell why? Think of Axioms 2 and 3. The same argument implies that some of the $n_X \times n_Y$ conditions in the general case are redundant. How many?

For independent random variable, we can express the expectation of the product as the product of expectations, i.e., if $Z = XY$ for indpendent random variables X and Y, then

$$E[Z] = E[XY] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} xy P_{XY}(x,y) \tag{1.47}$$

$$= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} xy P_X(x) P_Y(y) \tag{1.48}$$

$$= \sum_{x\in\mathcal{X}} x P_X(x) \sum_{y\in\mathcal{Y}} y P_Y(y) = E[X]E[Y]. \tag{1.49}$$

Note again that this *not* true in general of two random variables that are not independent. Independence is a sufficient condition but not a necessary condition for $E[XY] = E[X]E[Y]$ to hold. Random variables X and Y for which $E[XY] = E[X]E[Y]$ are called *uncorrelated.* Independent random variables are always uncorrelated, but it is possible in theory for two variables to be uncorrelated but not independent. Think of uncorrelation as a "cheap independence": it is easier to verify (just one condition as opposed to conditions for all x and y) and it hints at possible indepdence but without guaranteeing it.

Before we move on, we will give a quick thought to independence when more than two random variables are involved. If three or more random variables satisfy the condition

$$P_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = P_{X_1}(x_1) P_{X_2}(x_2) \cdots P_{X_n}(x_n) \tag{1.50}$$

for all $(x_1, x_2, \ldots, x_n)$ in $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$, the it is easy to see, by applying the sum rule, that any X_j and X_k for $j \neq k$ are independent. We say that the random variables are *mutually independent.* However, if

$$P_{X_j X_k}(x_j, x_k) = P_{X_j}(x_j) P_{X_k}(x_k) \tag{1.51}$$

for all j,k, and (x_j, x_k) in $\mathcal{X}_j \times \mathcal{X}_k$, but (1.50) is not fulfilled, then we say that the random variables are only *pairwise independent.*

Example 1.15. Let X and Y be binary random variables indicating "heads" in a random experiment involving two fair coins thrown independently. X is one if the first coin shows "heads" and zero otherwise, and Y is one if the second coin shows "heads" and zero otherwise. Let Z be a binary random variable obtained from XORing X and Y, where the XOR operation

---

gives 1 if either X or Y are 1 but not both, and 0 if both X and Y are zero or both are one. The joint distribution of X,Y,Z is

$$
\begin{cases}
P_{XYZ}(0,0,0) & = 1/4 \\
P_{XYZ}(0,0,1) & = 0 \\
P_{XYZ}(0,1,0) & = 0 \\
P_{XYZ}(0,1,1) & = 1/4 \\
P_{XYZ}(1,0,0) & = 0 \\
P_{XYZ}(1,0,1) & = 1/4 \\
P_{XYZ}(1,1,0) & = 1/4 \\
P_{XYZ}(1,1,1) & = 0
\end{cases}
\tag{1.52}
$$

We obtain $P_Z(0) = 1/2$ by marginalisation of the joint distribution. We see that the three random variables are *not mutually independent* by observing $P_X(0)P_Y(0)P_Z(0) = (1/2)^3 = 1/8$ which is not equal to $P_{XYZ}(0,0,0)$. This is not surprising since Z is a function of X and Y and hence X and Y fully determine Z and hence cannot be independent of Z.

On the other hand, we can obtain the joint distribution of X and Z by marginalising over Y,

$$
\begin{cases}
P_{XZ}(0,0) & = P_{XYZ}(0,0,0) + P_{XYZ}(0,1,0) = 1/4 \\
P_{XZ}(0,1) & = P_{XYZ}(0,0,1) + P_{XYZ}(0,1,1) = 1/4 \\
P_{XZ}(1,0) & = P_{XYZ}(1,0,0) + P_{XYZ}(1,1,0) = 1/4 \\
P_{XZ}(1,1) & = P_{XYZ}(1,0,1) + P_{XYZ}(1,1,1) = 1/4
\end{cases}
\tag{1.53}
$$

and hence verify that X and Z are independent. Y and Z are also independent by symmetry and X and Y are independent by definition, so we conclude that X, Y and Z are *pairwise independent*.

Example 1.16. What is the probability that two or more people have the same birthday in a group of n people? This problem is sometimes called the "birthday paradox" because the probability is higher than most people would expect. A few (wrong) assumptions are made to make the problem easy:

- Leap years and the 29 February are ignored[8].

- It is assumed that the probability that someone has their birthday on any day of the year is 1/365, i.e., birthdays are equally probable to be on any day of the year.

- The birthdays in the group are assumed to be independent.

Let us write $B_1, B_2, \ldots, B_n$ for the random variables corresponding to the birthdays of the n people in the group. The event that two or more people have the same birthday is the complement of the event that all people have different birthdays, whose probability can be computed as

$$
\sum_{k_1} \sum_{k_2 \notin \{k_1\}} \sum_{k_3 \notin \{k_1,k_2\}} \cdots \sum_{k_n \notin \{k_1,\ldots,k_{n-1}\}} P_{B_1}(k_1) P_{B_2|B_1}(k_2|k_1) \cdots P_{B_n|B_1\ldots B_{n-1}}(k_n|k_1 \ldots k_{n-1}).
\tag{1.54}
$$

Since all the random variables are independent and uniform, all probabilities in the expression are 1/365, so it is only a matter of taking sums over the sets. The sum over k_1 is over all

365 birthdays, the sum over k_2 over all 364 birthdays not equal to k_1, etc. The resulting expression is hence

$$p(n \text{ different birthdays}) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{366 - n}{365} \tag{1.55}$$

$$= \frac{365!}{365^n (365 - n)!} \tag{1.56}$$

Hence, the probability that two or more people have the same birthdays

- among all 309 IB students enrolled in Lent 2021 is $1 - 10^{-89}$, practically 1;

- among the 15 Homerton IB students is $1 - 0.75 = 0.25$ so $1/4$;

- among the 21 Trinity IB students is $1 - 0.56 = 0.44$, soi almost $1/2$ (the $1/2$ limit is passed for 22 students).

1.8 Summary

Introduction:

- Probability theory is a branch of mathematics that deals with uncertain events.

- Statistics is the analysis and interpretation of data.

The axiomatic approach:

- A *sample space* Ω is the set of possible outcomes of a random experiment.

- An *event* is a subset of Ω.

- A *probability measure* p is a function that assigns numbers in $\mathbb{R}$. to events.

- Axiom 1: For any event $A \subset \Omega$, $p(A) \geq 0$, i.e., the probability of any event is non-negative.

- Axiom 2: $p(\Omega) = 1$, i.e., the probability of the certain event is 1.

- Axiom 3: For any events A and B with empty intersection $A \cap B = \emptyset$, $p(A \cup B) = p(A) + p(B)$.

- $p(\emptyset) = 0$.

- Complement rule: $p(\Omega - A) = p(\bar{A}) = 1 - p(A)$.

- If $A \subseteq B$ then $p(A) \leq p(B)$.

- General addition rule: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

- $p(A \cap B)$ is sometimes denoted $p(A, B)$

- Sum Rule: $p(A, B) + p(A, \bar{B}) = p(A)$.

---

[8]Apologies to all those whose birthday is the 29 February!

- Product Rule: $p(A, B) = p(A|B)p(B)$, equivalent to the definition of conditional probability: $p(A \mid B) = p(A, B)/p(B)$.

- Bayes' theorem: $p(B \mid A) = p(A \mid B)p(B)/p(A)$.

Random variables:

- Probability mass function (PMF, or distribution): $P_X(x) = p(X = x)$

- Cumulative probability distribution: $F_X(x) = p(X \leq x)$

- Joint probability: $P_{XY}(x, y) = p(X = x \cap Y = y)$

- Product rule: $P_{XY}(x, y) = P_{Y|X}(y|x)P_X(x)$

- Sum rule (or marginalisation): $\sum_{y \in \mathcal{Y}} P_{XY}(x, y) = P_X(x)$

- Bayes' theorem: $P_{X|Y}(x \mid y) = \frac{P_{Y|X}(y \mid x)P_X(x)}{\sum_{x' \in \mathcal{X}} P_{Y|X}(y \mid x')P_X(x')}$

Expectation and entropy:

- Expectation: $E[f(X)] = \sum_{x \in \mathcal{X}} f(x)P_X(x)$, in particular $E[X] = \sum_{x \in \mathcal{X}} xP_X(x)$ and $E[X^2] = \sum_{x \in \mathcal{X}} x^2 P_X(x)$.

- Linearity of expectation: $E[X + Y] = E[X] + E[Y]$ and $E[cX] = cE[X]$ for any c

- Variance: $\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$

- Entropy: $H(X) = E[\log_2(1/P_X(x))] = -\sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x)$

Independence:

- Independent events: $p(A, B) = p(A \cap B) = p(A) \cdot p(B)$ and hence $p(A|B) = p(A)$

- Independent random variables: $P_{XY}(x, y) = P_X(x)P_Y(y)$ for all x, y and hence $P_{X|Y}(x|y) = P_X(x)$

- For independent random variables X and Y, $E[XY] = E[X]E[Y]$.

1.9   Problems

Having completed Chapter 1, you should be able to attempt Problems 1 to 5 of Examples Paper 5.

Chapter 2

Discrete Probability Distributions

The previous chapter was loaded with concepts that were new for many of you. This chapter will be much easier to digest as we simply aim to get familiar with a few common probability distributions and illustrate them with examples. Note that a comprehensive list of known distributions would take far more time that we can afford to spend in this course and learning them by heart would also add little educational value, so we picked a few essential distributions that are worth knowing about. If you ever need to know about other distributions that the ones presented in this lecture, a good place to start is the Wikipedia list of probability distributions.

2.1 The Bernoulli Distribution

We begin with a simple binary distribution. A binary random variable X with a probability distribution $P_X(1) = p$ and $P_X(0) = 1 - p$ is said to have a *Bernoulli* distribution[1] with parameter p, denoted

$$X \sim \text{Ber}(p). \tag{2.1}$$

Bernoulli distributions occur in many scenarios:

- As mentioned in the previous chapter, they are indicator random variables for events, e.g.,

    - Will it rain tomorrow?
    - Will the UK economy grow above expectation?
    - Will the message I transmit be received and decoded correctly?

- They occur naturally as answers to yes/no questions, e.g.

    - Is the product defective?
    - Did the defendant murder the victim?
    - Will you marry me?

- They also occur in their own right in digital communications, where information is often encoded into binary symbols.

- Probability textbooks often illustrate Bernoulli distributions using "biased coins". These are coins that have different probabilities of landing on "heads" or "tails".

---

[1]Named after the Swiss scientist Jacob Bernoulli (1665-1705).

How to make a biased coin: have tou ever seen a biased coin in reality? If you type "how to make a biased coin" into a search engine you get lots of ingenious suggestions, such as bending the coin. My doctoral supervisor Jim Massey was given an extreme version of a biased coin as a retirement present: this had been manufactured by filing down two coins and gluing them together to yield something looking like a coin giving a random variable $X \sim \text{Ber}(1)$ that would always indicate "heads". *Disclaimer:* please check the legality of defacing coins in your jurisidiction before rushing to produce your own "Massey coin". In the UK as far as I am aware it is legal to do so at the time of writing.

We've already established that

$$E[X] = P_X(1) = p \tag{2.2}$$

for any binary random variable in the previous chapter. The other expectations of interest are the second moment.

$$E[X^2] = P_X(0) \cdot 0^2 + P_X(1) \cdot 1^2 = p \tag{2.3}$$

and the variance,

$$\text{Var}(X) = E[X^2] - E[X]^2 = p - p^2 = p(1 - p). \tag{2.4}$$

The entropy of a Bernoulli random variable is known as the *binary entropy function* of p

$$H(X) \;=\; H_2(p) = \; P_X(0)\log_2\frac{1}{P_X(0)} + P_X(1)\log_2\frac{1}{P_X(1)} \;=\; p\log_2\frac{1}{p} + (1-p)\log_2\frac{1}{1-p} \tag{2.5}$$

with $H_2(0) = H_2(1) = 0$ using our rule of "$0\log 0 = 0$". A plot of the binary entropy function is represented in Figure 2.1. Notice that our uncertainty about a binary random variable is highest when it is equally likely to be 1 or 0, and at its lowest when it is certain to be 0 or 1. Our uncertainty is about half a bit when the probability of 1 or 0 is about 0.11.
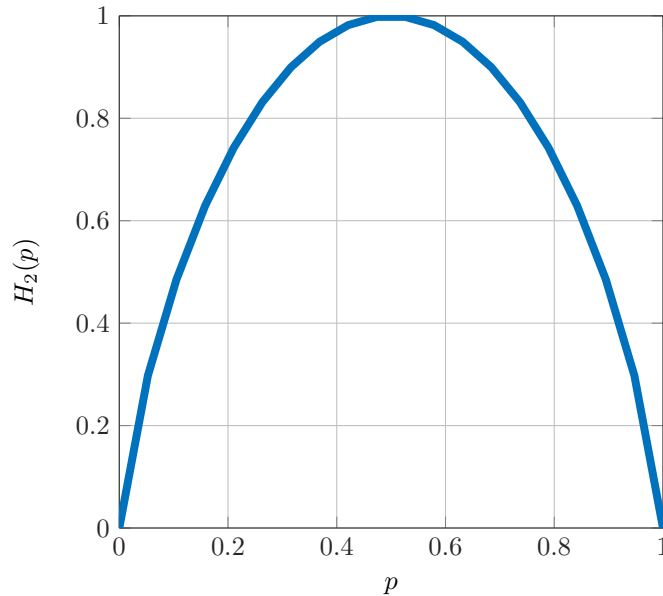

Figure 2.1: The binary entropy function $H_2(p)$

2.2 The Binomial Distribution

Consider n independent $\text{Ber}(p)$ distributed random variables $X_1, X_2, \ldots, X_n$. The random variable $Y = \sum_{k=1}^{n} X_k$ is said to follow a *binomial* distribution with parameters n and p, denoted

$$Y \sim B(n,p). \tag{2.6}$$

It is the number of ones, "yes" answers, or successes in n independent Bernoulli trials. Note that each individual sequence of values $x_1, x_2, \ldots, x_n$ has probability

$$P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{k=1}^{n} P_{X_k}(x_k) = p^{\sum_k x_k}(1-p)^{n-\sum_k x_k} \tag{2.7}$$

in other words its probability only depends on the number of ones and number of zeros. For example,

$$P_{X_1 X_2 X_3 X_4 X_5}(0,1,0,1,1) = p^3(1-p)^2 \tag{2.8}$$

because the sequence contains three ones and two zeros. The number of sequences of length n that have k ones and $n-k$ zeros (when order is irrelevant) is the number of combinations

$$\binom{n}{k} = {}^nC_k = \frac{n!}{(n-k)!k!} \tag{2.9}$$

where the first notation is preferred by engineers mathematicians and scientists, and the second notation is preferred by designers of calculators because the first notation doesn't fit on a calculator key. We will use both notations interchangeably.

We've now deduced an expression for the binomial distribution

$$P_Y(k) = \binom{n}{k}p^k(1-p)^{n-k} \tag{2.10}$$

for $k = 0, 1, \ldots, n$.[2] We can easily verify that the distribution sums to one using the binomial expansion[3] as follows

$$(p+1-p)^n = 1 = \binom{n}{0}p^n(1-p)^0 + \binom{n}{1}p^{n-1}(1-p) + \binom{n}{2}p^{n-2}(1-p)^2 + \ldots + \binom{n}{n}p^0(1-p)^n \tag{2.11}$$

where the expression on the right is precisely the sum of the binomial distribution.

Example 2.1. A hypothetical student "C." has a probability of $p = 1/3$ of failing to wake up in time for lectures on any given morning. What is the probability that "C." attends[4] k of the six IB Paper 7 "probability and statistics " lectures?

The random variable Y counting the number of lectures attended is binomial $B(6, 1/3)$. Figure 2.2 illustrates the distribution graphically. Observe that this particular distributon has two modes (largest probabilities) at 1 and 2 lectures. In general, the mode of a random variable $X \sim B(n,p)$ is the largest integer smaller than np, except when np is an integer as is the case here, when you get two modes at np and $np - 1$. We will compute the mean and standard deviation of binomial distributions below.

---

[2]Note that a binomial random variable $Y \sim B(n,p)$ which is the sum of n Bernoulli random variables can take on $n+1$ values from 0 to n.

[3]We use a slightly different binomial expansion than the one in your data book here, that can easily be derived from it: $(x+y)^n = x^n(1+y/x)^n = x^n(1+ny/x + {}^nC_2(y/x)^2 + \ldots) = x^n + nx^{n-1}y + {}^nC_2 x^{n-2}y^2 \ldots$

[4]*Disclaimer:* this example was not modified for the Pandemic year. In fact, this year, "C." oversleeps because they were up until 4 a.m. watching my fascinating video lectures, so please be assured of my warmest appreciation!
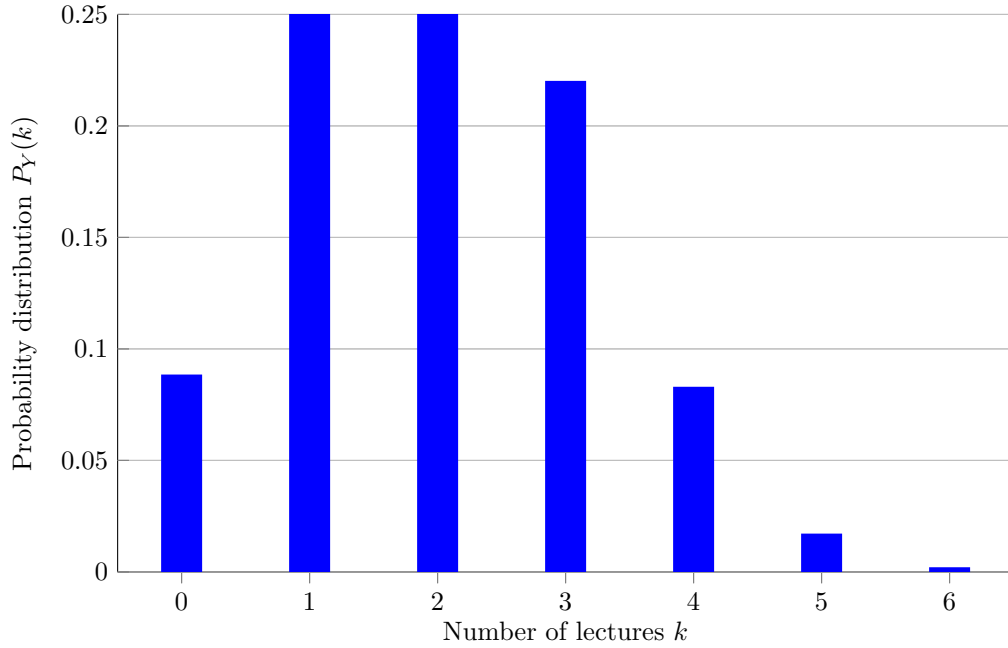
Figure 2.2: Binomial distribution $B(6, 1/3)$

Example 2.2. Sometimes the binomial distribution results in counter-intuitive conclusions, like the following example loosely borrowed from [Ros72]. Suppose airplanes are designed to remain airborne as long as half of their engines are still operational. This means that an airplane with two engines is able to fly with one, and an airplane with four engines is able to fly with two or more. Suppose the probability of failure for each engine is independent and equal to p. Would you prefer to fly with a two engine plane or a four engine plane?

The number Y of engines failing follows a binomial distribution $B(n, p)$ where n is the total number of engines. The probability of remaining airborne is the sum of the distribution for values half of n and less, i.e., for $n = 2$

$$p(\text{airborne}) = P_Y(0) + P_Y(1) = \binom{2}{0} p^0 (1-p)^2 + \binom{2}{1} p^1 (1-p)^1 = (1-p)^2 + 2p(1-p) = 1 - p^2,$$
(2.12)

and for $n = 4$,

$$p(\text{airborne}) = P_Y(0) + P_Y(1) + P_Y(2) = (1-p)^4 + 2p(1-p)^3 + 6p^2(1-p)^2 = 1 - 4p^3 + 3p^4. \quad (2.13)$$

Equating the two expressions leads to the surprising conclusion that four engines are not always better than two engines: there is a critical point at $p = 1/3$ above which it is preferrable to fly in a two engine plane. Personally, I would not be enclined to fly on any airplane whose engines have such a high probability of failure, but the example nonetheless illustrates an interesting point.

The expected value of the binomial distribution can be computed tediously using the definition of the expectation

$$E[Y] = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \tag{2.14}$$

or, alternatively, we can use the linearity properties developed in the previous section to note that, if $Y = X_1 + X_2 + \ldots + X_n$ where each X_k is $\text{Ber}(p)$, then

$$E[Y] = E[X_1] + \ldots + E[X_n] = nE[X_1] = np. \tag{2.15}$$

We can compute the second moment of a binomial distribution using the expectation product rule for independent random variables

$$E[Y^2] = E[(X_1 + \ldots + X_n)^2] = nE[X_1^2] + 2\binom{n}{2}E[X_1]^2 = np + n(n-1)p^2 \tag{2.16}$$

and the variance follows as

$$\text{Var}(Y) = E[Y^2] - E[Y]^2 = np + n(n-1)p^2 - (np)^2 = np(1-p). \tag{2.17}$$

The entropy of a binomial distribution is not easy to calculate.

Note that binomial distributions are difficult to estimate numerically for large n due to the difficulty of evaluating ${}^nC_k$ for large n. A few approximations are handy in addressing this. Stirling's approximation for factorials is

$$n! \simeq \sqrt{2\pi n}\exp(-n)n^n. \tag{2.18}$$

MacKay's book [Mac03] on page 2 gives the following approximation

$$\binom{n}{k} \simeq 2^{nH_2(k/n)} \tag{2.19}$$

following a simple derivation. Finally, the Poisson distribution introduced in the last section of this chapter gives a good approximation of the binomial distribution for small p and large n.

2.3 The Geometric Distribution

The next two distributions we will study concern random variables taking on values in the countably infinite set of integers $1, 2, 3, \ldots$ as opposed to the Bernoulli and binomial distributions that both had finite sets of possible values. The geometric distribution can also be derived from a collection of independent Bernoulli random variable as the distribution of the index of the first 1 in the sequence. Hence, Y is a geometric distributed random variable

$$Y \sim \text{Geom}(p) \tag{2.20}$$

derived from an infinite collection $X_1, X_2, \ldots$ of independent $\text{Ber}(p)$ random variables, and therefore

$$P_Y(k) = P_{X_1}(0)P_{X_2|X_1}(0\,|\,0)\cdots P_{X_k|X_1\ldots X_{k-1}}(1\,|\,0,\ldots,0) \tag{2.21}$$

$$= P_{X_1}(0)P_{X_2}(0)\cdots P_{X_{k-1}}(0)P_{X_k}(1) \tag{2.22}$$

$$= p(1-p)^{k-1}, \text{ for } k = 1, 2, 3, \ldots \tag{2.23}$$

i.e., the probability that the first $k-1$ variables are zero and that the k-th variable is one. We leave it as an exercise to verify that

$$E[Y] = \frac{1}{p} \tag{2.24}$$

Hint: use the expression for the sum of a geometric sequence in the maths data book and note that

$$\frac{d}{dr}S_\infty = \frac{d}{dr}\frac{a}{1-r} = \frac{d}{dr}\sum_{k=0}^{\infty}ar^k. \tag{2.25}$$

The variance can also be determined by repeated use of the derivation trick above to yield

$$\text{Var}(Y) = \frac{1-p}{p^2}. \tag{2.26}$$

The entropy of a geometric distribution can also be derived easily using sums of geometric distributions and the derivation trick to yield

$$H(Y) = \frac{H_2(p)}{p}. \tag{2.27}$$

Example 2.3. The geometric distribution often comes into play in repeated random experiments as the distribution of the number of attempted until success or failure. Consider for example trying to generate a random point in the disc of radius r by using Python's pseudo-random generator `random.uniform(-r,r)` twice, obtaining two random variables (the x coordinate and the y coordinate) ranging between $-r$ and r. For the purposes of this example, let us assume that Python's pseudo-random generator is a perfect generator of uniformly distributed random variables (these are continuous random variables, about which we will learn more in the next chapter). If the resulting point is not in the unit circle, i.e., if $x^2 + y^2 > r$, the attempt is discarded and two new random numbers need to be generated, and so forth until the point obtained is in the circle. Let Z be the number of attempts until we successfully generate a point in the circle. The probability of success at every attempt is the surface of the disk of radius r divided by the surface of the square of side $2r$, i.e.,

$$p(\text{"success"}) = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4} = 0.7854 \tag{2.28}$$

Z is geometrically distributed with $p = \pi/4$. The number of attempts needed averages $E[X] = 1/p = 4/\pi = 1.27$. The distribution of attempts is illustrated in Figure 2.3.

Note that some textbooks define the geometric distribution as $P_Y(k) = p(1-p)^k$ for $k = 0, 1, 2, 3\ldots$, so starting with 0 instead of 1 as we did. These definitions are equivalent in all respects except that you have to take 1 from the expectation if you start at 0 instead of 1, giving $E[X] = 1/p - 1$. The variance and entropy remain the same. Your mathematics data book has both options, called Geometric (1) (starting from zero) and Geometric (2) (starting from 1 as presented here.)

2.4 The Poisson Distribution

The Poisson distribution, named after the French scientist Siméon Poisson[5], is used to model the probability of the number of incidents[6] in a time interval when incidents happen independently at a given rate of λ incidents per time interval. The incidents in this context are assumed to be of zero duration.

---

[5]Pronounced "pwa sô" where the last vocal is a nasal o. I have attended a talk by a reputed mathematician who spent his life studying Poisson processes yet kept calling them "poison processes". "Poisson" means fish in French and there is nothing poisonous about fish as long as it's fresh.

[6]In most probability textbooks the Poisson distribution is described as modeling the number "events" that occur in a time interval. The problem is that we've given the word "event" a very precise meaning in probability theory, but in this case the word is used in its colloquial sense "a thing that happens at a given time". We will use the synonym "incident" to avoid confusion even though it feels a bit clonky.
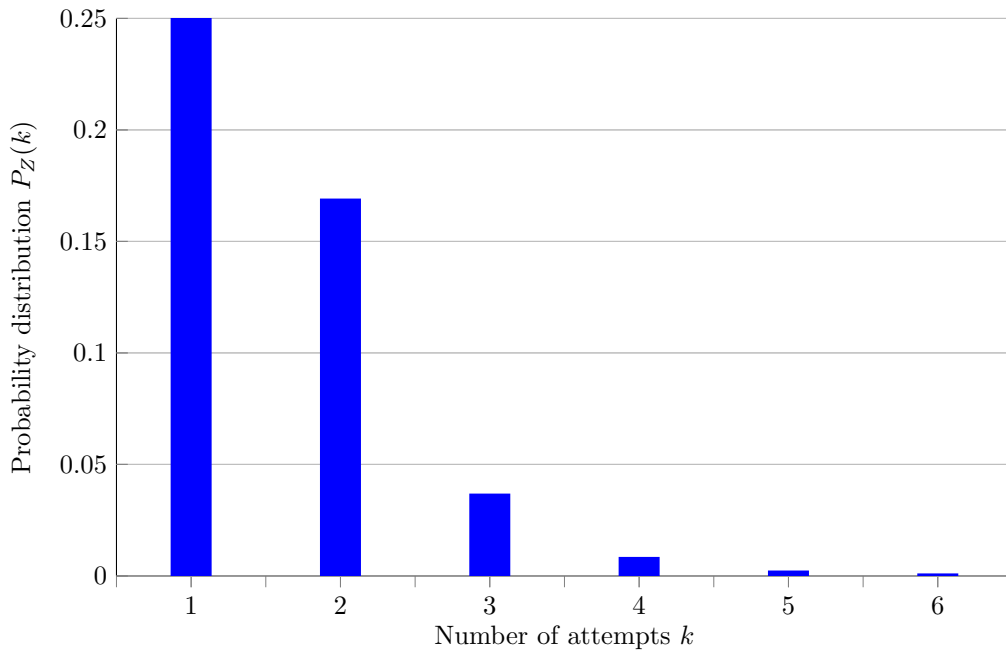
Figure 2.3: Geometric distribution for the number of attempts to generate a point in a circle

---

Example 2.4. A certain model of 10 Gb/s Ethernet router handles packets at an average rate of 5×10^6 packets per second. What is the probability distribution of the number of packets handled in any given microsecond (μs)?

The number of packets handled in a μs follows a Poisson distribution with parameter $\lambda = 5$.

Example 2.5. A bridge is being built for a new motorway linking Cambridge to Milton Keynes, that will carry traffic at an average rate of 10 vehicles per minute. What is the probability distribution for the number of vehicles the bridge has to carry in any given hour?

The number of vehicles in an hour follows a Poisson distribution with parameter $\lambda = 600$. Note that this ignores predictable rate fluctuations according to time of day, season, etc. The structural engineers building the bridge would do well to dimension the structure using more accurate traffic models than just the average rate of traffic.

Rather than postulate an expression for the Poisson distribution, we will derive it from first principles. Let Y be the random variable counting the number of incidents in the time interval of interest,

$$Y \sim \text{Poisson}(\lambda). \tag{2.29}$$

We will refer to the time interval of interest as the "unit interval" (e.g. $1\mu s$.) Let us now subdivide the unit interval into n equal sub-intervals. If there are λ incidents per unit interval on average, there will be λ/n incidents in each sub-interval on average. If we pick n large enough so that $\lambda/n << 1$, since each interval can only contain an integer number of incidents, most intervals will contain no incident and some may contain one. The probability that any interval contains two or more incidents is expected to go to zero as n gets large enough. Hence, we can define indicator random variables $X_1, X_2, \ldots, X_n$ that are 1 if an incident occurs in the corresponding
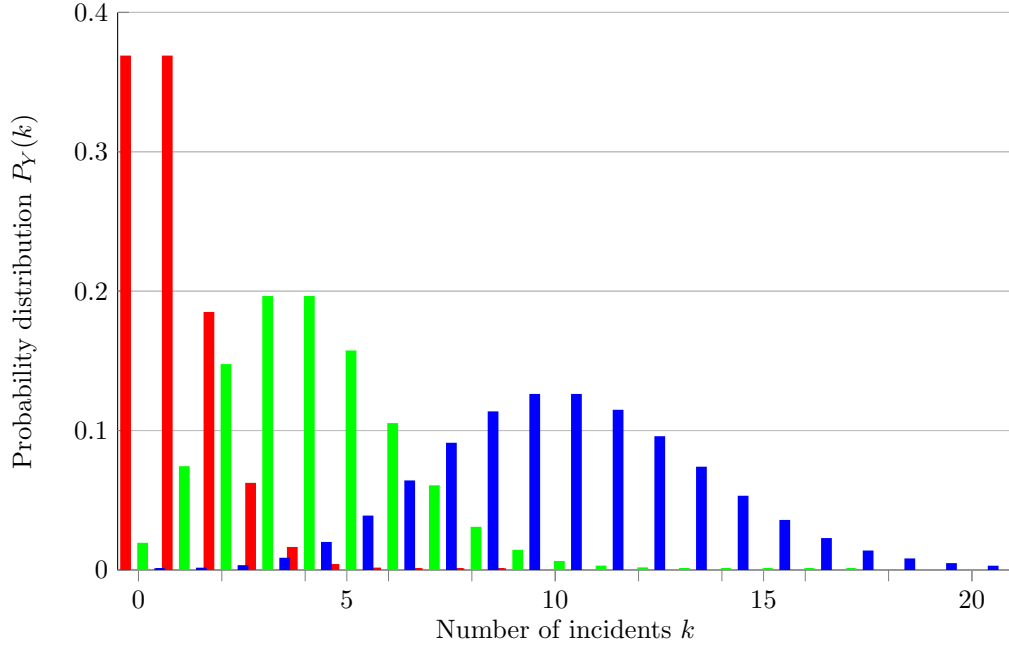
Figure 2.4: Poisson distributions for $\lambda = 1$ (red), $\lambda = 4$ (green) and $\lambda = 10$ (blue)

sub-interval and 0 otherwise. This is therefore well approximated by a sequence of independent Bernoulli random variables with parameter λ/n, i.e., $X_k \sim \text{Ber}(\lambda/n)$ for $k = 1, \ldots, n$, and the approximation is exact in the limit as n goes to infinity. Since $Y \simeq X_1 + X_2 + \ldots + X_n$, Y follows a binomial distribution $Y \sim B(n, \lambda/n)$. The Poisson distribution is the limit of a binomial distribution $B(n, \lambda/n)$ as n goes to infinity. We need to evaluate

$$P_Y(k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}. \tag{2.30}$$

For a fixed k,

$$\lim_{n\to\infty} \binom{n}{k} \frac{1}{n^k} = \lim_{n\to\infty} \frac{n!}{n^k (n-k)! k!} = \lim_{n\to\infty} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \frac{1}{k!} = \frac{1}{k!}. \tag{2.31}$$

Furthermore, using a limit from the Mathematics Data Book (page 2), we have

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n-k} = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} = e^{-\lambda} \cdot 1 \tag{2.32}$$

giving

$$P_Y(k) = \frac{\lambda^k}{k!} e^{-\lambda} \text{ for } k = 0, 1, 2, 3 \ldots \tag{2.33}$$

Since we've derived the Poisson distribution as an approximation to a binomial, it is clear that the Poisson distribution can also be used in reverse to approximate a binomial distribution. This works for small p and large n and is useful when the binomial distribution is difficult to compute numerically due to the difficulty of evaluating $\binom{n}{k}$. Hence, for large n and small p,

$$\binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{(pn)^k}{k!} e^{-pn}, \tag{2.34}$$

i.e., a Poisson distribution with parameter $\lambda = pn$.

Poisson distributions for various parameters λ are shown in Figure 2.4. The expectation of the Poisson distribution is λ by definition, since we assumed that λ was the average number of incidents per time interval when deriving the distribution. For peace of mind, it's worth double-checking by recomputing the expectation from first principles, which is left as an exercise (when doing so, look out for an expression that turns out to be the power series of the exponential function.)

The Poisson distribution unusally has identical expectation and variance (which we won't derive but is again a fun exercise if you enjoy juggling around with algebra and derivatives)

$$E[Y] = \text{Var}(Y) = \lambda. \tag{2.35}$$

This is a very useful property and often used in data analysis, as the expectation and the variance are easy to estimate from data and checking if the estimates are close is a good way to validate a hypothesis that the data source generates Poisson distributed random variables. The entropy is more difficult to derive and does not have simple expression.

We will revisit Poisson processes in the next chapter on continuous random variables when we study the probabilistic properties of the time interval between incidents in a Poisson process.

2.5 Summary

Bernoulli distribution:

- $X \sim \text{Ber}(p)$, $P_X(1) = 1 - P_X(0) = p$.
- $E[X] = p$, $E[X^2] = p$ and $\text{Var}(X) = p(1-p)$
- Binary entropy function: $H(X) = H_2(p) = -p \log_2 p - (1-p) \log_2(1-p)$

Binomial distribution:

- $Y \sim B(n,p)$, $P_Y(k) = \binom{n}{k} p^k (1-p)^{n-k}$
- Binomial coefficient: $\binom{n}{k} = {}^nC_k = \frac{n!}{(n-k)!k!}$
- $E[Y] = np$, $\text{Var}(Y) = np(1-p)$
- Stirling's approximation: $n! \simeq \sqrt{2\pi n}(n/e)^n$.
- Binary entropy approximation: $\binom{n}{k} \simeq 2^{nH_2(k/n)}$
- The binomial distributions models the number of ones in n independent Bernoulli random variables with Bernoulli paramater p

Geometric distribution:

- $Y \sim \text{Geom}(p)$, $P_Y(k) = p(1-p)^{k-1}$ for $k = 1, 2, \ldots$
- $E[Y] = 1/p$, $\text{Var}(Y) = (1-p)/p^2$, $H(Y) = H_2(p)/p$

Poisson distribution

- $Y \sim \text{Poisson}(\lambda)$, $P_Y(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for $k = 0, 1, 2, \ldots$
- $E[Y] = \text{Var}(Y) = \lambda$
- Approximation of a binomial as a Poisson: if $Z \sim B(p,n)$ for a small p and a large n, then $P_Z(k) \simeq \frac{(pn)^k}{k!} e^{-pn}$, i.e., approximately $Z \sim \text{Poisson}(pn)$
- The Poisson distribution models the number of incidents in a unit interval when the incidents occur at a rate of λ per unit interval

2.6   Problems

Having completed Chapter 2, you should be able to complete problems 1 to 7 in Examples Paper 5.

Chapter 3

Continuous Random Variables

In Chapter 1, we introduced the axiomatic approach but reassured you that, for many engineering applications, it would be sufficient to think about joint distributions of random variables. A simplified probability theory is sufficient for tackling a finite number of discrete random variables defined over finite alphabets. For continuous random variables on the other hand, the axiomatic approach will help us form a better intuition. Mathematicians normally study probability theory hand in hand with measure theory, a more general framework for assigning measures (numbers) to sets, and a probability measure is just one example of such a measure. As engineers, we need to understand some continuous probabilty theory because there are topics within that theory that are of central relevance to engineering applications. We will , but we do so without studying measure theory. We hope in the two introductory sections to give you sufficient depth of understanding for you to imagine why such a theory exists and why mathematicians require it to provide a precise framework for all possible continuous random variables.

3.1   Fundamentals of Continuous Random Variables

In Section 1.5, we defined random variables as functions of outcomes of a random experiment. Discrete random variables were seen as a natural extension of the concept of "event":

- an event A partitions the sample space Ω into two regions: the set A of outcomes where the event is "true", and the complementary set $\bar{A}$ of outcomes where the event is "false";

- a discrete random variable X partitions Ω into n regions, where $n = |\mathcal{X}|$ is the number of elements in the set of possible values $\mathcal{X}$ of X. Each region corresponds to the subset of outcomes in Ω for which the random variable takes on one value in $\mathcal{X}$ (e.g., $X = 0$ for outcomes $\{1, 2\}$, $X = 1$ for outcomes $\{3, 4\}$, and $X = 2$ for outcomes $\{5, 6\}$ of a fair dice throw.)

For continuous random variables, the picture remains the same but is more difficult to imagine graphically. For one, the sample space itself must contain a continuum of possible outcomes in order for random variables defined on it to be truly continuous, so we are dealing with more complex underlying random experiments and corresponding sample spaces. The probability measure $p(.)$ will in general assign positive probabilities only to *regions* of Ω. Although single outcomes are still technically events (e.g., the atomic event $\{a\}$ containing only the outcome a of the random experiment), the probability measure for any event containing only a finite number or countably infinite number of outcomes will in general be zero in those sample apaces. For random variables, this means that the event $\{X = x\}$ that a random variable X takes on a single value x will in general have probability zero. Exceptions to this are for example random variables that remain constant in a region of Ω and hence may have positive probability for certain values. These are called *mixed discrete/continuous* random variables.

Example 3.1. A vehicle's speed can be considered a continuous random variable. What is the probability that a vehicle should be passing a speed camera at a speed *precisely* equal to the speed limit of 30.00000... mph?

   The speed X can take on a continuum of values and the probability of a single outcome $\{X = 30.00000\ldots\}$ is zero. We expect the event $\{29.9 \leq X \leq 30.1\}$ on the other hand to have non-zero probability. Note that if a car passed one speed camera A driving below 30 mph and the next speed camera B a mile down the road driving above 30 mph, then there must have been a position between A and B where the car was driving at exactly 30 mph. The fact that an event has zero probability does not necessarily mean that it will never occur given an suitably infinite set of random variables, such as the infinite set of random variables corresponding to all the speeds of the car over the interval $[A, B]$.

We will in general refrain from considering events with a finite number of outcomes and hence not consider probability distributions $P_X(x) = p(X = x)$ as we did for discrete random variables. This distribution would in general be zero everywhere and hence be of little use. However, we can still consider events corresponding to intervals, and in particular, use the cumulative probability function defined in Chapter 1

$$F_X(x) = p(X \leq x). \tag{3.1}$$

This is still well defined and satisfies the following properties

$$\begin{cases} F_X(x) \geq 0, \\ \lim_{x \to -\infty} F_X(x) = 0, \\ \lim_{x \to \infty} F_X(x) = 1, \\ F_X(x) \text{ is non-decreasing in } x. \end{cases} \tag{3.2}$$

where the first property follows from Kolmogorov's first axiom because $F_X(x)$ is the probability of a well defined event; the second and third properties follow from the fact that the limit of the sets defined by $X \leq x$ as x goes to $-\infty$ and $+\infty$ is the empty set and the sample space Ω, respectively ; and the last property follows because if $x_1 \leq x_2$ then the event $\{X \leq x_1\}$ is a subset of the event $\{X \leq x_2\}$, since for every outcome for which X is less than x_1, X is also less than x_2, and we proved in Chapter 1 that $A \subseteq B$ implies $p(A) \leq p(B)$ and hence $p(X \leq x_1) \leq p(X \leq x_2)$ or $F_X(x_1) \leq F_X(x_2)$.

   The probability of falling within an interval $[a, b]$ can be expressed in function of the cumulative probability function, i.e.,

$$p(a \leq X \leq b) = p(X \leq b) - p(X < a) = F_X(b) - F_X(a). \tag{3.3}$$

The careful reader will have noticed that we brushed over the difference between '$<$' and '$\leq$' in the expression above. This will in general make no difference as the events $\{X < a\}$ and $\{X \leq a\}$ differ in only a single point. We have seen that single outcomes or any finite number of outcomes will in general have probability zero for continuous random experiments. This distinction can become important in the case of mixed discrete/continuous random variables, and one would have to carefully adjust the expression above if a were a value of X corresponding to a region of Ω with non-zero probability.

Example 3.2. The probability that the car speed recorded by speed camera A is zero or less (assuming that the camera records the magnitude of the speed only) is 0, hence $F_X(x) = 0$ for all $x \leq 0$. The probability that the speed is less than 1000 mph is 1 as there are no cars that can drive that fast, hence $F_X(1000) = \lim_{x \to \infty} F_X(x) = 1$.

The probability that the car's speed is between 29.9 and 30.1 mph is $F_X(30.1) - F_X(29.9)$. If some cars are equipped with speed limiters set for 29.9 that physically limit the speed to that value no matter how much the driver accelerates, then the above expression may no longer apply and one would need to carefully consider whether we mean the "open" or "closed" interval between 29.9 and 30.1 and what the (now non-zero) probability of driving exactly $29.900000\ldots$ mph is.

Joint cumulative probability functions are defined similarly to joint distributions for discrete random variables

$$F_{XY}(x, y) = p(X \le x \cap Y \le y) \tag{3.4}$$

and independence of continuous random variables is defined as the independence of all events associated with the random variables, e.g.,

$$F_{XY}(x, y) = p(X \le x \cap Y \le y) = p(X \le x)p(Y \le y) = F_X(x)F_Y(y) \tag{3.5}$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where the condition "for all" specifies an infinity of conditions in the continuous case.

There are no product or sum rules for cumulative probability functions. The former is a pure formal issue: if we defined $F_{Y|X}(y|x)$ as the probability $p(Y \le y \,|\, X \le x)$, then there would be no problem to state a product rule and Bayes' theorem with conditional probability functions. However, this is not normally done due to the danger of interpreting $F_{Y|X}(y|x)$ as $p(Y \le y \,|\, X = x)$, which is not defined if X is a true continuous random variable since $p(X = x) = 0$. As for the sum rule or marginalisation, it is not possible to write that with cumulative probabilites because, unlike the events $X = x$, the events $X \le x$ are *not* mutually exclusive. They intersect, and hence it isn't possible to make use of Kolmogorov's Axiom 3 to state a sum rule.

You will learn about a product and a sum and rule for continuous variables in the next section.

3.2 The Probability Density Function

For a discrete random variable X taking values over a set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ such that $x_1 \le x_2 \le \ldots \le x_n$, we have

$$P_X(x_k) = F_X(x_k) - F_X(x_{k-1}) \text{ for } k = 2, 3, \ldots, n. \tag{3.6}$$

Can we apply a similar approach to continuous random variables in order to obtain something along the lines of a probability distribution from the well defined cumulative probability function? If we try to apply the expression above to a continuous random variable for an infinitesimal interval, we obtain in general

$$\lim_{\Delta x \to 0} (F_X(x + \Delta x) - F_X(x)) = 0 \tag{3.7}$$

which is in line with our claim for truly continuous random variables that the probability of single outcomes or values is zero. If instead we divide the expression by the size of the interval, we get the definition of a derivative, which is known as the *probability density function* (PDF):

$$f_X(x) = \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = \frac{dF_X}{dx} = F_X'(x). \tag{3.8}$$

The cumulative probability function for continuous random variables is often called the *cumulative density function* (CDF) by analogy with the PDF.

The PDF is also often called the *continuous probability distribution* and I will do my best to avoid following that trend but be aware that you might find it in past exams and in your data book. The following word of caution may explain my concerns.

A short digression into dimensional analysis: the temptation is great to consider a probability density function as a sort of probability of a "small interval". The reason for our long-winded introduction of this section, rediscovering derivatives from first principles rather than just plainly defining the PDF as a derivative, is to give you a better understanding of the subtle difference between the PDF and the probability of a small interval. The difference is in the *division by the length* of the small interval.

Say for example that X is the speed of a car in the previous examples. The unit of speed that we used is the mile per hour (mph). Probability is a unitless quantity. The probability $p(a \leq X \leq b) = F_X(b) - F_X(a)$ is just a number without a unit. The cumlative probability function $F_X(x)$ and the probability distribution $P_X(x)$ of a discrete random variable are all probabilities and hence unitless as well. The probability density function (PDF) however is divided by Δx. In our example, Δx has the unit *mph*. Hence, the unit of $f_X(x)$ is hpm=mph^{-1}. In other words, whatever the unit u of X, the PDF always has the unit u^{-1} and hence it is clearly not a probability, which must always be unitless.

For continuous random variables, think of $f_X(x)dx$ as the infinitesimal equivalent of a probability, but never of $f_X(x)$ itself as this can lead to serious confusion.
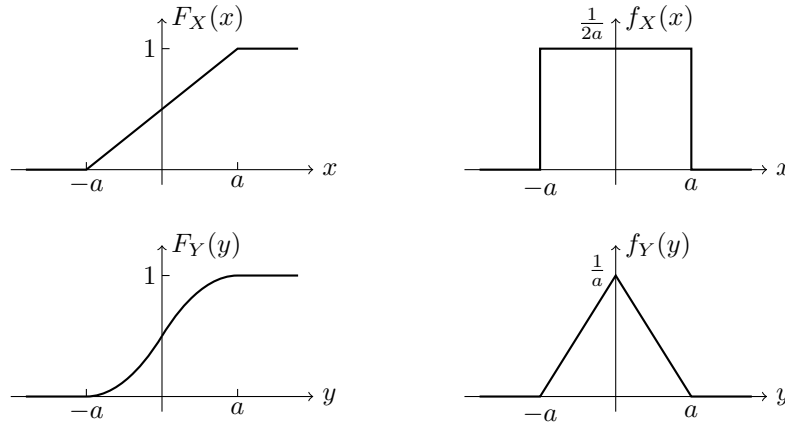

Figure 3.1: Cumulative probability function (left) and probability density function (right) for two random variables, X uniform and Y triangular

The probability density function is a very useful quantity that often takes precedence over the cumulative probability function in people's perception because it acts like a probability distribution in helping us visualise the probabilistic behaviour of random variables. Figure 3.1 shows the cumulative probability distribution and the probability density function for two random variables. When using continuous variables, it is worth remembering the following two rules

1. $f_X(x)dx$ is the limiting "probability" that X takes values in the interval $[x, x + dx]$, *not* $f_X(x)$; and

2. when matters get difficult and challenge your intuition, always revert back to the well defined cumulative probability function, which has a "physical" meaning the probability of an event.

A few fundamental properties that follow immediately from the definition of the probability density function are

$$f_X(x) \geq 0 \text{ for all } x \tag{3.9}$$

following from the fact that $F_X(x)$ is non-decreasing. Also,

$$p(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx, \tag{3.10}$$

i.e., the probability of an interval is the integral of the PDF over the interval, and hence

$$\int_{-\infty}^{\infty} f_X(x)dx = \lim_{x \to \infty} F_X(x) - \lim_{x \to -\infty} F_X(x) = 1 - 0 = 1, \qquad (3.11)$$

i.e., the probability density function integrates to 1. One would be tempted to conclude that the density behaves like a probability measure since it's non-negative and integrates to 1, but note that the density can be larger than 1, as illustrated for example by considering the uniform density in the first example of Figure 3.1 for $a = 1/4$.

Joint probability density functions are obtained from the cumulative density function through a multiple differentiation, e.g.,

$$f_{XY}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{XY}(x,y) \qquad (3.12)$$

and independent random variables satisfy

$$f_{XY}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_X(x)F_Y(y) = f_X(x)f_Y(y). \qquad (3.13)$$

In fact, this condition implies independence because there is no integration constant when going from the PDF to the cumulative probability function, since every cumulative probability function must begin at 0 and end at 1. Finally, we will state without proof[1] that the "sum rule" becomes an "integral rule" or, in other words, that marginalisation applies to probability density functions

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy. \qquad (3.14)$$

Conditional probability density functions can be defined as

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}, \qquad (3.15)$$

yielding the product rule

$$f_{XY}(x,y) = f_{Y \mid X}(y \mid x)f_X(x) \qquad (3.16)$$

and Bayes' theorem applies

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y \mid x')f_X(x')dx'}. \qquad (3.17)$$

So in short, despite all our warnings about not mistaking densities for probabilities, in fact the PDF (Probability Density Function) acts very much as a PMF (Probability Mass Function) for all practical purposes since it fulfils a sum rule and a product rule.

The probability density function can also be used to compute expectations

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx \qquad (3.18)$$

for any function $f(.)$. In particular, we will be interested in the mean (also called "first moment")

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx, \qquad (3.19)$$

the "second moment"

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)dx \qquad (3.20)$$

---

[1]The proof isn't hard as such but is notationally tedious and of little educational value.

and the n-th moment

$$E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx. \tag{3.21}$$

The variance remains $\text{Var}(X) = E[X^2] - E[X]^2$. Continuous random variables do not have an entropy in the sense defined for discrete random variables. They do have a quantity known as the "differential entropy" but we will not study it in this course.

We are now ready to study a few important PDFs and CDFs.

3.3   The Exponential Density

The exponential density can be derived from the Poisson distribution we studied in the previous lecture. Consider the example we gave of a 10 Gb/s Ethernet router handling packets at an average rate of 5×10^6 packets per second. Packets arrive at the router at irregular intervals. Assume that the time of arrivals and intervals between packets are independent. This is a fairly reasonable model for packet arrival times in a router, but would not be a good model for London buses for example who famously tend to arrive in convoys[2]. What is the (continuous) probability density function of the time interval X between consecutive packets? This is known as the *exponential density* for the continuous random variable X,

$$X \sim \text{Exp}(\lambda) \tag{3.22}$$

In order to derive this density, consider Y_t the Poisson distributed random variable giving the number of arrivals in a given time interval of length t. Let λ be the rate of packet arrivals per time unit, so that λt is the rate for the interval of length t. In our example, $\lambda = 5 \times 10^5$. Remember from Section 2.4 that

$$P_{Y_t}(k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{3.23}$$

Note that the λ in our original expression for the Poisson distribution has been replaced by λt in this expression simply because we are considering an interval of length t and λ is the rate of arrival per *unit* of time, whereas in Section 2.4 it was the rate of arrival for the time interval considered.

We express the cumulative probability function of the inter-arrrival time as

$$F_X(t) = p(X \le t) = 1 - p(X > t). \tag{3.24}$$

X can only be larger than t if no arrivals occur in the interval, implying

$$p(X > t) = P_{Y_t}(0) = e^{-\lambda t} \tag{3.25}$$

and hence

$$F_X(t) = 1 - e^{-\lambda t}. \tag{3.26}$$

We now derive[3]the probability density function of X

$$f_X(t) = \frac{d}{dt} F_X(t) = \lambda e^{-\lambda t}, \text{ where } t \ge 0, \lambda \ge 0. \tag{3.27}$$

As mentioned above, the exponential density is used to model the time intervals in a Poisson process with independent arrival times. Figure 3.2 shows the probability density functions for exponentially distributed random variables for various values of λ.

---

[2]I only mention this because bus arrival times are often used in textbooks to illustrate the exponential density. My personal experience with London buses leads me to believe that this a poor example.

[3]For those of you who are comfortable with infinetisimal calculus, there is an alternative derivation of the exponential density that gets the PDF directly and provides an alternative insight. We know that $f_X(t)dt$ is the limiting probability that the next packet in a Poisson process will arrive at time t. This can be seen as the probability of the event {no packet will arrive in the interval $[0, t)$ AND a packet will arrive in the interval $[t, t+dt)$} in the limit. We can write this as follows

$$f_X(t)dt = p(Y_t = 0 \cap Y_{dt} = 1) = P_{Y_t}(0)P_{Y_{dt}}(1) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} \frac{(\lambda dt)^1}{1!} e^{-\lambda dt} = \lambda e^{-\lambda t} e^{-\lambda dt} dt. \tag{3.28}$$

Since $\lim_{dt \to 0} e^{-\lambda dt} = 1$, we obtain the exponential density.
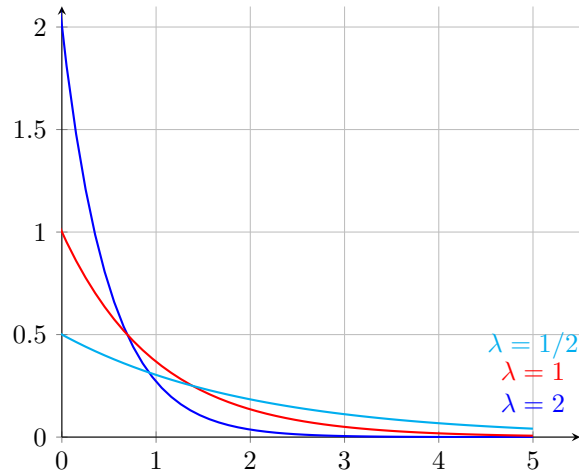
Figure 3.2: Exponential probability density functions for $\lambda = 2, 1$ and $1/2$

The mean of an exponential density function can be calculated easily using integration by parts

$$E[X] = \int_0^\infty t f_X(t) dt = -te^{-\lambda t}\big|_0^\infty + \int_0^\infty e^{-\lambda t} dt \tag{3.29}$$

$$= -\frac{1}{\lambda} e^{-\lambda t}\Big|_0^\infty = \frac{1}{\lambda} \tag{3.30}$$

and the second moment

$$E[X^2] = \int_0^\infty t^2 f_X(t) dt = -t^2 e^{-\lambda t}\big|_0^\infty + 2\int_0^\infty t e^{-\lambda t} dt \tag{3.31}$$

$$= \frac{2E[X]}{\lambda} = \frac{2}{\lambda^2} \tag{3.32}$$

and hence the variance

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \tag{3.33}$$

The relationship between the variance and the mean for Poisson and exponential random variables can be exploited as a tool to verify if an arrival process, say for example the packet arrival in our 10 Gbit/s Ethernet router, is a Poisson process:

- verify whether mean and variance of the *number of arrivals per unit interval* obeys $E[Y] \simeq \mathrm{Var}(Y)$ as should be the case for a Poisson distributed random variable, and

- verify whether mean and variance of the *time intervals between arrivals* obeys $E[X]^2 \simeq \mathrm{Var}(X)$ as should be the case for exponential distributed continuous random variables.

The Poisson waiting paradox: the derivation of the exponential density we gave does not rely anywhere on the last occurrence that happened. Although we described the exponential density as a model of the time intervals *between* occurrences, in fact the memoryless nature of the process means that the *time until the next occurrence* is exponentially distributed with parameter λ, no matter when we start counting time. Our derivation only required there to be no occurrences in the interval $[0, t)$ and one occurrence in the interval $[t, t + dt)$. It did not require there to have been an occurrence at or near time 0. The average waiting time until the next occurrence is $1/\lambda$.

> Now consider our derivation in reverse, asking the question of how much *time has elapsed since the last occurrence* rather than how much time remains until the next occurrence. Since our derivation was symmetric in time, the answer would be the same: the time elapsed since the last occurrence follows an exponential density with parameter λ. The average time elapsed since the last occurrence is $1/\lambda$.
>
> This is known as the waiting time paradox:
>
> - the average time between occurrences is $1/\lambda$,
>
> - the average time until the next occurrence is $1/\lambda$,
>
> - the average time since the last occurrence is $1/\lambda$,
>
> - if you pick any reference time, it would appear that the average time between the last occurrence and the next occurrence is $2 \times 1/\lambda = 2/\lambda$.
>
> Be sure to remember this paradox next time you wait for a bus that's meant to arrive every 12 minutes on average and has not arrived for the past 24 minutes...

3.4 The Gaussian Density

The Gaussian density, named after the German mathematician Carl Friedrich Gauss (1777-1855), is of central importance in the study of probability[4] It arises whenever a quantity is the sum of many smaller effects. It is a good model (sometimes provably accurate) for many quantities, such as

- the velocities of particles in an ideal gas;

- noise added to a signal by a variety of parasitic effects in a communications receiver;

- measurement noise;

- quantities in biology (e.g. height or weight of humans)

- quantities in finance (price indices),

- quantities in social studies (IQ test results),

- marks for the IB Paper 7 (Mathematical Methods) tripos examination.

We will understand why the Gaussian density is such a good model in Section 4.4.

If Y follows a Gaussian density with parameters μ and σ^2

$$Y \sim \mathrm{N}(\mu, \sigma^2) \tag{3.34}$$

its probability density function (PDF) is

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}. \tag{3.35}$$

---

[4]The Gaussian density is also called the *normal* density. The name was introduced by the English mathematician Karl Pearson who believed that other mathematicians than Gauss deserved to share the credit for its invention. He later regretted the name, because it may lead people to believe that other densities are abnormal. Gauss's first use of the density is undisputed nowadays and all the mathematicians involved in Pearson's attempted historical re-engineering have had important mathematical concepts named after them (e.g., the Laplace transform). Sadly, the name "normal" has stuck, and it is important for you to be aware of this alternative name as you may encouter it frequently.
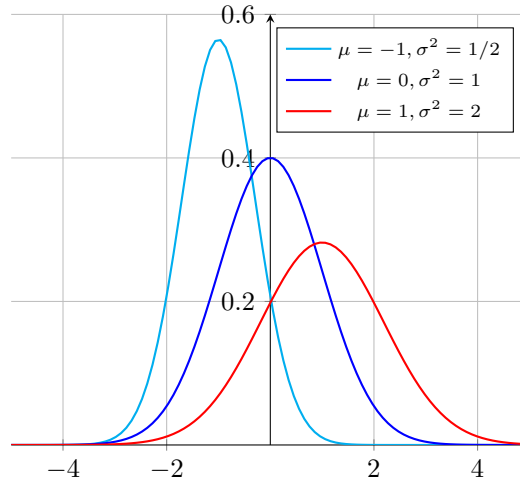
Figure 3.3: Gaussian probability density functions for various parameters

The mean is $E[Y] = \mu$ and the variance is $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = \sigma^2$. Figure 3.3 shows Gaussian probability density functions for a few choices of parameters.

A zero mean, unit variance Gaussian random variable

$$X \sim \mathrm{N}(0, 1) \tag{3.36}$$

is said to follow a *standard Gaussian* density. We will show in the next chapter that, if X is standard Gaussian, then $Y = \sigma X + \mu$ is Gaussian with mean μ and variance σ^2. Conversely, for any Gaussian random variable $Y \sim \mathrm{N}(\mu, \sigma^2)$, $X = (Y - \mu)/\sigma$ is standard Gaussian.

The cumulative probability function or CDF of a standard Gaussian, often denoted $\Phi(x)$,

$$F_X(x) = \Phi(x) = p(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x'^2}{2}} dx' \tag{3.37}$$

can not be expressed in closed form and is hence often tabulated or implemented as a numerical integral. It is tabulated on Page 29 of your mathematics data book. Most programming and computation environments such as Python, MATLAB, etc. have library functions for the so-called "error function" (e.g. `math.erf(x)` and `scipy.special.erf(x)` in Python, `erf(x)` in MATLAB) which is the probability that a Gaussian random variable with mean 0 and variance $1/2$ (!!!^{5}) lies within the interval $[-x, x]$, giving

$$\mathtt{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-x'^2} dx', \tag{3.38}$$

so the standard Gaussian cumulative can be computed from this function as

$$F_X(x) = \Phi(x) = \frac{1}{2} + \frac{1}{2} \mathtt{erf}(x/\sqrt{2}). \tag{3.39}$$

Another often tabulated or implemented function is called the Q function, defined as

$$Q(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x'^2}{2}} dx' = 1 - \Phi(x), \tag{3.40}$$

i.e., the probability that a standard Gaussian random variable exceeds the value x. The Gaussian PDF is an even function $f_X(x) = f_X(-x)$ and hence the CDF satisfies

$$F_X(x) = \Phi(x) = 1 - \Phi(-x) = 1 - F_X(-x). \tag{3.41}$$

---

[5]Why?? No idea... Computer scientists can be strange sometimes.

A few useful values to remember are, for any $Y \sim N(\mu, \sigma^2)$ or $X \sim N(0, 1)$,

$$\begin{cases} p(\mu - \sigma \leq Y \leq \mu + \sigma) = F_X(1) - F_X(-1) = 2\Phi(1) - 1 \simeq 2/3, \\ p(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) = F_X(2) - F_X(-2) = 2\Phi(2) - 1 \simeq 0.95 \end{cases} \tag{3.42}$$

so a 2σ deviation from the mean is often called the "95% confidence interval".

3.5 The Beta Density

The Beta density is a continuous probability density function whose range is limited to a finite interval, normally $[0, 1]$ but its range can be shifted to other intervals if required. It is used to model parameters that have a finite range in various disciplines (biology, physics, medicine, operations research.) More importantly for us, it can be used to model the parameter p of a Bernoulli distributed random variable.

Example 3.3. As captain of your football team, you have the difficult task of picking a player to shoot the penalty your team has just been awarded. You carefully consider the binary sequence of n past penalty try-outs for player A ('1' for score, '0' for fail) and would like an estimate of the probability that player A will score the next penalty.

The sequence of past penalty try-outs can be considered as a sequence of independent Bernoulli distributed random variables with parameter p (we ignore the "learning effect" or "anxiety effect" that may introduce some dependency in the data) and hence the number of scores ('1's) Y follows a binomial distribution $Y \sim B(n, p)$.

The unknown quantity in this case is the parameter of the Bernoulli distribution. It is a continuous random variable limited to a finite range $[0, 1]$. We shall call it π to distinguish it from a fixed parameter p, so we can write for example $F_\pi(p)$ for the probability $p(\pi \leq p)$. Hence the number of scores is now $Y \sim B(n, \pi)$ where we see for the first time a random variable π as a parameter for the distribution of another random variable Y.

Most people if given a count of Y scores in n trials would be tempted to conclude that $p = Y/n$. There are obvious problems with this approach:

- It yields a fixed value for the parameter without accounting for our uncertainty. Probability theory is the mathematical theory of uncertainty and hence one should expect a probability density for π rather than a fixed value p.

- If $n = 0$ (no data has been seen yet), the estimator above is $p = 0/0$ and hence not determined.

- If $Y = 1$ and $n = 1$ (only one 1 seen so far and nothing else) the estimator above yields $p = 1$ or a probability of 1 to see another 1. This does not seem like a reasonable estimate given that we've seen only one data point so far.

So how can we correctly make statements about π given an observed count of Y scores in n trials? We will return to this question after introducing the Beta density.

We will say that a random variable π follows a Beta density with parameters α, β, i.e.,

$$\pi \sim \text{Beta}(\alpha, \beta) \tag{3.43}$$

when it has the probability density function

$$f_\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1} (1 - p)^{\beta - 1}, \text{ with } 0 \leq p \leq 1 \text{ and } \alpha, \beta \geq 0. \tag{3.44}$$
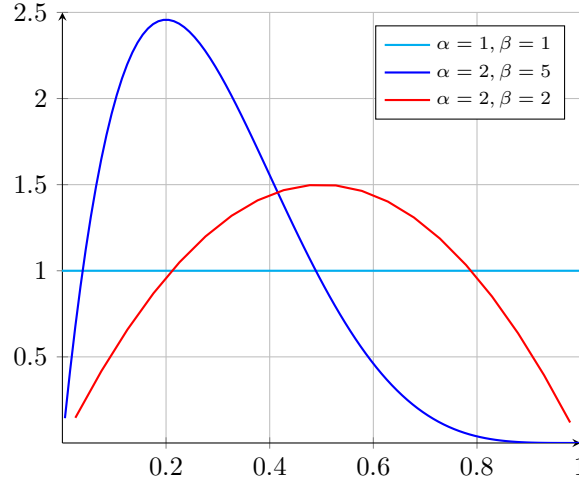
Figure 3.4: Beta probability density functions for various parameters

The Gamma function $\Gamma(x)$ is an extension of the factorial function to complex and real numbers and is defined via the integral

$$\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy \tag{3.45}$$

for which no closed-form expression exists. For integer arguments n it is equivalent to the factorial

$$\Gamma(n) = (n-1)! \tag{3.46}$$

We will in general consider Beta densities with integer parameters, so the Beta probability density function in this case can be expressed as

$$f_\pi(p) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}p^{\alpha-1}(1-p)^{\beta-1} = (\alpha+\beta-1)\binom{\alpha+\beta-2}{\alpha-1}p^{\alpha-1}(1-p)^{\beta-1}. \tag{3.47}$$

Figure 3.4 shows Beta densities for various parameters. Note that Beta(1,1) is the *uniform* probability density function over the interval $[0,1]$.

The expected value of the Beta density is

$$E[\pi] = \frac{\alpha}{\alpha+\beta}. \tag{3.48}$$

Its variance can also be expressed in closed form but we won't give it here. There is no closed form for its cumulative probability function.

Example 3.4. We return to our football example. We begin by selecting a *prior* for the unknown paramater π determining the probability that player A scores a goal at each penalty attempted. This prior is a probability density function for π given that no data has been collected ($n = 0$). At this point, it would be acceptable to chose any prior that reflects the captain's experience, belief, gut feeling, or intuition about player A's abilities. Let us assume that the captain is inexperienced or would prefer not to exert any prejudice on player A's ability to score a penalty before seeing her in action. The captain assumes a uniform probability density function $f_\pi(p) = 1$ for $p \in [0,1]$, or $\pi \sim \text{Beta}(1,1)$.

Having observed Y scores in n trials for player A, the captain calculates the conditional probability[6]

$$P_{Y|\pi}(k\,|\,p) = \binom{n}{k}p^k(1-p)^{n-k} \tag{3.49}$$

which is a binomial distribution, and is now in a position to use Bayes' theorem

$$f_{\pi|Y}(p\,|\,k) = \frac{P_{Y|\pi}(k\,|\,p)f_\pi(p)}{P_Y(k)}.\tag{3.50}$$

For a uniform prior, $f_\pi(p) = 1$, it is easy to guess (or verify by a long process of integration by part that you can do as an exercise if you wish) that $P_Y(k) = 1/(n+1)$ for $k = 0, 1, \ldots, n$. Hence we get

$$f_{\pi|Y}(p\,|\,k) = (n+1)\binom{n}{k}p^k(1-p)^{n-k}\tag{3.51}$$

which is the Beta density Beta$(k+1, n-k+1)$. The expected value of π given the data is

$$E[\pi\,|\,Y = k] = \frac{k+1}{n+2}.\tag{3.52}$$

In other words, if you've seen player A score four goals in five trials, you should estimate the probability density function of the Bernoulli coefficient to be $\pi \sim \text{Beta}(5, 2)$. If you want the expectation of the Bernoulli coefficient, it would be 5/7, which is what most people would guess was the probability of scoring five goals (one more than scored) in seven penalty trials (two more than attempted.) This expected value estimator with a uniform prior is called the *Laplace* estimator after the French mathematician Pierre-Simon Laplace (1749-1827) and has the advantage that it gives reasonable answers for the disturbing cases mentioned in the previous examples box:

- when you haven't seen any trials, the expected value of the Bernoulli coefficient is $(0+1)/(0+2) = 1/2$;

- when you've seen only one trial and it was a success, the expected value of the Bernoulli coefficient is $(1+1)/(1+2) = 2/3$.

The method above is an example of the *Bayesian approach* to probablities and statistics. The football team captain could improve the estimate by using a prior more representative of their intuition, belief or experience, and hence become more able to take decisions effectively when new players join the team.

3.6   Summary

Fundamentals:

- Cumulative Probability Distribution (CDF): $F_X(x) = p(X \le x)$

- Properties: $F_X(x) \ge 0$, non-decreasing function of x from 0 at $-\infty$ to 1 at $+\infty$

- $p(a \le X \le b) = F_X(b) - F_X(a)$

- Joint cumulative: $F_{XY}(x, y) = p(X \le x \cap Y \le y)$

- Independence: $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all (x, y).

---

[6]We are stretching our definitions here to conditional probability of discrete random variables conditioned on continuous random variables with a word of caution that such constructs work for most well behaved engineering examples but would arise serious suspicion among mathematicians.

Probability density:

- Probability Density Function (PDF): $f_X(x) = \frac{dF_X}{dx} = F'_X(x)$

- Properties: $f_X(x) \geq 0$ for all x, $p(a \leq X \leq b) = \int_a^b f_X(x)dx$, and $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

- Joint PDF: $f_{XY}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{XY}(x,y)$

- Independence: $f_{XY}(x,y) = f_X(x)f_Y(y)$ for all (x,y) (necessary and sufficient condition)

- Sum rule: $f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy$

- Product rule: $f_{XY}(x,y) = f_{Y\,|\,X}(y\,|\,x)f_X(x)$ (or definition of conditional PDF)

- Bayes' theorem: $f_{X|Y}(x\,|\,y) = \frac{f_{Y|X}(y\,|\,x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y\,|\,x')f_X(x')dx'}$

- Expectation: $E[f(X)] = \int_{-\infty}^{\infty} f(x)f_X(x)dx$, n-th moment $E[X^n]$.

Exponential Density:

- $X \sim \mathrm{Exp}(\lambda)$

- $F_X(t) = 1 - e^{-\lambda t}$, $f_X(t) = \lambda e^{-\lambda t}$, where $t \geq 0$

- $E[X] = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$

- The exponential density models inter-arrival times in a Poisson process

Gaussian (or normal) Density:

- $Y \sim \mathrm{N}(\mu, \sigma^2)$, $E[Y] = \mu$ and $\mathrm{Var}(Y) = \sigma^2$

- $f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$.

- Standard Gaussian: $X \sim \mathrm{N}(0,1)$

- $F_X(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-\frac{x'^2}{2}}\,dx' = 1 - Q(x) = \frac{1}{2} + \frac{1}{2}\mathtt{erf}(x/\sqrt{2})$

Beta density:

- $\pi \sim \mathrm{Beta}(\alpha, \beta)$ (π is the name of the random variable, not the number $3.1415\ldots$)

- $f_\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}$, with $0 \leq p \leq 1$ and $\alpha, \beta \geq 0$

- $\Gamma(x) = \int_0^{\infty} y^{x-1}e^{-y}dy$ and for integers n, $\Gamma(n) = (n-1)!$

- $E[\pi] = \frac{\alpha}{\alpha+\beta}$

- The Beta distribution models the density of a Bernoulli parameter after having observed $\alpha - 1$ ones and $\beta - 1$ zeros, assuming a uniform prior

3.7 Problems

Having completed this chapter, you should be able to complete Examples Paper 5 and start working on Examples Paper 6.

Chapter 4

Manipulating and Combining Distributions

Having learned about discrete and continuous random variables, we are now ready to start manipulating them. This is where probability theory really gets interesting.

4.1 Functions of random variables

The first operation we will consider is taking functions of random variables. Let X be a random variable and let $Y = g(X)$ for some function f mapping the domain $\mathcal{X}$ of X to the domain $\mathcal{Y}$ of Y. The question we will be interested in is to determine the distribution of Y when the distribution of X is known. Note that, since Y is a function of X, the two random variables automatically share a common sample space and have common randomness.

4.1.1 Functions of Discrete Random Variables

We first consider discrete random variables X. Our aim is to determine the probability $P_Y(y) = p(Y = y)$ for all values $y \in \mathcal{Y}$. There are two cases of interest:

1. if there is only one value $x \in \mathcal{X}$ such that $g(x) = y$, then $P_Y(y) = p(Y = y) = p(X = x(y)) = P_X(x(y))$ where $x(y) = g^{-1}(y)$ is the only x such that $g(x) = y$.

2. otherwise, let $\mathcal{X}_y$ be the subset of $\mathcal{X}$ containing all values x such that $g(x) = y$, then

$$P_Y(y) = p(Y = y) = \sum_{x \in \mathcal{X}_y} p(X = x) = \sum_{x \in \mathcal{X}_y} P_X(x) \tag{4.1}$$

Of course, the expression just stated for the second case in fact describes both cases: the set $\mathcal{X}_y$ simply contains only one element in the first case.

Example 4.1. Let X be the value of a random dice throw (equal to the outcome). The set of values is $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.

- Let $Y = f_1(X) = X^2$. In this case, all values of $\mathcal{X}$ are mapped to distinct values in $\mathcal{Y}$ and hence $P_Y(y) = P_Y(x^2) = P_X(x) = P_X(\sqrt{y})$ for all $y \in \mathcal{Y}$.

- Let $Z = f_2(X)$ be 0 if X is even and 1 if X is odd. In this case,

$$P_Z(0) = P_X(2) + P_X(4) + P_X(6) = \frac{3}{6} = \frac{1}{2}. \tag{4.2}$$

Figure 4.1 illustrates the two example functions and how they map $\mathcal{X}$ to $\mathcal{Y}$ and $\mathcal{Z}$. In this type of graphical representation of functions, when there are "collisions" (values mapping to the same image), the probabilities accumulate.
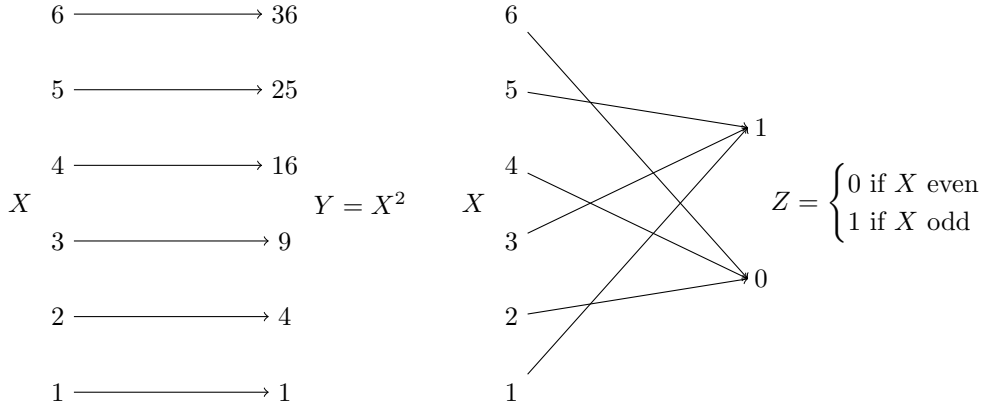

Figure 4.1: Y has the same probabilities as X, whereas for Z the probabilities sum to $(1/2, 1/2)$

.

Although the theory for discrete random variables is very simple, it can already lead to interesting problems, as the following examples show.

Example 4.2. Let X have a binomial distribution $X \sim B(99, p)$. Let Y be a summary of X where the 100 possible values of X are grouped into 10 bins, i.e., $Y = 1$ if $0 \leq X \leq 9$, $Y = 2$ if $10 \leq X \leq 19$, etc. Picture for example Y representing a 10-bin "histogram" for X, or giving a dart player a mark Y between 1 to 10 according to the number of times they hit bull's eye in 100 shots. What is the probability distribution of Y?

The answer

$$P_Y(y) = \sum_{m=10y}^{10y+9} \binom{99}{m} p^m (1-p)^{99-m} \tag{4.3}$$

cannot be brought into a form that can be calculated easily, other than evaluating the sums using an appoximation for the binomial coefficients.

Example 4.3. Let X have a geometric distribution $P_X(k) = p(1-p)^{k-1}$ for $k = 1, 2, \ldots$. Let $Y = (X - 1) \bmod m$, i.e., the value of $X - 1$, *modulo* some given number m. What is the probability distribution of Y?

X and Y could occur for example in the following scenario. Consider a ripple or synchronous counter that you've studied in 1A Paper 3 Digital Circuits last year, that counts from 0 to $m - 1$ and then starts again at zero. Say the clock used for the counter has a probability p of failure at each count. The number of clock cycles X until failure follows a geometric distribution. We are interested in the probability distribution of the clock state Y when the clock fails, which is $(X - 1) \bmod m$.

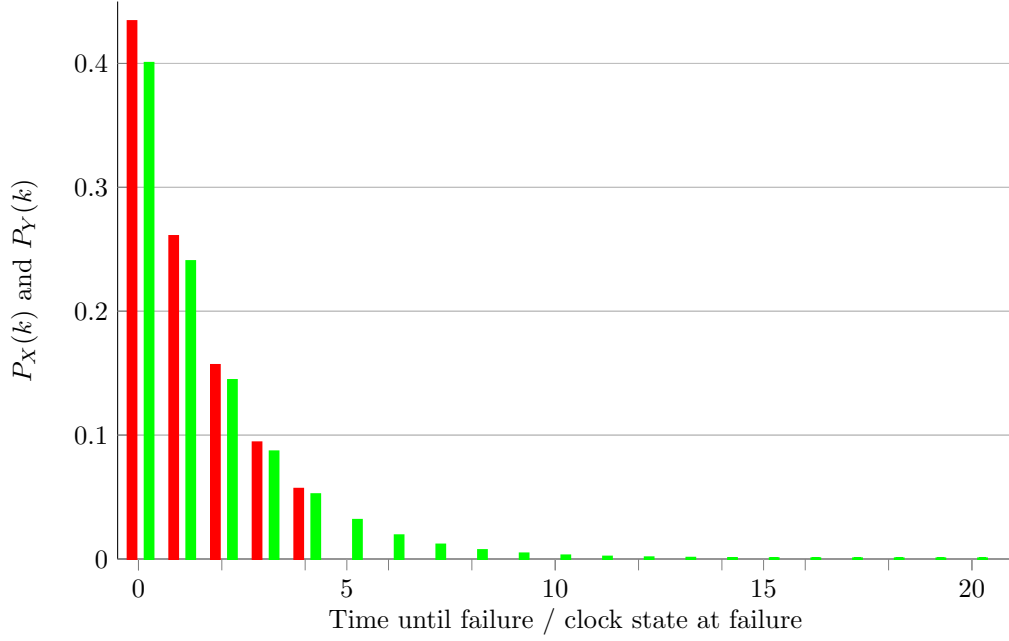Observe that Y will take on the value 0 if X is 1, $m+1$, $2m+1$, $3m+1$, etc. The resulting

Figure 4.2: Distribution of the time X until failure (green) and clock state Y at failure (red) when clock operates modulo $m = 5$.

distribution does have a closed form obtained via the sum of geometric sequences

$$P_Y(y) = \sum_{r=0}^{\infty} p(1-p)^{rm+y} = \frac{p(1-p)^y}{1-(1-p)^m} \tag{4.4}$$

for $y = 0, 1, \ldots, m - 1$, which is a re-scaled geometric distribution truncated to the positions 0 to $m - 1$. This is illustrated in Figure 4.2.

4.1.2 Functions of Continuous Random Variables

Taking functions of continuous random variables is a thornier matter. The two cases we considered in the previous section also apply, but even the "simple" case of a function that maps only one value x to an image $g(x) = y$ is now interesting and requires some theory.

Before proceeding, we will make two assumptions about the function $g(.)$ that will simplify our initial treatment. We will later relax the second assumption somewhat to make the theory more general:

- $g(.)$ is "smooth", i.e., continuous and differentiable over the domain $\mathcal{X}$ of X.

- The derivative $g'(.)$ is positive over $\mathcal{X}$ except at most in single points, i.e., the function is non-decreasing and does not remain constant over any intervals[1].

The second condition implies that the function $g(.)$ is "one-to-one" in the sense that for every image y there is only one $x \in \mathcal{X}$ such that $g(x) = y$. Valid functions that we may consider at this stage are for example

---

[1]If $g(.)$ remains constant over intervals $Y = g(X)$ is in general a mixed continuous-discrete random variable for which the theory can be more difficult to establish, although such random variables are by no means inaccessible to an engineer with a bit of imagination.
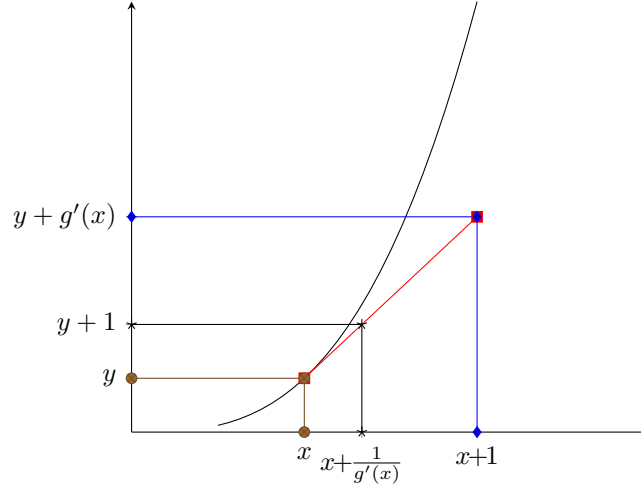
Figure 4.3: The derivative of g and g^{-1} seen as coordinates of the tangent vector

- $g(x) = \alpha x$ for $\alpha > 0$

- $g(x) = x^2$ for $x \geq 0$

- $g(x) = e^x$

Since the function does not decrease anywhere in $\mathcal{X}$, we conclude that the cumulative probability function of the random variable $Y = g(X)$ is

$$F_Y(y) = p(Y \leq y) = p(X \leq x(y)) = F_X(x(y)) \tag{4.5}$$

where $x(y) = g^{-1}(y)$ is the only x such that $g(x) = y$. We obtain the probability density function by differentiating, using the chain rule,

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))}. \tag{4.6}$$

The inverse of the derivative on the right is the derivative of $g^{-1}(y)$, which you can either see as a consequence of

$$\frac{dF_Y}{dy} = \frac{dF_X(x)}{dy} = \frac{dF_X}{dx}\frac{dx}{dy} = \frac{dF_X}{dx}\left(\frac{dy}{dx}\right)^{-1} = \frac{dF_X}{dx}\frac{1}{g'(x)} \tag{4.7}$$

or by picturing the tangent vector at (x, y) that can be expressed either as $\mathbf{t_1} = (1, g'(x))$, or as $\mathbf{t_2} = \mathbf{t_1}/g'(x) = (1/g'(x), 1)$. The first expression $\mathbf{t_1}$ is the one that matters in the coordinate system (x, y) relevant to the function g, whereas the second expression $\mathbf{t_2}$ is the one that matters in the reverse coordinate system (y, x) relevant to the function g^{-1} that maps y to x. This is illustrated in Figure 4.3.

Example 4.4. Let X be uniformly distributed over $\mathcal{X} = [0, 1]$, i.e.,

$$f_X(x) = 1, \quad x \in [0, 1]. \tag{4.8}$$

By integration, we have
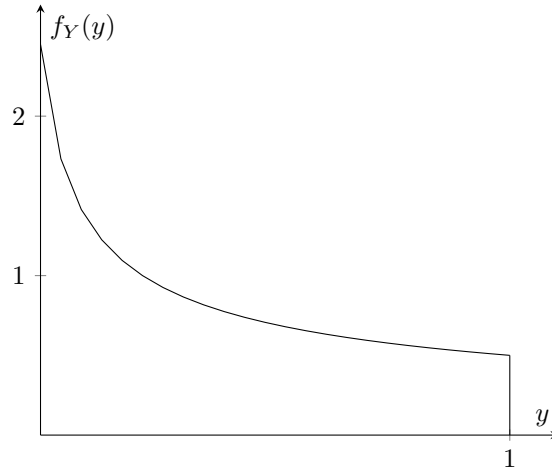
$$F_X(x) = x, \quad x \in [0, 1]. \tag{4.9}$$

Figure 4.4: The density $f_Y(y)$ for $Y = X^2$ and X uniform between 0 and 1.

Let $Y = g(X) = X^2$. We have $g'(x) = 2x \geq 0$ for $x \in [0, 1]$, and hence

$$F_Y(y) = F_X\left(g^{-1}(y)\right) = \sqrt{y} \tag{4.10}$$

and

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \frac{1}{g'\left(g^{-1}(y)\right)} = \frac{1}{2\sqrt{y}}. \tag{4.11}$$

The density is illustrated in Figure 4.4.

Now that we have obtained expressions for $F_Y(y)$ and for $f_Y(y)$ for functions $Y = g(X)$ such that $g'(x) \geq 0$ for all $x \in \mathcal{X}$, let us consider the case $g'(x) \leq 0$ for all $x \in \mathcal{X}$ (where again, we assume that $g'(x)$ is zero at most in single points so that the function is not constant on any interval). These are one-to-one functions that do not increase anywhere and hence our first step would have differed from the previous case in that

$$F_Y(y) = p(Y \leq y) = p(X{\geq}x) = 1 - F_X(x) \tag{4.12}$$

where $x = g^{-1}(y)$. The differentiation to obtain the probability density function would have an added minus sign, which is just as well since $dx/dy \leq 0$ so our original expression without the minus sign would have resulted in a negative density.

We can combine the two density expressions for $g'(x) \geq 0$ or $g'(x) \leq 0$ everywhere in $x \in \mathcal{X}$ into a single expression

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = f_X\left(g^{-1}(y)\right) \frac{1}{\left| g'\left(g^{-1}(y)\right) \right|} \tag{4.13}$$

that is valid for all functions $Y = g(X)$ that are either increasing or decreasing everywhere in $\mathcal{X}$ or zero in single points at most. It is possible to extend this expression to functions that are increasing and decreasing in well-defined intervals, giving you the ability to tackle even quite complex problems such as the one in the following example.

Example 4.5. The following example is fairly complex and involves a function that is neither decreasing everywhere nor increasing everywhere. Consider a microprocessor that resets
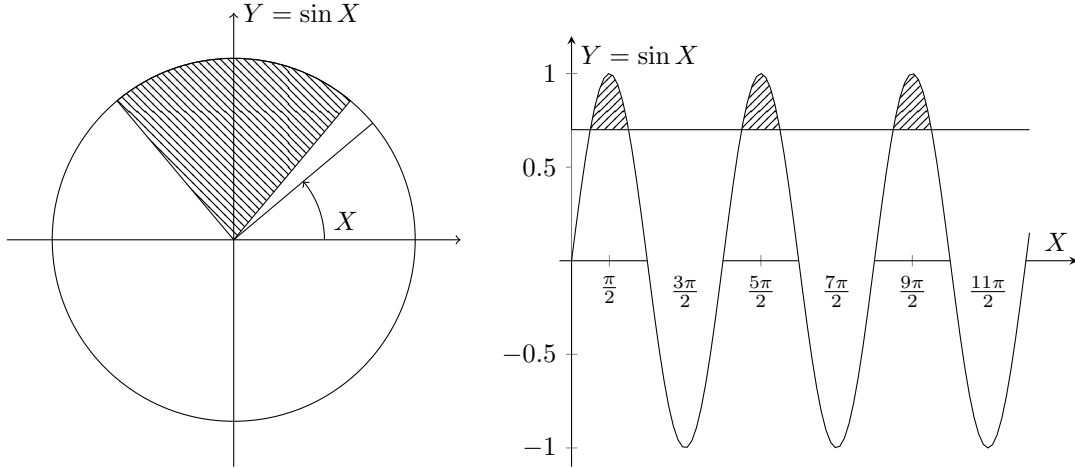
Figure 4.5: For $Y = \sin X$, the union of the non-shaded segments is the event $\{Y \le y\}$

(syncs) an oscillator every time an interrupt occurs. The time X until the next interrupt is an exponential random variable. We wish to know the probability density function of the oscillator state $\sin X$ when the next interrupt occurs.

In other words, X is an exponential random variable with parameter λ, i.e.,

$$f_X(x) = \lambda e^{-\lambda x} \text{ and } F_X(x) = 1 - e^{-\lambda x} \tag{4.14}$$

and $Y = \sin X$. Can we determine $F_Y(y)$ and $f_Y(y)$? Since $\sin(x)$ is neither increasing nor decreasing everywhere, we cannot simply apply Equation 4.13. It is easiest to tackle this problem via the cumulative probability function.

The cumulative function $F_Y(y) = p(Y \le y)$ is the probability of the event $\{Y \le y\}$ that $\sin X$ takes on values less than or equal to y. Since $\sin x$ is invertible over the interval $[-\pi/2, \pi/2]$, we use the convention that $\sin^{-1} y$ is the value x in the interval $[-\pi/2, \pi/2]$ for which $\sin x = y$. The event $\{Y \le y\}$ is best visualised on a graph as in Figure 4.5 where the *non-shaded* region of the circle and of the sinusoid corresponds to the region for which $\sin X \le y$. From examining this region, we conclude that, for $x = \sin^{-1} y$,

$$F_Y(y) = p(Y \le y)$$

$$= p(X \le x) + \sum_{k=1}^{\infty} p\left((2k-1)\pi - x \ \le \ X \ \le \ 2k\pi + x\right)$$

$$= p(X \le x) + \sum_{k=1}^{\infty} \left[F_X(2k\pi + x) - F_X\left((2k-1)\pi - x\right)\right]$$

$$= p(X \le x) + \sum_{k=1}^{\infty} \left[1 - e^{-\lambda x - 2\lambda k\pi} - 1 + e^{\lambda x + \lambda \pi - 2\lambda k\pi}\right]$$

$$= p(X \le x) + \left[e^{\lambda x} e^{\lambda \pi} - e^{-\lambda x}\right] \sum_{k=1}^{\infty} e^{-2\lambda k\pi}$$

$$= p(X \le \sin^{-1} y) + \left[e^{\lambda(\sin^{-1} y + \pi)} - e^{-\lambda \sin^{-1} y}\right] \frac{e^{-2\lambda \pi}}{1 - e^{-2\lambda \pi}}$$

where we have used the "sum of geometric sequence" formula from your Mathematics Data Book. The first term is zero for negative[2] $y \in [-1, 0)$ and $p(X \le \sin^{-1} y) = F_X(\sin^{-1} y) =$

$1 - e^{-\lambda \sin^{-1} y}$ for non-negative $y \in [0, 1]$. Since $F_X(0) = 0$, the cumulative probability $F_Y(y)$ remains continuous at $y = 0$, but its derivative is discontinuous and we will not be able to derive an expression for $f_Y(0)$. Hence, for negative $y \in [-1, 0)$,

$$f_Y(y) = \frac{d}{dx} F_Y(x) \frac{dx}{dy} = \frac{\lambda}{\sqrt{1 - y^2}} \frac{e^{\lambda(\sin^{-1} y + \pi)} + e^{-\lambda \sin^{-1} y}}{e^{2\lambda \pi} - 1}, \tag{4.15}$$

and for positive $y \in (0, 1]$,

$$f_Y(y) = \frac{d}{dx} F_Y(x) \frac{dx}{dy} = \frac{\lambda}{\sqrt{1 - y^2}} \left[e^{-\lambda \sin^{-1} y} + \frac{e^{\lambda(\sin^{-1} y + \pi)} + e^{-\lambda \sin^{-1} y}}{e^{2\lambda \pi} - 1} \right]. \tag{4.16}$$

Note how these expressions could have been obtained directly from $f_X(x)$ by using the rules we introduced for "increasing" and "decreasing" functions respectively for all the points x such that $g(x) = y$, then summing all the resulting terms. The cumulative probability function $F_Y(y)$ and the density $f_Y(y)$ are plotted in Figure 4.6 for $\lambda = 1$. It is fairly surprising that such a seemingly simple problem yields a discontinuous probability density function for Y, and one that is so blatantly asymmetric between positive and negative values. Can you explain why this is the case?
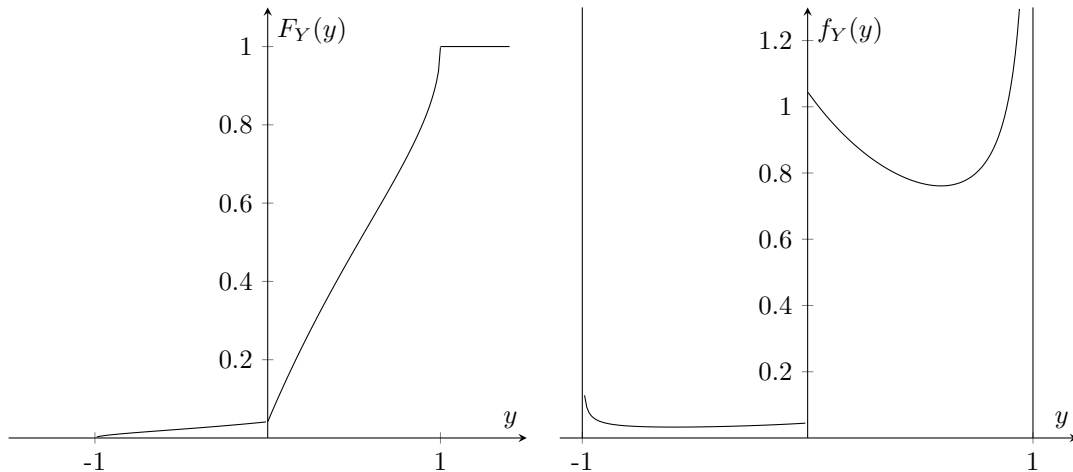
Figure 4.6: Cumulative and density of $Y = \sin X$ for $X \sim \text{Exp}(1)$

Before we move on to other manipulations of random variables, we will consider a very simple function that you will find very useful in practice, e.g., when solving the examples papers. Let X be any random variable with mean $\mu = E[X]$ and variance $\sigma^2 = \text{Var}(X)$, where $\sigma \geq 0$. Now consider the random variable

$$Y = \frac{X - \mu}{\sigma}. \tag{4.17}$$

The function $g(x) = (x - \mu)/\sigma$ is increasing everywhere and $g^{-1}(y) = \sigma y + \mu$, so we can use the expressions we derived for increasing functions to state directly that

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dx}{dy} = \sigma f_X(\sigma y + \mu). \tag{4.18}$$

---

[2]The notation $[a, b)$ specifies an interval closed on the left (including a) and open on the right (excluding b).

We can now compute the moments of Y,

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} y \sigma f_X(\sigma y + \mu) dy \tag{4.19}$$

$$= \int_{-\infty}^{\infty} \frac{x - \mu}{\sigma} f_X(x) dx = \frac{1}{\sigma} \left(\int_{-\infty}^{\infty} x f_X(x) dx - \mu \int_{-\infty}^{\infty} f_X(x) dx \right) \tag{4.20}$$

$$= (E[X] - \mu)/\sigma = 0 \tag{4.21}$$

where we used a variable substitution $y = g(x)$ in the integration (noting that $dx = \sigma dy$), and similarly,

$$E[Y^2] = \int_{-\infty}^{\infty} y^2 f_Y(y) dy = \int_{-\infty}^{\infty} y^2 \sigma f_X(\sigma y + \mu) dy \tag{4.22}$$

$$= \int_{-\infty}^{\infty} \frac{(x - \mu)^2}{\sigma^2} f_X(x) dx = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx \tag{4.23}$$

$$= E[(X - \mu)^2]/\sigma^2 = \text{Var}(X)/\sigma^2 = 1. \tag{4.24}$$

Furthermore, $\text{Var}(Y) = E[Y^2] - E[Y]^2 = 1 - 0^2 = 1$. We conclude that Y is a random variable with mean 0 and variance 1. This is true in particular for a Gaussian random variable $X \sim N(\mu, \sigma^2)$, as already stated when we introduced the Gaussian distribution. This implies for example that if you need to evaluate the cumulative distribution $F_Y(y)$ of a variable $Y \sim N(\mu, \sigma^2)$ at y, you can read out the value $\Phi(x)$ for $x = (y - \mu)/\sigma$ in the table on the last page of your Mathematics Data Book. In 2P6 Communications later this term, your lecturer will make use of this fact in Handout 4 to compute the probability of error in the presence of Gaussian noise, obtaining expressions of the form $Q(A/\sigma) = 1 - \Phi(A/\sigma)$ for an error threshold A and zero-mean Gaussian noise of variance σ^2. It is easy to show that, conversely, if you take any random variable Y with mean zero and variance 1, the random variable $X = aY + b$ has mean b and variance a^2.

4.2 Sums of Random Variables

Let X and Y be two random variables. We are interested in the probability distribution of the sum $S = X + Y$ of the random variables X and Y. This will in general take a form of a probability distribution P_S if X and Y are discrete, or cumulative probability F_S and density function f_S if X and Y are continuous. In general, we will always assume in this section that X and Y are *independent* random variables although we will begin by looking at expectations and moments, for which we make no such assumption.

4.2.1 Mean and variance of sums

The first statements we will make concern the mean and variance of S and follow directly from the linearity of expectation shown in Section 1.6 for discrete random variables and easily extended to continuous random variables:

$$E[S] = E[X + Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{XY}(x, y) \, dx \, dy \tag{4.25}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) \, dx \, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) \, dx \, dy \tag{4.26}$$

$$= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f_{XY}(x, y) \, dy \right) dx + \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{XY}(x, y) \, dx \right) dy \tag{4.27}$$

$$= \int_{-\infty}^{\infty} x f_X(x) \, dx + \int_{-\infty}^{\infty} y f_Y(y) \, dy = E[X] + E[Y]. \tag{4.28}$$

Hence the expectation of a sum S is the sum of the expectations. Note that we have not used independence anywhere in this derivation and hence this statement holds true for any andom variables X and Y, dependent or independent. We now look at variances, again making no assumption about indepedence, and write

$$\text{Var}(S) = \text{Var}(X + Y) = E[(X + Y)^2] - (E[X + Y])^2 \tag{4.29}$$

$$= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 \tag{4.30}$$

$$= E[X^2] - (E[X])^2 + E[Y^2] - (E[Y])^2 + 2(E[XY] - E[X]E[Y]) \tag{4.31}$$

$$= \text{Var}(X) + \text{Var}(Y) + 2(E[XY] - E[X]E[Y]) \tag{4.32}$$

where we used (4.28) in every step of the derivation. The last term in the expression is known as the covariance

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] \tag{4.33}$$

so the expression can be summarised as

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y). \tag{4.34}$$

It is easy to see that the covariance is zero when X and Y are independent, by verifying that $E[XY] = E[X]E[Y]$ for independent random variables, which was shown in (1.49) for discrete random variables and can easily be shown similarly for continuous random variables, so we conclude that

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \text{ when } X \text{ and } Y \text{ are independent,} \tag{4.35}$$

i.e., the variance of a sum of independent random variables is the sum of the variances.

Another measure related to the covariance is the *correlation coefficient*

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} \tag{4.36}$$

and satsifies $-1 < \rho < 1$. When $\rho > 0$ the variables are called *correlated* and when $\rho < 0$ they are anti-correlated. When $\rho = 0$ the variables are called *uncorrelated*. This is often taken as a "poor man's independence" because it is easier to verify experimentally than full independence. It is important to understand that independent random variables are always uncorrelated, but uncorrelated random variables can be dependent. The condition for variables to be uncorrelated is a lot weaker than the condition for them to be independent.

We had already observed these properties of mean and variance for discrete random variables and used them to compute the expectation and variance of the binomial distribution that was introduced as the probability distribution of a sum of independent Bernoulli random variables with parameter p, but we now know that they apply to all random variables. Having sorted out means and variances, we are ready to take on the more difficult problem of determining the full probability distribution or density function of S given those of X and Y. For this, we will handle the discrete and continuous cases separately.

4.2.2 Sums of Independent Discrete Random Variables

For discrete random variables, we assume a sum $S = X + Y$ of two *independent* random variables taking values in sets of numbers $\mathcal{X}$ and $\mathcal{Y}$. We can write the distribution as a probability of the associated event

$$P_S(s) = p(S = s) = p(X + Y = s) \tag{4.37}$$

$$= \sum_{(x,y):x+y=s} P_{XY}(x, y) \tag{4.38}$$

$$= \sum_{(x,y):x+y=s} P_X(x)P_Y(y) \tag{4.39}$$

where the last step follows if X and Y are independent. At this point, it is worth pausing to reflect about what we mean by '+' when we say $S = X+Y$. I hear you worry that I am about to embark on a primary school revision exercise, re-deriving addition from first principles. However, there are many applications in computer science, communications and information theory where we don't mean integer addition when we write '+'. You have already encoutered the "modulo" addition in an example earlier in this chapter, and computer, electronic, or communications engineers may well mean "modulo n" addition when they write "$X + Y$". You also learned about alternative addition rules in your IA Paper 4 Computing course and in the IA Paper 3 Digital Circuits and Information Processes.

For now, we will restrict ourselves to plain integer addition. In order to further develop the expression above, we also need to specify that the sets $\mathcal{X}$ and $\mathcal{Y}$ of possible values of X and Y consist of a range of consecutive integers, possibly infinite and possibly including negative numbers, and any sum we write on elements of the set is taken in order from the smallest to the largest integer. If so, we can continue

$$P_S(s) = \sum_{(x,y):x+y=s} P_{XY}(x,y) = \sum_{x \in \mathcal{X}} P_{XY}(x, s-x) \tag{4.40}$$

$$= \sum_{x \in \mathcal{X}} P_X(x) P_Y(s-x) \tag{4.41}$$

where the condition about consecutive integers ensures in the last line that $s - x$ is a valid value in $\mathcal{Y}$ for the random variable Y. Pause a while to reflect on (4.41). We will return to it in the next section after taking a first look at the sum of continuous random variables.

Example 4.6. Let X and Y be random variables indicating the value of each dice in a two-fair-dice throwing experiment, and $S = X + Y$. We have

$$\begin{cases} P_S(2) & = P_X(1)P_Y(1) = 1/36 \\ P_S(3) & = P_X(1)P_Y(2) + P_X(2)P_Y(1) = 2/36 \\ P_S(4) & = P_X(1)P_Y(3) + P_X(2)P_Y(2) + P_X(3)P_Y(1) = 3/36 \\ P_S(5) & = P_X(1)P_Y(4) + P_X(2)P_Y(3) + P_X(3)P_Y(2) + P_X(4)P_Y(1) = 4/36 \\ P_S(6) & = P_X(1)P_Y(5) + P_X(2)P_Y(4) + P_X(3)P_Y(3) + P_X(4)P_Y(2) + P_X(5)P_Y(1) = 5/36 \\ P_S(7) & = P_X(1)P_Y(6) + P_X(2)P_Y(5) + P_X(3)P_Y(4) \\ & \quad + P_X(4)P_Y(3) + P_X(5)P_Y(2) + P_X(6)P_Y(1) = 6/36 \\ P_S(8) & = P_X(2)P_Y(6) + P_X(3)P_Y(5) + P_X(4)P_Y(4) + P_X(5)P_Y(3) + P_X(6)P_Y(2) = 5/36 \\ P_S(9) & = P_X(3)P_Y(6) + P_X(4)P_Y(5) + P_X(5)P_Y(4) + P_X(6)P_Y(3) = 4/36 \\ P_S(10) & = P_X(4)P_Y(6) + P_X(5)P_Y(5) + P_X(6)P_Y(4) = 3/36 \\ P_S(11) & = P_X(5)P_Y(6) + P_X(6)P_Y(5) = 2/36 \\ P_S(12) & = P_X(6)P_Y(6) = 1/36. \end{cases}$$

$$\tag{4.42}$$

4.2.3   Sums of Independent Continuous Random Variables

For continuous random variables, the derivation is similar if you use infinitesimal probability calculus. Despite my insistence so far to show you how results can be derived directly from probabilities using the cumulative probability function, the following is one result that cannot be derived in this manner without unnecessarily complicating things. We will hence resort to infinitesimal calculus to compute the density of $S = X + Y$ given the densities of the *independent* continuous random variables X and Y. Note that X and Y are always assumed to be real-valued

in the continuous case and there is no ambiguity about the nature of the sum as there was in the discrete case.

Recall that $f_X(x)\,dx$ is the infinitesimal probability of being in the interval $[x, x+dx)$. Hence,

$$f_S(s)ds = \int_{(x,y):x+y=s} f_X(x)\,dx f_Y(y)\,dy \tag{4.43}$$

$$= \left(\int_{-\infty}^{\infty} f_X(x) f_Y(s-x)\,dx \right) ds \tag{4.44}$$

where we have used a change of variable $s = y+x$ to obtain the second expression, and noted that s being a constant in the first expression no integration results over the variable s. This leads to

$$f_S(s) = \int_{-\infty}^{\infty} f_X(x) f_Y(s-x)\,dx. \tag{4.45}$$
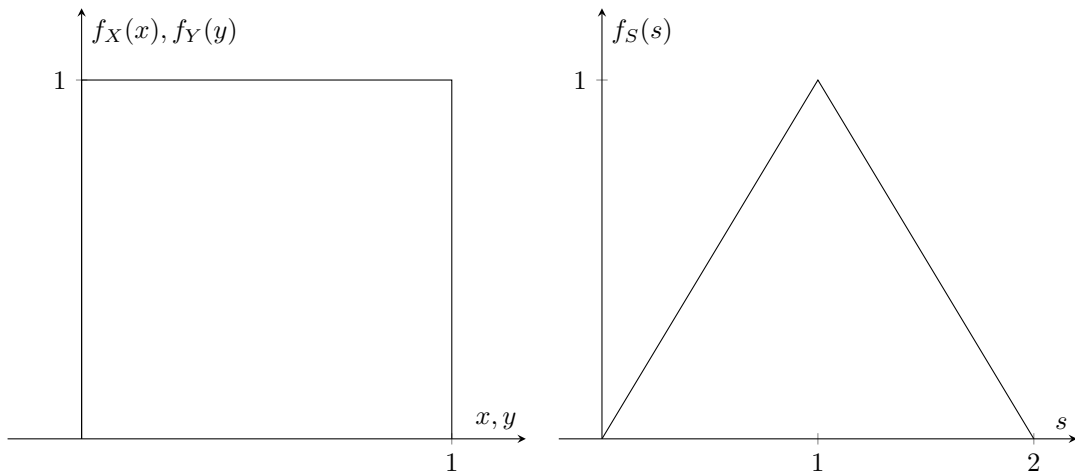
Figure 4.7: Sum of two uniformly distributed random variables

Example 4.7. Let X and Y be independent and uniform between 0 and 1, i.e.,

$$f_X(x) = f_Y(x) = \begin{cases} 1 \text{ for } x \in [0,1], \\ 0 \text{ otherwise} \end{cases} \tag{4.46}$$

and $S = X + Y$.

We use the expression above to get

$$f_S(s) = \int_0^1 f_X(x) f_Y(s-x)\,dx = \int_0^1 f_Y(s-x)\,dx = \begin{cases} \int_0^s 1\,dx = s \text{ for } s \in [0,1] \\ \int_{s-1}^1 1\,dx = 2-s \text{ for } s \in [1,2] \end{cases}$$
$$\tag{4.47}$$

giving a triangular distribution as illustrated in Figure 4.7.

The careful reader will have identified the expression obtained for the sum of two continuous random variables as a *convolution* $f_X \star f_Y$ of the probability density functions f_X and f_Y, as learned in IA Paper 4. The expression $\sum_x P_X(x) P_Y(s-x)$ in (4.41) for discrete random variable also appears to be a sort of discrete convolution. You may not have encountered a discrete

Variable	Sum limits	Transform name	Taught in...	Application
z^{-1}	0 to ∞	z-Transform	Part IIA	Digital control and signal processing
D	0 to $n-1$	D-Transform	Part IIB	Coding theory (error correction)
$e^{j\frac{2\pi}{n}}$	0 to $n-1$	Discrete Fourier Transform	IB Paper 6	Signal processing (cyclic convolution!)
z	all $x \in \mathcal{X}$	PGF	This course	Discrete probability distributions

Table 4.1: Common discrete tranforms

convolution in your studies in Part I so far, but it is in fact very common, for example when studying the discrete ("digital") equivalent of the linear time-invariant systems, and you may see many examples of this in Part II. It should therefore come as no surprise that we will use transforms of distributions and densities in the next section to tackle convolutions efficiently.

4.3   Transforms of distributions

You have already encountered a number of transforms in IA Paper 4 (Fourier Series, Laplace Transform), IB Paper 6 (Fourier Transform), and are about to encounter more. Each transform has its own special properties that make it suitable for one or the other application, but they all have the "convolution property" in common (see Information data book, table on Page 1, second to last row.) We will introduce two transforms, one for sums of discrete random variables, and another one for sums of continuous random variables.

4.3.1   The Probability Generating Function (PGF)

Before introducing yet another transform, let us first consider a generic model for a discrete transform of a sequence $u_1, u_2, u_3, \ldots$

$$U(z) = \sum_k u_k z^k \tag{4.48}$$

where "z" is the variable in the "transform domain". If we multiply the transforms of two sequences $u_1, u_2, \ldots$ and $v_1, v_2, \ldots$ with each other, we obtain

$$U(z)V(z) = \left(\sum_m u_m z^m\right)\left(\sum_n v_n z^n\right) \tag{4.49}$$

$$= \sum_m \sum_n u_m v_n z^{m+n} \tag{4.50}$$

$$= \sum_k \left(\sum_m u_m v_{k-m}\right) z^k \tag{4.51}$$

where we used the substitution $k = m + n$ in the last step. We see that the result is the transform of the convolution $(u_1, u_2, \ldots) \star (v_1, v_2, \ldots)$. The transforms you've learned or have yet to learn may have different names for the transform domain variable "z", but they all have the convolution property as a consequence of the simple derivation above. Table 4.1 provides a list of common discrete transforms, most of which are in your information and mathematics data books[3]. The last entry in the table is the transform we will be using for discrete probability distributions:

---

[3]Note that we are only providing this table to help you place the new transforms we introduce within the rich world of transforms that share the "convolution property." You are not expected to learn or understand all the transforms for this course and only need to learn the PGF and MGF.

Name	Parameters	Distribution $P_X(x)$	PGF $g_X(z) = E[z^X]$
Bernoulli	$\text{Ber}(p)$	$(p, 1-p)$	$1-p+pz$
Binomial	$\text{B}(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$(1-p+pz)^n$
Geometric (1)	$\text{Geom}(p)$	$p(1-p)^{x-1}, x = 1, 2, \ldots$	$\frac{pz}{1-(1-p)z}$
Geometric (2)	p	$p(1-p)^x, x = 0, 1, \ldots$	$\frac{p}{1-(1-p)z}$
Poisson	$\text{Po}(\lambda)$	$\frac{\lambda^x}{x!}e^{-\lambda}$	$e^{\lambda(z-1)}$

Table 4.2: Probability Generating Functions (PGF) for various discrete distributions

---

Probability Generating Function (PGF): for a discrete random variable X taking values on the set $\mathcal{X}$, the probability generating function (PGF) is defined as

$$g_X(z) = \sum_{x \in \mathcal{X}} P_X(x)z^x = E[z^X]. \tag{4.52}$$

Note the expectation notation that piggybacks the definition of the PGF on the definition of the expectation. It is often the preferred expression in textbooks, although it fails to emphasise the transform nature of the PGF, which is its main feature in my view. A table of discrete probability distributions and their PGFs is given in Table 4.2. A similar table is also available in your mathematics data book on page 27. All of these expressions are easy to derive using power series and the sum of geometric series from your data book, and you are encouraged to do so as an exercise.

The PGF of the binomial deserves a special mention. We have introduced the binomial as the probability distribution of the sum $Y = X_1 + X_2 + \ldots + X_n$ of n independent Bernoulli random variables with parameter p. Hence the probability distribution of Y results from the convolution of the $\text{Ber}(p)$ distribution n times with itself

$$P_Y(y) = (P_X \star P_X \star \ldots \star P_X)(y) \tag{4.53}$$

which is tedious to evaluate or write down as a multiple summation. We introduced the PGF as a means to evaluate convolutions efficiently as multiplications in the transform domain, and hence we see that, given that the Bernoulli PGF is

$$g_X(z) = 1 - p + pz, \tag{4.54}$$

we obtain the result in one step

$$g_Y(z) = (g_X(z))^n = (1 - p + pz)^n. \tag{4.55}$$

It is interesting to note that the binomial and Poisson distributions are both *closed* under addition: consider a sum $S = X + Y$ of two random variables X and Y. If X and Y are binomial $X \sim B(n_1, p)$ and $Y \sim B(n_2, p)$ with the *same parameter* p, then

$$g_S(z) = g_X(z)g_Y(z) = (1-p+pz)^{n_1}(1-p+pz)^{n_2} = (1-p+pz)^{n_1+n_2} \tag{4.56}$$

so S is binomial $S \sim B(n_1 + n_2, p)$. This not surprising given that S is the sum of $n_1 + n_2$ independent Bernoulli trials. If X and Y are Poisson $X \sim \text{Po}(\lambda_1)$ and $Y \sim \text{Po}(\lambda_2)$ with PGFs

$$g_X(z) = e^{\lambda_1(z-1)} \quad \text{and} \quad g_Y(z) = e^{\lambda_2(z-1)}, \tag{4.57}$$

then

$$g_S(z) = g_X(z)g_Y(z) = e^{\lambda_1(z-1)}e^{\lambda_2(z-1)} = e^{(\lambda_1+\lambda_2)(z-1)} \tag{4.58}$$

Variable	Integration limits	Transform	Taught in...	Application
$-s$	0 to ∞	Laplace Tranform	IA Paper 4	Control theory
$-j\omega$	$-\infty$ to ∞	Fourier Transform	IB Paper 6	Signal processing
$-j\frac{2\pi n}{T}$	0 to T	Fourier Series	IA Paper 4	Periodic functions
s	$-\infty$ to ∞	MGF	This course	Continuous probability densities

Table 4.3: Common continuous tranforms

which implies that S is a Poisson distributed random variable $S \sim \text{Po}(\lambda_1 + \lambda_2)$. Again not surprising if you consider two parallel Poisson processes with rates λ_1 and λ_2 of occurrence, and you then mix the two processes so that all occurrences of one or the other process count equally. This is clearly then a Poisson process with rate $\lambda_1 + \lambda_2$ occurrences per time interval.

A property of the PGF that is also stated in your databook is the following

$$
\begin{cases}
g_X'(1) = \sum_{x \in \mathcal{X}} x P_X(x) z^{x-1}\big|_{z=1} = E[X] \\
g_X''(1) = \sum_{x \in \mathcal{X}} x(x-1) P_X(x) z^{x-2}\big|_{z=1} = E[X^2] - E[X] \\
g_X^{(k)}(1) = E[X(X-1)(X-2)\cdots(X-k+1)]
\end{cases}
\tag{4.59}
$$

Although this ability to express moments from its derivatives is described in many textbooks as the "raison d'être" of the PGF, it is in fact a trivial consequence of the definition of the PGF and rarely useful. It is easy to compute the first and second moments of a sum of independent random variables without the PGF by using the properties described at the beginning of this section. One is rarely interested in higher moments of a distribution.

4.3.2   The Moment Generating Function (MGF)

For continuous random variables, the approach is very similar to the discrete case. We can again consider a "generic" transform for a function $f(x)$ of the form

$$
F(s) = \int f(x) e^{sx} \, dx.
\tag{4.60}
$$

Note that we used the exponential function rather than z^x as we did in the discrete case, because z^x is sometimes ill defined for continuous x. We've left the integration limits unspecified in our generic transform but each transform will have specific limits (as was the case for discrete transforms) and these can be infinite. Note that we called our generic "transform variable" s instead of z as in the discrete case simply to follow convention. Again, it is easy to show the convolution property for two function $g(x)$ and $f(x)$,

$$
G(s)F(s) = \left(\int g(x) e^{sx} \, dx \right) \left(\int f(y) e^{sy} \, dy \right)
\tag{4.61}
$$

$$
= \int \int g(x) f(y) e^{s(x+y)} \, dx \, dy
\tag{4.62}
$$

$$
= \int \left(\int g(x) f(z-x) \, dx \right) e^{sz} \, dz
\tag{4.63}
$$

where we have applied the change of variable $z = x + y$ and note that the integration limits may need adapting as a result (but these are infinite in many cases). As in the discrete case, the transforms you learned may have different names for the tranform domain variable "s", but they all have the convolution property. Table 4.3 provides a list of common continuous tranforms. The last entry in the table is the transform we will be using for continuous probability density functions:

Name	Parameters	Density $f_X(x)$	MGF $g_X(s) = E[e^{sX}]$
Uniform	over $[0, 1]$	1	$\frac{1}{s}(e^s - 1)$
Exponential	$\text{Exp}(\lambda)$	$\lambda e^{-\lambda x}$	$\frac{\lambda}{\lambda - s}$
Gaussian	$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$e^{s\mu + \sigma^2 s^2/2}$
Standard Gaussian	$N(0, 1)$	$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$e^{s^2/2}$

Table 4.4: Moment Generating Functions (MGF) for various continuous densities

---

Moment Generating Function (MGF): for a continuous random variable X, the moment generating function (MGF) is defined[4] as

$$g_X(s) = \int_{-\infty}^{\infty} f_X(x) e^{sx} \, dx = E[e^{sX}]. \tag{4.64}$$

The MGF can be seen as a two-sided generalisation of the Laplace transform. Note again the expectation notation that provides a compact definition but hides the transform nature of the MGF[5].

A table of continuous probability distributions and their MGFs is given in Table 4.4. A similar table is also available in your mathematics data book on page 28. These expression are mostly easy to derive and you are encouraged to do so. We derive the MGF of the standard Gaussian distribution

$$g_X(s) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{xs} \, dx \tag{4.65}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2 - 2xs}{2}} \, dx \tag{4.66}$$

$$= e^{s^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-s)^2}{2}} \, dx = e^{s^2/2} \tag{4.67}$$

where the last step follows by recognising that the integral is that of a Gaussian density with variance 1 and mean s and hence integrates to 1. It is easy to show that, for $Y = \alpha X$

$$g_Y(s) = g_X(\alpha s) \tag{4.68}$$

and, for $Y = X + \beta$,

$$g_Y(s) = e^{\beta s} g_X(s) \tag{4.69}$$

and hence determine the MGF of general Gaussian variables from the standard Gaussian expression above.

---

[4]If you have an old pre-2017 mathematics databook, the MGF was defined as $E[e^{-sX}]$. This is not the standard definition given in the probability textbooks that I checked. The MGF is normally defined as $E[e^{sX}]$ (no minus in the exponent.) This has been corrected in the current version of the databook.

[5]You may be wondering why we could not just pick existing transforms to perform convolutions of probability distributions (e.g. z-transform) or densities (e.g. Fourier transform.) The answer is in part that we needed slightly different properties (e.g., sum of probabilities over the values in $\mathcal{X}$) and in part that other transforms are in fact used as well. In particular, the Fourier transform of a density is called its *characteristic function*, can be expressed as $g_X(j\omega)$, and can be used interchangeably with the MGF in most contexts. Furthermore, when tackling other types of addition for discrete random variables such as *modulo* addition, the Discrete Fourier Transform and other related transforms are used. Your mobile phone, your broadband router, and your digital TV receiver ("freeview", cable, etc.), are all evaluating Discrete Fourier Transforms of probability distributions thousands of times per second in order to tackle modulo sums of Bernoulli random variables when decoding incoming signals. You can learn more about this in the Part II module 3F7 "Information Theory".

Now consider two independent Gaussian random variables $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and their sum $Z = X + Y$. The MGF of the sum is

$$g_Z(s) = e^{s\mu_1 + \frac{s^2\sigma_1^2}{2}} e^{s\mu_2 + \frac{s^2\sigma_2^2}{2}} = e^{s(\mu_1+\mu_2) + \frac{s^2(\sigma_1^2+\sigma_2^2)}{2}} \tag{4.70}$$

showing that Z is a Gaussian random variable with the sum of the means and the sum of the variances, $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Hence the Gaussian density is *closed* under addition of random variables.

Finally, like the PGF, the MGF can be used to compute moments of random variables. Note that

$$g'_X(0) = \int_{-\infty}^{\infty} x f_X(x) e^{sx} \, dx \bigg|_{s=0} = E[X]. \tag{4.71}$$

This argument can be repeated recursively to obtain

$$\begin{cases} E[X] = g'(0) \\ E[X^2] = g''(0) \\ E[X^n] = g^{(n)}(0). \end{cases} \tag{4.72}$$

Hence the name "moment generating function", Note that the last expression is true for all n as can easily be verified, and even for $n = 0$ since

$$g(0) = \int_{-\infty}^{\infty} f_X(x) e^{0x} \, dx = \int_{-\infty}^{\infty} f_X(x) \, dx = 1 = E[X^0] = E[1]. \tag{4.73}$$

4.4 The Central Limit Theorem

Having analysed sums of random variables and introduced the MGF, we are ready to state and prove one of the pivotal results of probability theory:

The Central Limit Theorem: let $X_1, X_2, \ldots$ be independent random variables with means $\mu_1, \mu_2, \ldots$ and variances $\sigma_1^2, \sigma_2^2, \ldots$. Assume that the random variables are all continuous with any probability density functions whose MGF exist. In particular, the densities can all be different. Then the random variable

$$Y_n = X_1 + X_2 + \ldots + X_n \tag{4.74}$$

tends to a Gaussian random variable Y,

$$Y \sim N\left(\mu_1 + \mu_2 + \ldots + \mu_n, \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2\right) \tag{4.75}$$

as n grows to infinity.

The central limit theorem is the reason why the Gaussian distribution was described as being "of central importance" to probability theory when we first introduced it. Whenever quantities emerge as a sum of many effects, they tend to be Gaussian. Having learned that the density of a sum is the convolution of the densities of its terms, this also means more generally that if you pick any functions that have a well-behaved transform, their convolution will tends towards a function of the form e^{-x^2}, which is rather surprising.

We will first make a few observations about the theorem and then provide a proof. First, note that, since we've already shown that the mean of the sum is always the sum of the means, and the sum of the variances of independent random variables is the sum of the variances, we immediately see that

$$E[Y_n] = \mu_1 + \mu_2 + \ldots + \mu_n \quad \text{and} \quad \text{Var}(Y_n) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2. \tag{4.76}$$

Hence, the "surprising" part of the theorem is the fact that Y_n tends to a Gaussian random variable rather than that its mean and variances are as stated.

What does it mean that Y_n "tends to" Y? This is in fact a subtle question and one that mathematicians would approach with careful formalism. For our purposes, it suffices to imagine that the density of Y_n is "close" to the density of Y in some sense, and that it becomes "closer" as n grows. The proof below will give us a better understanding of this process.

The central limit theorem is often presented in the form of an "average" of random variables instead of a sum. If we multiply the sum by a factor α, since $\alpha Y_n = \alpha X_1 + \alpha X_2 + \ldots + \alpha X_n$, we conclude that αY_n has mean $E[\alpha X_1] + E[\alpha X_2] + \ldots + E[\alpha X_n]$ and variance $\operatorname{Var}(\alpha X_1) + \operatorname{Var}(\alpha X_2) + \ldots + \operatorname{Var}(\alpha X_n)$. Now,

$$E[\alpha X_k] = \alpha E[X_k] = \alpha \mu_k \tag{4.77}$$

for all k, and

$$\operatorname{Var}(\alpha X_k) = E[(\alpha X_k)^2] - E[\alpha X_k]^2 = \alpha^2 \operatorname{Var}(X_k) \tag{4.78}$$

for all k. Hence, we obtain an alternative version of the central limit theorem by picking $\alpha = 1/n$, to yield that

$$\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \longrightarrow Y \sim \mathrm{N}\left(\frac{\mu_1 + \mu_2 + \ldots + \mu_n}{n}, \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n^2}\right). \tag{4.79}$$

Note that if $\mu_1 = \mu_2 = \ldots = \mu_n = \mu$ and $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2$, i.e., the random variables $X_1, X_2, \ldots$ are indepedent with equal mean and variance, the equation above implies that

$$\frac{1}{n}(X_1 + X_2 + \ldots + X_n) \longrightarrow Y \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \tag{4.80}$$

i.e., the mean is the same as the mean of each variable, while variance becomes smaller as more random variables are added. This is the basis for averaging a sequence of random variables over time, as the probability that the observed value is close to the mean increases as the number of observations averaged increases. We will re-visit this fact at the end of this section to discuss a number of probabilistic characteristics that can be estimated statistically from data.

Another choice $\alpha = 1/\sqrt{n}$ would give a sum whose mean grows as $\mu\sqrt{n}$ and whose variance remains constant at σ^2. This is of no practical relevance but we will use this $\alpha = 1/\sqrt{n}$ in our proof of the central limit theorem below.

Proof of the central limit theorem: without loss of generality, let us assume that $\mu_1 = \mu_2 = \ldots = \mu_n = 0$ and that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = 1$. Our approach[6] will seek to prove that

$$Y_n = \frac{1}{\sqrt{n}}(X_1 + X_2 + \ldots + X_n) \tag{4.81}$$

tends to a standard Gaussian distribution $N(0,1)$. We already know from the discussion above that $E[Y_n] = \mu\sqrt{n} = 0$ and $\operatorname{Var}(Y_n) = \sigma^2 = 1$. Given that every variable X_k for $k = 1, 2, \ldots$ has an MGF, we can write this MGF as a Taylor series around 0, to yield

$$g_{X_k}(s) = g_{X_k}(0) + g'_{X_k}(0)s + g''_{X_k}(0)\frac{s^2}{2} + g^{(3)}(0)\frac{s^3}{3!} + \ldots \tag{4.82}$$

Recall that $g_{X_k}^{(m)}(0) = E[X_k^m]$. Hence we can re-write the expression above as

$$g_{X_k}(s) = E[X_k^0] + E[X_k]s + E[X_k^2]\frac{s^2}{2} + o(s^3) \tag{4.83}$$

$$= 1 + \mu_k s + (\sigma_k^2 + \mu_k^2)\frac{s^2}{2} + o(s^3) \tag{4.84}$$

$$= 1 + \frac{s^2}{2} + o(s^3). \tag{4.85}$$

Note how, despite the fact that the densities of X_k may all be different, the fact that they all have mean 0 and variance 1 implies that the terms up to the quadratic in the Taylor series of their MGF must be the same. We can now use the convolution property to observe that

$$g_{Y_n}(s) = \prod_{k=1}^{n} g_{X_k}\left(\frac{s}{\sqrt{n}}\right) \tag{4.86}$$

$$= \prod_{k=1}^{n}\left[\left(1 + \frac{s^2}{2n}\right) + o\left(\frac{s^3}{n^{3/2}}\right)\right] \tag{4.87}$$

$$= \left(1 + \frac{s^2}{2n}\right)^n + o\left(\frac{s^3}{n^{3/2}}\right). \tag{4.88}$$

For any given s,

$$\lim_{n\to\infty} g_{Y_n}(s) = \lim_{n\to\infty}\left(1 + \frac{s^2}{2n}\right)^n + o\left(\left(\frac{s}{\sqrt{n}}\right)^3\right) = e^{s^2/2} \tag{4.89}$$

where we have used the limit $(1 + x/n)^n \to e^x$ from your Mathematics Data Book page 2, and the fact that all terms of power 3 and above of $s/\sqrt{n}$ tend to zero as n grows large. We have shown that the MGF of Y_n tends to the MGF of a standard Gaussian $e^{s^2/2}$ as n grows, indicating that the density of Y_n approaches the standard Gaussian density and hence proving the theorem.

The central limit theorem (CLT) and a more general but weaker result called the *law of large numbers* (LLN) form the basis for statistical estimation of probabilities from data. We have already had a foretaste of estimating probablity parameters from data when we studied the Beta density in Section 3.5. Both the CLT and the LLN can be used to show that, for a sequence of independent random variables $X_1, \ldots, X_n$ that habe the same mean and variance, the time average of any function of the variables

$$Z_n = \frac{1}{n}(g(X_1) + g(X_2) + \ldots + g(X_n)) \tag{4.90}$$

is a random variable with expectation $E[g(X)]$ whose variance decreases with n, and hence computing Z_n from data gives an estimate of $E[g(X)]$ whose accuracy increases with n. This ability to estimate probabistic characteristics from data can be used to ascertain that data follows a given probabilistic model. There are many characteristics of probability distributions or densities that can be estimated from data, such as:

- the mean or expectation $E[X] = \mu$

- the variance $\text{Var}(X) = \sigma^2 = E[X^2] - \mu^2$ and the standard deviation $\sigma = \sqrt{\text{Var}(X)}$

- the skewness $E[(X - \mu)^3]/\sigma^3$. If the skewness is positive, the distribution (or density) is *skewed to the right.* Informally the 'tail' of the distribution (or density) is longer to the right.

There are other characteristics of probability distributions or densities that can be estimated from data but don't rely on the CLT or the LLN:

- the mode $\max_x P_X(x)$ or $\max_x f_X(x)$

- the median M such that $F_X(M) = 1/2$ can be estimated by the data middle: order all n data points in decreasing order and pick the $n/2$ point

---

[6]Note that this is equivalent to the more general theorem because we can shift and scale all random variables and track the effect on Y_n to show that the sum is a general Gaussian random variable, and inversely we can map the general case to this special case by shifting all random variables so their means are zero and their variances 1.
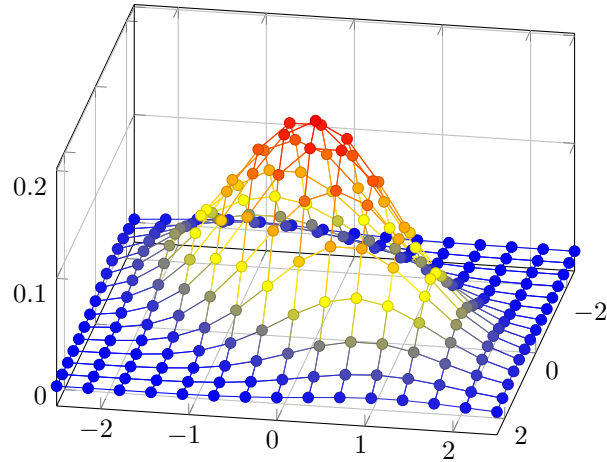
Figure 4.8: Two-variable Gaussian distribution with zero means, unit variances, and $E[X_1 X_2] = 1/2$

- the quartiles Q_1 and Q_3 such that $F_X(Q_1) = 1/4$ and $F_X(Q_3) = 3/4$ and the interquartile range $Q_3 - Q_1$ can be estimated similarly to the median

---

Example 4.8. The distribution of income in Britain is skewed to the right. The average income is very different from the median, since a few people have very large incomes. The logarithm of the income is much less skewed.

Sometimes the *variance* of the distribution can be heavily influenced by very few observations. The *interquartile range* also quantifies the *spread* of a distribution, but it is said to be *more robust toward outliers*.

4.5 Multivariate Gaussians

Although we have introduced joint probabilities and densities and learned how to manipulate them, we have not studied any specific joint densities so far as we have univariate densities (exponentials, Gaussian etc.) One multivariate density that plays am important role in engineering applications is the multivariate Gaussian. The random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is multivariate Gaussian

$$\mathbf{X} \sim \mathrm{N}(\mu, \Sigma), \quad \text{if} \quad f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)} \tag{4.91}$$

were Σ is the $n \times n$ *covariance matrix* whose elements are

$$\Sigma_{km} = E[(X_k - \mu_k)(X_m - \mu_m)] = \begin{cases} \mathrm{Cov}(X_k, X_m) \text{ for } k \neq m \\ \mathrm{Var}(X_k) \text{ for } k = m \end{cases} \tag{4.92}$$

and where $\mu_k = E[X_k]$ for all k and

$$\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T \tag{4.93}$$

is the *mean vector*. Figure 4.8 illustrates a bi-variate Gaussian density with $\mu = (0, 0)$ and

$$\Sigma = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}. \tag{4.94}$$

The notation for multivariate Gaussian densities may be intimidating at first. Let us analyse the expressions stated:

- remember that $\mathbf{X}$ is a random vector and its possible values $\mathbf{x}$ are also vectors. $\mathbf{x}^T$ denotes transposition.

- the density is a scalar function as before. We can interpret it as the infinitesimal probability $f_{\mathbf{X}}(\mathbf{x})dx_1 dx_2 \ldots dx_n$ that $\mathbf{X}$ lies in the region $[x_1, x_1+dx_1)\times[x_2, x_2+dx_2)\times\ldots\times[x_n, x_n+dx_n)$ of n dimensional space. For example, for a 2-dimensional density, $f_{\mathbf{X}}((1,2))dxdy$ gives the infinitesimal probability of being in a square of dimensions $dx \times dy$ whose bottom left corner is $(1,2)$.

- The letter $\mathbf{\Sigma}$ normally used for summation is used to denote a matrix here. We have followed convention in adopting this notation even though I admit that it is rather bizarre.

- The position of the variance in the factor $1/\sqrt{2\pi\sigma^2}$ of the univariate Gaussian distribution is now occupied by the *determinant* of $\mathbf{\Sigma}$ in the multivariate expression.

- The expression in the exponent is a vector-matrix-vector multiplication but you can persuade yourself that the result is scalar, so the exponential is simply the scalar exponential.

- The position of the variance in the exponent $1/(2\sigma^2)$ of the univariate Gaussian density is now taken up by the *inverse* $\mathbf{\Sigma}^{-1}$of $\mathbf{\Sigma}$.

Consider the covariance matrix $\mathbf{\Sigma}$. It diagonal elements $\Sigma_{kk} = E[X_k^2] - \mu_k^2 = \text{Var}(X_k)$ are simply the variances of the individual component random variables X_k. For any k and m, $\Sigma_{km} = \Sigma_{mk}$ since the definition of the matrix elements is symmetric, and hence the covariance matrix is symmetric. If two components X_k and X_m are independent, then $\Sigma_{km} = E[X_k]E[X_m]-\mu_k\mu_m = 0$.

The entry Σ_{km} in the covariance matrix is the *covariance* $\text{Cov}(X_k, X_m)$ of X_k and X_m, as defined in section 4.2.1, where we had observed that independent random variables always had zero covariance but it was possible for variables with zero covariance not to be independent. However, for multi-variate Gaussians it is easy to see that uncorrelated implies independent and vice-versa. The covariance matrix $\mathbf{\Sigma}$ in that case is diagonal and the density is simply the product of univariate Gaussian densities.

It can be shown that the marginal density of any component of a multivariate Gaussian is a univariate Gaussian, i.e.,

$$f_{X_n}(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_{n-1} = \frac{1}{\sqrt{2\pi\Sigma_{nn}}} e^{\frac{(x-\mu_k)^2}{2\Sigma_{nn}}}. \quad (4.95)$$

In fact, if we perform any partial marginalisation of a multi-variate Gaussian, we obtain another multi-variate Gaussian with reduced dimensions. For example, the vector $(X_1, \ldots, X_k)$ for $k < n$ is multi-variate Gaussian

$$(X_1, \ldots, X_k) \sim \text{N}(\mu_{1,\ldots,k}, \mathbf{\Sigma}_{k \times k}) \quad (4.96)$$

where $\mathbf{\Sigma}_{k \times k}$ is the $k \times k$ submatrix of $\mathbf{\Sigma}$ consisting of its first k rows and first k columns.

Furthermore, conditional densities of variables that are jointly multi-variate Gaussian are also always uni-variate or multi-variate Gaussian. If we write

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix} \quad (4.97)$$

where $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$ are block matrices of dimensions $k \times k$, $k \times (n - k)$ and $(n - k) \times (n - k)$, respectively, and

$$\mu = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \quad (4.98)$$

where $\mathbf{a}$ and $\mathbf{b}$ are vectors of length k and $(n - k)$, respectively then one can show that

$$f_{(X_1,\ldots,X_k)|(X_{k+1},\ldots,X_n)}(\mathbf{x}\,|\,\mathbf{y}) \sim \text{N}(a + CB^{-1}(\mathbf{y} - \mathbf{b}), A - CB^{-1}C^T). \quad (4.99)$$

Note how the conditional mean vector of $(X_1, \ldots, X_k)$ depends on the value $\mathbf{y}$ of $(X_{k+1}, \ldots, X_n)$.

4.6   Summary

Functions $Y = g(X)$ of random variables:

- X discrete, $P_Y(y) = \sum_{x \in \mathcal{X}_y} P_X(x)$ where $\mathcal{X}_y$ is the set of all x such that $g(x) = y$

- X continuous, g monotone increasing or decreasing with zero derivative in at most single points, $f_Y(y) = f_X\left(g^{-1}(y)\right) \frac{1}{|g'(g^{-1}(y))|}$

- for any X with mean μ and variance σ^2, $Y = (X - \mu)/\sigma$ has mean 0 and variance 1

- for any X with mean 0 and variance 1, $Y = aX + b$ has mean b and variance a^2.

Mean and variance of sums of random variables:

- $E[X + Y] = E[X] + E[Y]$

- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$, where $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$ is zero if X and Y are independent

- Correlation coefficient $\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$

Sums of discrete random variables:

- $P_{X+Y}(s) = \sum_x P_X(x)P_Y(s - x)$

- Probability generating function: $g_X(z) = \sum_x z^x P_X(x) = E[z^X]$

- $g_{X+Y}(z) = g_X(z)g_Y(z)$, $g_X^{(k)}(1) = E[X(X - 1) \cdots (X - k + 1)]$

- Sums of binomial random variables are binomial.

- Sums of Poisson random variables are Poisson.

Sums of continuous random variables:

- $f_{X+Y}(s) = f_X \star f_Y(s) = \int_{-\infty}^{\infty} f_X(x)f_y(s - x)\, dx$

- Moment generating function: $g_X(s) = \int_{-\infty}^{\infty} f_X(x)e^{sx}\, dx = E[e^{sX}]$

- $g_{X+Y}(s) = g_X(s)g_Y(s)$, $g^{(n)}(0) = E[X^n]$

- Sums of Gaussian random variables are Gaussian.

- Central Limit Theorem: the sum of n independent random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ tends to a Gaussian random variable with mean $\sum_i \mu_i$ and variance $\sum_i \sigma_i^2$ as n goes to infinity.

Characterising probability distributions and densities:

- Mean $E[X] = \mu$, variance $\mathrm{Var}(X) = \sigma^2$, skewness $E[(X - \mu)^3]/\sigma^3$

- Mode $\max_x P_X(x)$ or $\max_x f_X(x)$

- Median M and quartiles Q_1, Q_3, $F_X(M) = 1/2$, $F_X(Q_1) = 1/4$, $F_X(Q_3) = 3/4$, and inter-quartile range $Q_3 - Q_1$

Multi-variate Gaussians:

- $\mathbf{X} \sim \mathrm{N}(\mu, \mathbf{\Sigma})$, $f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}|\mathbf{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$

- Uncorrelated Gaussians are independent.

- Marginals and conditionals of multi-variate Gaussians are Gaussian.

Chapter 5

Decision, Estimation and Hypothesis Testing

So far in the course, we have introduced probability as a branch of mathematics dealing with uncertain events. In all the theory you've seen so far, you've learned how to quantify uncertainty, manipulate measures of uncertainty, but never how to resolve uncertainty into certainty. Indeed, in our example in Chapter 3 where a football captain needed the probability that a player would score a penalty, we learned that the probabilist's preferred approach is to compute a density for the unknown parameter p, rather than decide a specific value for it from the data.

There are occasions in practice when one has no choice but to make statements about uncertain events that don't involve densities or probability distributions, for example:

- A decision must be taken based on the data observed. For example, a patient has been tested for a condition and the doctor and patient must now *decide* whether to treat or not to treat for the condition based on the test result.

- An estimate is needed for an unknown quantity based on measured data. For example, we create a shear wave in a fluid using a vibrating probe and obtain sensor measurements of the dissipated energy, from which the viscosity of the fluid can be estimated.

- Public information: many statisticians and those funding them believe that the wider public is not able to understand probabilistic statements. This patronising view of public understanding of mathematics leads some to demand certain statements about uncertain measurements. For example, "The public wants to be told whether global warming is happening. They won't understand a probability density function of the predicted temperatures over the next decade." Recent U.S. presidents have made statements such as "Global warming is a fact!" or "The concept of global warming was created by and for the Chinese in order to make U.S. manufacturing non-competitive.". The role of statisticians is to prove or disprove the global warming hypothesis, depending on which political patrons they serve.

In this chapter, we will give a brief introduction to decision and estimation theory (there is much more to come in Part II modules such as 3F3) then proceed in more depth to discuss hypothesis testing.

5.1   Decision and Estimation theory

The model for both decision and estimation theory is represented in Figure 5.1. X and Y are two jointly distributed random variables. In decision theory, X is a discrete random variable and the role of the decision block $d(.)$ is to decide the value of X based on the observation Y, which may be discrete or continuous. In estimation theory, X is a continuous random variable and $d(.)$ aims
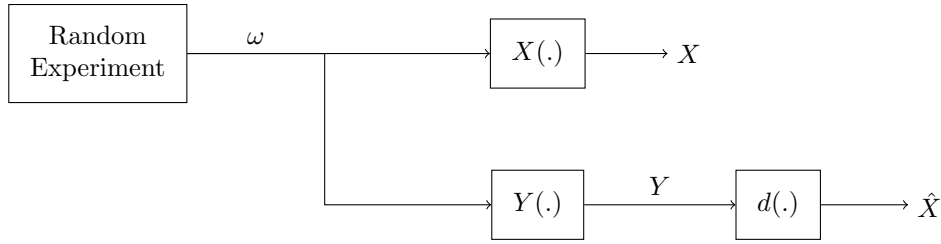
Figure 5.1: Model for decision and estimation theory.

to provide an estimate of X based on the observation Y. In all cases, the conditional probability distribution $P_{Y|X}(.|.)$ or density $f_{Y|X}(.|.)$ is known to the decider or estimator.

Example 5.1. Data bits are transmitted over a twisted copper cable to your games console via your home broadband modem. The data bits, valued zero or one, are transmitted as $+A$ Volt for a zero, and $-A$ Volt for a one. The received observations are independent given the data and Gaussian distributed with mean $+A$ or $-A$, i.e.,

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(y-(-1)^x A)^2/(2\sigma^2)} \text{ for } x \in \{0,1\}. \tag{5.1}$$

Hence, if the data bit is a Bernoulli random variable $P_X(1) = 1 - P_X(0) = p$, then

$$P_{X|Y}(1|y) = \frac{f_{Y|X}(y|1)p}{f_{Y|X}(y|0)(1-p) + f_{Y|X}(y|1)p} = \frac{1}{1 + \frac{f_{Y|X}(y|0)}{f_{Y|X}(y|1)}\frac{1-p}{p}}. \tag{5.2}$$

For a given observation y, it seems like a good idea to decide $\hat{X} = 1$ if $P_{X|Y}(1|y) > 1/2$ and to decide $\hat{X} = 0$ otherwise, or in other words

$$d(y) = \mathbf{1}\left(\frac{f_{Y|X}(y|0)}{f_{Y|X}(y|1)}\frac{1-p}{p} < 1\right) = \mathbf{1}\left(\frac{1-p}{p}e^{((y+1)^2-(y-1)^2)/(2\sigma^2)} < 1\right) \tag{5.3}$$

where the function $\mathbf{1}(.)$ is called an *indicator function* and returns a 1 if the condition in the brackets is true and a 0 otherwise. Note that if $P_X(1) = p = 1/2$, then the rule can be further simplified to

$$d(y) = \mathbf{1}\left((y+1)^2 - (y-1)^2 < 0\right) = \mathbf{1}(y < 0), \tag{5.4}$$

i.e., decide $\hat{X} = 1$ if the observation is negative, and $\hat{X} = 0$ if the observation is positive. If the observation is exactly zero, we could decide either way but since the probability of this event is zero it does not change the expected performance of our decision.

In this example, a decision for the value of bit X is necessary if your games console requires the data bits to keep your game going, e.g., it is not set up to receive probabilistic statements specifying the probability distribution of X given the observation. In fact, note that the "decision approach" to communications is overly simplistic. All of your communications devices nowadays are set up to receive probabilistic statements about data bits from the transmission line, as this allows for a much better data encoding and decoding to maximise the data transmission rate.

The example above used the so-called Maximum A-Posteriori (MAP) rule:

MAP rule: for an observation $Y = y$, pick $\hat{X} = x$ to maximise $P_{X|Y}(x|y)$.

The rule for $p = 1/2$ is an instance of the Maximum Likelihood (ML) rule:

ML rule: for an observation $Y = y$, pick $\hat{X} = x$ to maximise $P_{Y|X}(y|x)$ (or $f_{Y|X}(y|x)$ for continuous observations.)

The ML rule is equivalent to the MAP rule when X is uniform, but is often also used in cases where the prior distribution P_X is unknown to the decider.

In estimation problems, X is continuous and we cannot hope to reconstruct X exactly. Hence, our aim in general is to find a $\hat{X}$ that approximates X as closely as possible given the observation Y, and one must define what is meant by "as closely as possible." One common way to define this closeness is to minimise the Mean Squared Error (MSE) $E[(\hat{X} - X)^2|Y = y]$, where the conditional expectation is defined just like the expectation but using the conditional probability density function. Such an estimator is called the Minimum Mean Squared Error (MMSE) estimator. Let $m_y = E[X|Y = y]$, i.e., the expected value of X given the observation y, and note that

$$E[(\hat{X} - X)^2|Y = y] = E[(X - \hat{X})^2|Y = y] = E[(X - m_y + m_y - \hat{X})^2|Y = y] \tag{5.5}$$

$$= E[(X - m_y)^2 + (\hat{X} - m_y)^2 - 2(X - m_y)(\hat{X} - m_y)|Y = y] \tag{5.6}$$

$$= E[(X - m_y)^2|Y = y] + (\hat{X} - m_y)^2 - 2(\hat{X} - m_y)E[X - m_y|Y = y] \tag{5.7}$$

$$= \text{Var}[X|Y = y] + (\hat{X} - m_y)^2 - 2(\hat{X} - m_y)(E[X|Y = y] - m_y) \tag{5.8}$$

$$= \text{Var}[X|Y = y] + (\hat{X} - m_y)^2 \geq \text{Var}[X|Y = y] \tag{5.9}$$

with equality if and only if $\hat{X} = m_y = E[X|Y = y]$. Hence the mean squared error is minimised by picking the conditional expectation of X given the observation as our estimate, giving the following rule:

MMSE estimator: $\hat{X} = E[X|Y = y]$.

While we have used a notation implying that the observation Y is a single random variable, in practice Y is often a vector of random variables. We have only scratched the surface of decision and estimation theory in this section. Many extensions of these theories are possible. For example, decision theory can be extended to take a cost associated with every decision into account. Estimation theory can be restricted to linear estimators, i.e., what is the best estimator for X given a vector $\underline{Y}$ when you are only allowed to use estimators of the form $\hat{X} = \underline{c}.\underline{Y}$ for a given vector $\underline{c}$?

5.2 Hypothesis testing: simple hypotheses

Hypothesis testing is a branch of classical statistics that establishes rules for making certain statements about uncertain events, sometimes qualifying them with a soft "*p*-value". The use of hypothesis testing is by no means uncontroversial and students are strongly encouraged to read Chapter 37 in MacKay's book [Mac03] on this topic. However, hypothesis testing is the standard technique used in many areas of science, government and engineering. To date, it is impossible to publish research in most medical or natural science journals with claims not validated using classical stastistics and hypothesis testing with *p*-values. There are voices even within the medical world [Goo99] that argue against this practice. Hypothesis testing is included in this course because you need to be aware of it, even though the probability theory that you have learned so far is sufficient for you to make probabilistic statements about hypotheses that are often more accurate than those described in this section and the next.

For an observed random variable Y, a *simple* hypothesis H is one for which the probabilities

$$p(Y = y \,|\, H) \text{ and } p(Y = y \,|\, \bar{H})$$

are well defined, where $\bar{H}$ is the complement of H. H is often called the *null hypothesis* $H_0 = H$, and $\bar{H}$ the *alternative hypothesis* $H_1 = \bar{H}$.

Example 5.2. A hypothetical researcher has discovered a simple pregnancy test based on a retina scan that can be administered easily and inexpensively by every optician. The test gives a numerical outcome that is Gaussian $N(1, 1/4)$ for a pregnant test person, and Gaussian $N(0, 1)$ for a test person who is not pregnant.

In the example above, the null hypothesis H_0 "the test person is pregnant" is a simple hypothesis. It is a well defined event that has a probability and for which there is a conditional probability of the observation. Its converse $H_1 = \bar{H}$ "the test person is not pregnant" is equally well defined.

The outcome of a hypothesis test is a statement concluding either

$$H_0 \text{ is true (i.e., } H_1 \text{ is false)}$$

or

$$H_0 \text{ is false (i.e., } H_1 \text{ is true),}$$

possibly with a numerical p-value indicating the strength of the statement. Let the random variable X be an indicator random variable for our statement, i.e., $X = 1$ if we claim that H_0 is true, and $X = 0$ if we claim that H_0 is false. It is useful to distinguish between the types of error that we can make in our statement:

X \ H_0	false	true
0	$\checkmark$	type I
1	type II	$\checkmark$

This table highlights the asymmetry of the hypothesis testing problem. There may be different consequences to a type I and a type II error. In the example above, if we tell the test person that she's likely to be pregnant, we will recommend that she attend a GP for a more accurate test. If we tell the test person that she's not pregnant, there will be no follow-up action. The damage done by a type I error (undetected pregnancy) is a lot worse than the cost of the type II error (unnecessary visit to a GP.)

Example (continued): One simple strategy for the developer of the iris scan pregnancy test is to set a threshold at a value β and pick $X = 1$ (H_0 is true) if $Y > \beta$ and $X = 0$ (H_0 is false) if $Y \leq \beta$. The type I error probability is defined as

$$\varepsilon_{\mathrm{I}} = p(X = 0 \,|\, H_0) = p(Y < \beta \,|\, H_0),$$

i.e., the probability that the data lie in a region for which we will state "H_1 is true" given that H_0 is true. We know that the density of Y given H_0 is Gaussian $N(1, 1/4)$. Hence, supposing that the Medicines and Healthcare products Regulatory Agency (MHRA) requires a Type I error probability of 0.01 for this type of test, we have

$$p(Y < \beta \,|\, H_0) = p(2(Y - 1) < 2(\beta - 1) \,|\, H_0) = p(2(Y - 1) > 2(1 - \beta) \,|\, H_0)$$

where $2(Y - 1)$ is standard Gaussian $N(0, 1)$ (see Section 4.1.2) so we can read out the value z for which $\Phi(z) = 1 - 0.01 \simeq 0.99$ on page 29 of your Mathematics Data book to obtain

$$2(1 - \beta) = 2.33$$

and hence $\beta = -0.165$. Hence, the test will return a result of "pregnant" or "probably pregnant" for $Y > -0.165$, and "not pregnant" for $Y < -0.165$.
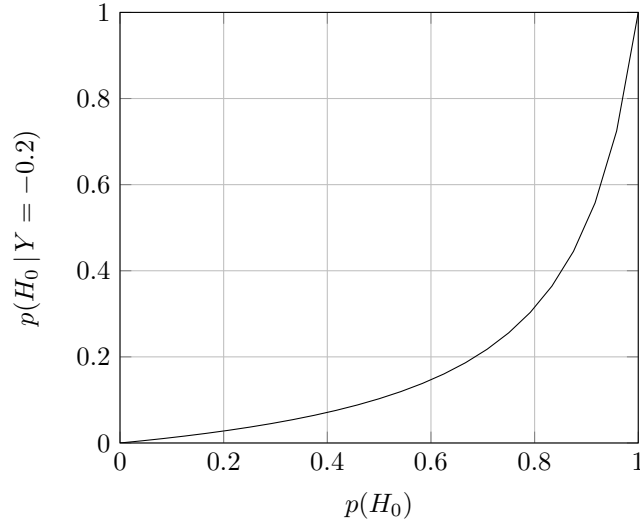
Figure 5.2: Probability of a hypothesis given the test result in function of the a-priori probability

There is considerable confusion about the meaning of the numerical value associated with a hypothesis testing statement. In the example above, we required a "not pregnant" result with an upper bound on the type I error probability of 1%. Most test persons not versed in hypothesis testing would assume that this means that if they get a "not pregnant" outcome, the probability that they are not pregnant is 99%. *This is wrong!* The correct interpretation of this figure is that the result of the test fell within a region whose total probability is less than 1% for a pregnant test person.

Example 5.3. Consider the example of a test person whose numerical test result was -0.2 and was given a "not pregnant" outcome. We would now like to depart from the framework of Hypothesis testing and compute the probability that the person is pregnant given the test result. In order to use Bayes' theorem we need the prior probability $p = p(H_0)$ that a person taking the test is pregnant. We can write

$$p(H_0|Y = -0.2) = \frac{f_Y(-0.2\,|\,H_0)dy\,p(H_0)}{f_Y(-0.2\,|\,H_0)dy\,p(H_0) + f_Y(-0.2\,|\,H_1)dy\,p(H_1)}$$

$$= \frac{p\frac{1}{\sqrt{\pi/2}}e^{-(-1.2)^2 \times 2}}{p\frac{1}{\sqrt{\pi/2}}e^{-(-1.2)^2 \times 2} + (1-p)\frac{1}{\sqrt{2\pi}}e^{-(-0.2)^2/2}}$$

$$= \frac{p \times 0.1123}{p \times 0.1123 + (1-p) \times 0.9802}$$

This probability is plotted in Figure 5.2 and we observe that the probability that the test person with a numerical result of -0.2 is pregnant varies widely with the a-priori probability $p = p(H_0)$. It is easy to calculate that the test person is in fact only 99% sure of not being pregnant if the a-priori probability is 8.1%.

A central theorem of hypothesis testing known as the Neyman-Pearson theorem shows that the type of threshold test we applied in our example is in fact optimal in terms of minimising the Type I probability for a given Type II probability.

5.3   Hypothesis testing: composite hypotheses

The second type of hypotheses that come into play in hypothesis testing are composite hypotheses[1], or hypotheses for which it is not easy or impossible to express a probability distribution of the data. These are often complements of a simple null hypothesis. Consider the following example.

Example 5.4. A hypothetical cafe "HN" has recently opened its doors opposite the department of Engineering. Its hypothetical owner "S" wonders whether customers know the difference between a "flat white" and a "cappucino". He proposes to use a double-blind statistical experiment in which customers are given a drink and asked to guess what it is, and has called upon the probability experts at the department to help him evaluate the test.

Of the 10 customers who participated in the test so far, 7 correctly identified their drink and 3 did not. We would like to test the following hypotheses:

- H_0: the data is "random"

- H_1: the data is "not random"

H_0 can be made more precise, i.e., the customers' answers were independent and $\mathrm{Ber}(1/2)$ (independent of each other and of the drinks tasted.)

In the example, H_0 is a simple hypothesis: given that the decisions are $\mathrm{Ber}(1/2)$ independent of the drinks, it is possible to evaluate the probability of the data observed. The complement hypothesis H_1 is a composite hypothesis: what does it mean for the data to be "not random"? Do we mean that it is Bernoulli with a different parameter than $1/2$? Or not Bernoulli at all? Or perhaps the customers' decisions were not independent and we must consider their joint distribution? Since the hypothesis is not well defined, it is hard or impossible to talk about the probability of the data given the hypothesis. The approach described in the previous section is no longer applicable.

In this case, classical statistics resorts to one-sided statistical tests, simply evaluating the probability of the data given H_0 and deciding that H_0 is true if the data falls within a pre-determined confidence range. Hypothesis H_1 is ignored, but the statement "H_0 is false" is nonetheless often translated into "H_1 is true" despite the fact that H_1 is not well defined. We continue our example.

Example 5.5. In the "cappucino" vs. "flat white" battle, we decide to admit the null hypothesis H_0 if the probability of the data observed *or any more extreme data* is more than 5% (a very common threshold for statistical significance.) The sum of Bernoulli random variables is binomial distributed, and hence

$$p(Y \geq 7 \,|\, H_0) = \binom{10}{7}\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^3 + \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1 + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0$$

$$= \frac{1}{2^{10}}\left(\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right) = 0.1719$$

This is far more than 5% and hence we state that H_0 is true, or "the data is random". In other words, "customers are unable to distinguish between a cappucino and a flat white."

Note that the size of the experiment has a bearing on our decision. If we increase the number of test persons involved but assume that the proportion of those correctly identifying their drinks remains constant, we get for a 20 customer experiment $p(Y \geq 14) = 0.0577$

---

[1]The appelation "simple hypothesis" vs. "composite hypothesis" is not widespread and is borrowed from [RHB06].

so much closer to the threshold but still favouring H_0, and for a 30 customer experiment $p(Y \geq 21) = 0.0214$ so we now decide that H_0 is false. The conclusion "the data is not random" or "H_1 is true" is risky since we have not clearly decided what we mean by H_1 in this context, but most would now conclude "customers can distinguish between a flat white and a cappucino".

Note that we could use the Beta distribution approach introduced in Section 3.5 to make statements about H_1 if we simply assume that "not random" means independent Bernoulli with an unknown parameter and hence compare H_1 against H_0 using Bayesian methods.

Note in the example that we use the probability of the data observed *or more extreme data* rather than just the probability of the data. It is in the nature of statistical experiment that the probability $p(Y = y \mid H_0)$ of the data will decrease with the size of the experiment. For example, if our experiment always yielded half correct guesses and half incorrect guesses (assuming an even experiment size), the probability of the observed data would be

$$p(Y = y \mid H_0) = \binom{n}{n/2} \frac{1}{2^n}$$

where n is the number of customers in the measurement. This is the mode (largest value) of the binomial distribution, but nonetheless decreases[2] with n. Hence it would appear that the probability of the data being random decreases with the size of the experiment, and we would reject the null hypothesis in all cases if the experiment size grew sufficiently large. This is the reason why statisticians insist on measuring the probability of the measured data *or more extreme data*. In the extreme example just presented, that value would be $1/2$ for all experiment sizes and hence always accepts the null hypothesis.

Example 5.6. A bus company claims that on a certain route there is a service every 20 minutes. Three people complain that this claim is false:

- A: had to wait 45 minutes on a particular day

- B: had to wait 45 minutes on both Monday and Tuesday

- C: had to wait 45 minutes on two days of last week

Are any of these claims statistically significant?

Null hypothesis: bus arrivals are random, Poisson, with intensity $\lambda = 3$ buses per hour. Thus, waiting times are exponentially distributed $\text{Ex}(\lambda = 3)$.

$$p(\text{wait} > 3/4) = \int_{3/4}^{\infty} 3 \exp(-3t)dt = \Big[-\exp(-3t) \Big]_{3/4}^{\infty}$$
$$= \exp(-9/4) \simeq 0.105$$

- A: had to wait 45 mins on a particular day. Since $p(A \mid H_0) = 0.105$ this is not hugely unlikely, and cannot be used to reject H_0 at a 5% level.

- B: had to wait 45 mins both Monday and Tuesday. Both events happen independently, so $p(B \mid H_0) = 0.105^2 = 0.011$. This seems quite unlikely under the null hypothesis, so we can reject the companies claim, say at a 5% level.

---

[2]As an exercise, verify that it decreases by considering two consecutive even test sizes n and $n + 2$ and showing that

$$\binom{n+2}{(n+2)/2} 2^{-(n+2)} - \binom{n}{n/2} 2^{-n} < 0.$$

- C: had to wait 45 mins on two days of last week.

$$p(C \mid H_0) \;=\; \sum_{k=2}^{5} \binom{5}{k} 0.105^k (0.895)^{5-k} \;\simeq\; 0.089,$$

again, not sufficiently strong evidence.

This example shows how sensitive the methods of composite hypothesis testing are to the exact wording of the hypothesis. A court statistician in cases involving customers A, B and C against the bus company would award damages to customer B and reject the claims of customers A and C even though common sense does not detect significant differences between these customers' experiences.

The final conclusion of this chapter is that one has to be extremely careful when interpreting statements obtained using hypothesis testing. The questions to ask are:

- what is the null hypothesis?

- what is the alternative hypothesis?

- what was the criterion used for accepting or rejecting the null hypothesis.

Statements should ideally be carefully worded to describe exactly what has been calculated, although if this is followed to the letter the statements would become so unintelligible as to little no sense to most people. Perhaps this is a good place to insist once again that computing the probability of the hypothesis given the data using Bayes' theorem results in much clearer statements that may not be pleasant to hear but are in fact informative and accurate, e.g., the probability that you have cancer given the test result is 85%.

5.4   Summary

Decision and estimation:

- Maximum A-Posteriori (MAP) decision rule: for an observation $Y = y$, pick $\hat{X} = x$ to maximise $P_{X|Y}(x|y)$.

- Maximum Likelihood (ML) decision rule: for an observation $Y = y$, pick $\hat{X} = x$ to maximise $P_{Y|X}(y|x)$ (or $f_{Y|X}(y|x)$ for continuous observations.)

- Minimum Mean Square Estimator (MMSE): $\hat{X} = E[X|Y = y]$.

Hypothesis testing:

- Error types for simple hypotheses:

Test outcome \\ H_0	false	true
H_0 is false	$\checkmark$	type I
H_0 is true	type II	$\checkmark$

- For a specified worst allowable probability of Type I error, the optimal minimising the probablity of a Type II error is a threshold-type test (Neyman-Pearson)

- For composite hypotheses (complementary hypothesis H_1 not well defined), always set a threshold to admit the null hypothesis H_0 based on the probability ("p-value") that the data is as observed *or more extreme*

Bibliography

[Bil95]   Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, 1995.

[FLS64]   Richard P Feynman, Robert B Leighton, and Matthew Sands. *The Feynman lectures on physics*. Addison-Wesley Publishing Company, 1964.

[Goo99]   Steven N Goodman. Towards evidence-based medical statistics. The P value fallacy. *Annals of Internal Medicine*, 130:995–1004, June 1999.

[Mac03]   David J C MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.

[RHB06]   K F Riley, M P Hobson, and S J Bence. *Mathematical methods for physics and engineering*. Cambridge University Press, 2006.

[Ros72]   Sheldon M Ross. *Introduction to Probability Models*. Academic Press, 1972.