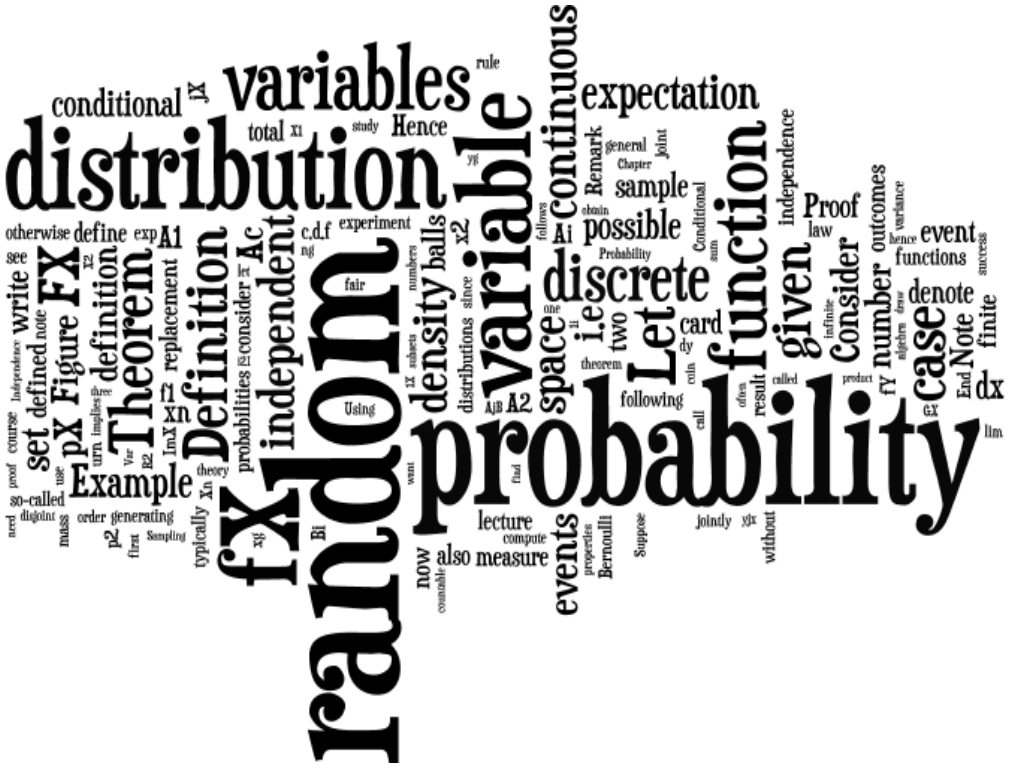
MATH40005: Probability and Statistics

Almut E. D. Veraart
Room 551, Huxley Building
Department of Mathematics
Imperial College London
180 Queen's Gate, London, SW7 2AZ
E-Mail: a.veraart@imperial.ac.uk

Autumn 2019



(Last Update: Thursday 24th October, 2019)

Contents

Chapter 1

Introduction

1.1 Why do we study probability?

Probability

- is a beautiful branch of mathematics with a long history going back to the early works by Cardano (16th century), Fermat and Pascal (17th century), Laplace (19th century). Modern (axiomatic) probability theory, however, is a much younger discipline which goes back to the influential work by Kolmogorov published in 1933,
- is a very dynamic discipline with a strong interplay between theory and applications,
- is ubiquitous in every day life and in most sciences,
- is the foundation for statistics,
- enables us to interpret and quantify uncertainty.

1.2 Complementary reading

- These lecture notes are self contained. They are mainly based on the textbooks Grimmett & Welsh (1986), Blitzstein & Hwang (2019) and Anderson et al. (2018).

  You can get the first and third book from our library and the second book is available on-line at https://projects.iq.harvard.edu/stat110/about, where you can also find additional exercises and solutions.
- Complementary reading material can be found in the following textbooks: Ross (2014).

1.3 Course overview (Autumn term)

1. Interpretations of probability; limiting frequency; classical (symmetry between equally likely outcomes) ; subjective (degree of personal belief)
2. Counting: multiplication principle; binomial coefficients; the inclusion-exclusion principle; stars and bars arguments
3. Formal probability: probability axioms; conditional probability; Bayes theorem; independence
4. Random variables: mass and density functions; common discrete and continuous distributions, transformations of random variables, expectation and variance; probability and moment generating functions

5. Multivariate random variables: Joint mass and density functions; independence; covariance
6. Conditional distribution: Conditional probability mass function, conditional density, conditional expectation, law of total expectation

1.4 Expectations and feedback

In order to successfully complete the course and build up a solid foundation in probability theory for term 2 and the subsequent years, please actively engage with the course and the material:

- You are strongly encouraged to ask questions in person (and not by email!) before/during/after the lectures, problem classes, tutorials and during my office hour.
- Take notes during the lectures.
- Try to answer the questions on the problem sheet yourself and also work in pairs or small groups to revise the material.
- Ideally you should read the material to be covered in each lecture before and after the relevant lecture.
- Test your knowledge and understanding regularly. For instance, write down a short summary after each lecture without consulting the lecture notes.
- Use Panopto wisely: Attend the lectures in person and possibly review sections of the lectures which you found particularly difficult by watching the relevant section in the video.
- If you have any concerns regarding the course, please approach me directly or speak to the year group representatives. This course is for you and your opinion matters!

1.4.1 Feedback

You will obtain feedback

- through one-to-one interactions with me before/after lectures or during my office hours.
- Through quizzes and homework assignments.
- Through the mid-term and January tests.

1.4.2 Assessment

The assessment plan for the two-term course MATH40005 is as follows:

- Portfolio element (10% of the course mark):
  - Term 1: Three Blackboard quizzes worth 1% each,
  - Term 2:
    - Two Blackboard quizzes worth 1% each,
    - One coursework worth 5%,
- Midterm exam in term 1 worth 5% (two-stage test (TBC))
- January test on the material from term 1 worth 10%,
- Midterm exam in term 1 worth 5% (two-stage test (TBC))
- Three-hour summer exam worth 70%.

Chapter 2

Sample spaces and interpretations of probability

The material of this chapter is based on Blitzstein & Hwang (2019), p.1-8, Anderson et al. (2018), p.1-5, Proschan & Shaw (2016), p.9-10.

2.1 Notation

Throughout the lecture notes we denote the natural numbers by $\mathbb{N} = \{1, 2, \dots\}$, the integers by $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$, the real numbers by $\mathbb{R}$. For real numbers $a < b$ we write $[a, b]$ for closed intervals and (a, b) for open intervals.

2.2 The sample space Ω

Probability theory is based on *set theory* which was introduced in the course *Introduction to University Mathematics*. We will now explain how set theory enters in probability theory and review some of the key concepts briefly.

Definition 2.2.1 (Sample space)**.** *The* sample space Ω *is defined as the set of all possible outcomes of an experiment. The elements of* Ω *are typically denoted by* ω *and called* sample points.

Example 2.2.2. *We start with the classical example of flipping a fair coin. We write* H *for heads and* T *for tail. The sample space is given by* $\Omega = \{H, T\}$.

Figure 2.1: Flipping a fair coin.

Example 2.2.3. *Consider the experiment where we roll a standard six-sided fair die. The sample space associated with this experiment is given by* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definition 2.2.4 (Cardinality)**.** *For any set A, we define the* cardinality *of A as the number of elements in A. We typically write $\mathrm{card}(A)$ or simply $|A|$ (the latter should not be confused with the absolute value!).*

Example 2.2.5. *The cardinality of the sample spaces Ω considered in the two examples above are 2 (flipping of a fair coin) and 6 (rolling a fair die), respectively.*

We note that the sample space of an experiment can be finite (e.g. $\Omega = \{1, \dots, 6\}$, countably infinite (e.g. $\Omega = \mathbb{N} = \{1, 2, \dots\}$), or uncountably infinite (e.g. $\Omega = [0, 1]$).

Definition 2.2.6 (Finite, countably infinite and uncountably infinite sets)**.** *A set A is said to be* finite*, if it has a finite number of elements. A set A is called* countably infinite *if there is a 1-1 correspondence (also called a bijection*[1]*) between the elements of A and the natural numbers $\mathbb{N} = \{1, 2, \dots\}$. If the set A is neither finite nor countably infinite, we call it* uncountably infinite*.*

Let's recap some concepts from set theory:

- Subsets of Ω are collections of elements of Ω and called *events*. Notation: A is a subset of Ω can be written as $A \subseteq \Omega$ meaning that every element of A is also an element of Ω.
- We write $\omega \in A$ if the element ω is a member of A and $\omega \notin A$ if the element ω is not a member of A.
- We denote the empty set by $\emptyset$. Note that the empty set contains no points, i.e. $\omega \notin \emptyset$ for all $\omega \in \Omega$.
- Every subset A of the sample space Ω satisfies $\emptyset \subseteq A \subseteq \Omega$.

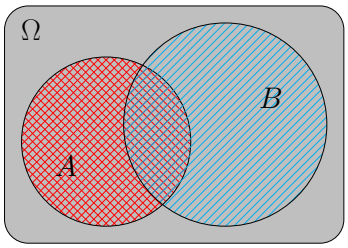
Suppose that $A, B \subseteq \Omega$ are events, then

- the union $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the event that at least one of A and B occurs (this is the *inclusive* "or"),
- the intersection $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ is the event that both A and B occur,
- the complement $A^c = \Omega \setminus A = \{\omega \in \Omega : \omega \notin A\}$ is the event that occurs if and only if A does not occur.

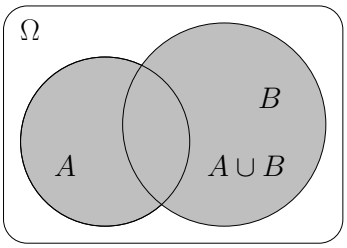
2.2.1 Venn diagrams

We typically use so-called *Venn diagrams* to illustrate concepts from set theory such as the union, intersection and complement of sets introduced above. Consider a sample space Ω with subsets $A, B \subseteq \Omega$.

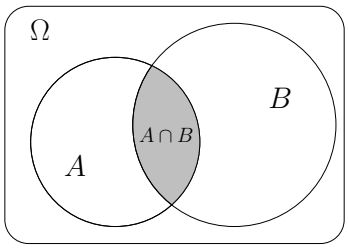
[1] This definition will be covered in the Analysis course.



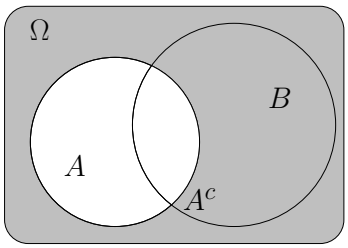
(a) $A, B \subseteq \Omega$



(b) $A \cup B$



(c) $A \cap B$



(d) A^c

Figure 2.2: We consider a sample space Ω with subsets $A, B \subseteq \Omega$ which are depicted in Figure 2.2a. The grey area in Figure 2.2b depicts the union $A \cup B$. The grey area in Figure 2.2c depicts the intersection $A \cup B$, and the grey area in Figure 2.2d depicts the complement A^c.

End of lecture 1.

2.3 Interpretations of probability

Let us briefly discuss the three main interpretations of probability, for an extended survey please see Hájek (2012).

2.3.1 Naive definition of probability - classical interpretation

Consider the case when the sample space Ω is finite, i.e. $\text{card}(\Omega)$ and suppose you want to assign a probability to the event $A \subseteq \Omega$. A naive definition of a probability is obtained when we count the elements in A and divide by the total number of elements in Ω:

Definition 2.3.1 (Naive definition of probability)**.** *Suppose that the sample space Ω is finite, i.e. $\text{card}(\Omega) < \infty$ and consider an event $A \subseteq \Omega$. Then the* naive probability *of A is defined as*

$$\text{P}_{\text{Naive}}(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}.$$

In addition to the assumption that the sample space is finite, the naive definition of a probability also assumes that each possible outcome has the same weight. When is such a definition applicable? In *symmetric settings* when all outcomes are equally likely (for instance when we toss a fair coin or roll a fair die), or in settings where the outcomes are equally likely due to the *design* of a study (for instance when I randomly select 10 students out of the entire year group assuming that the selection mechanism is such that all subsets of 10 students are equally likely).

Example 2.3.2. *Consider the example of rolling a six-sided fair die. What is the (naive) probability that I roll either a 1 or a 2? We have* $\Omega = \{1, \ldots, 6\}$ *and* $A = \{1, 2\}$. *Hence*

$$\mathrm{P}_{\mathrm{Naive}}(A) = \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)} = \frac{2}{6} = \frac{1}{3}.$$

Let us consider the complement: $A^c = \{3, 4, 5, 6\}$, *which can be computed as*

$$\mathrm{P}_{\mathrm{Naive}}(A^c) = \frac{\mathrm{card}(A^c)}{\mathrm{card}(\Omega)} = \frac{\mathrm{card}(\Omega) - \mathrm{card}(A)}{\mathrm{card}(\Omega)} = 1 - \frac{\mathrm{card}(A)}{\mathrm{card}(\Omega)} = 1 - \mathrm{P}_{\mathrm{Naive}}(A) = \frac{2}{3}.$$

Note that for $A \subseteq \Omega$ we always have that $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$, not just in the case of the naive probability.

The classical interpretation applies when we have outcomes that are equally likely. In our naive definition above, we have only covered the case when $\mathrm{card}(\Omega) < \infty$. If Ω is uncountably infinite, but of finite area, e.g. choose a disk of radius 1: $\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$ and the event A is some subset of Ω, then we could assume that the probability of the event A should be uniform on Ω, i.e.

$$\mathrm{P}(A) = \frac{\text{area of } A}{\text{area of } \Omega}.$$

For instance, for $A = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 0.5^2\}$, we have

$$\mathrm{P}(A) = \frac{\text{area of } A}{\text{area of } \Omega} = \frac{0.5^2\pi}{\pi} = 0.25.$$

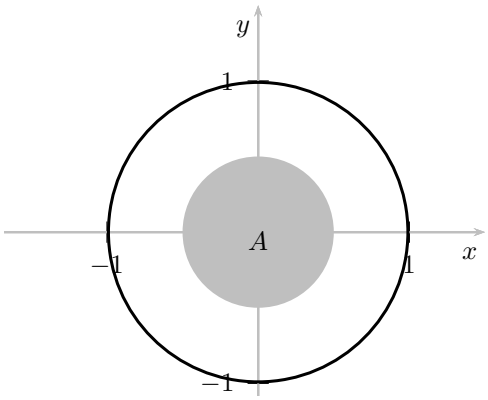


Figure 2.3: Illustration of the classical interpretation of probability in the case when Ω is uncountably infinite.

Remark 2.3.3. *In order for the classical/naive definition to work, we need that the number of elements in* Ω *is either finite, or, if* Ω *is uncountably infinite, then we require that the area of* Ω *is finite. In either scenario we can then define a uniform distribution, which we will study in more detail later in the course. Note that there is no uniform distribution on* $\mathbb{N}$ *or on* $\mathbb{R}$.

2.3.2 Limiting frequency

Consider n_{total} replications of an experiment and let n_A denote the number of times event A occurs (out of n_{total}). Then we could interpret the probability of event A occuring as

$$\mathrm{P}(A) = \lim_{n_{\mathrm{total}} \to \infty} \frac{n_A}{n_{\mathrm{total}}}.$$

The problem with this interpretation is that $n_{\mathrm{total}} \to \infty$ may be difficult to conceive, and any finite version may not be representative.

Example 2.3.4. *Consider a fair coin toss and let $A = \{H\}$ denote the event that heads appears. Figure[2] 2.4 illustrates a possible outcome of the experiment.*

Figure 2.4: One possible outcome when tossing a fair coin repeatedly.

Let us compute the relative frequency of A and report and plot them in Table 2.1 and Figure 2.5, respectively.

n_{total}	1	2	3	4	5	6	7	8	9	10	$\cdots$
$\frac{n_A}{n_{\text{total}}}$	0/1	0/2	1/3	2/4	2/5	3/6	4/7	5/8	6/9	6/10	$\cdots$

Table 2.1: Relative frequencies of heads when repeatedly tossing a fair coin.

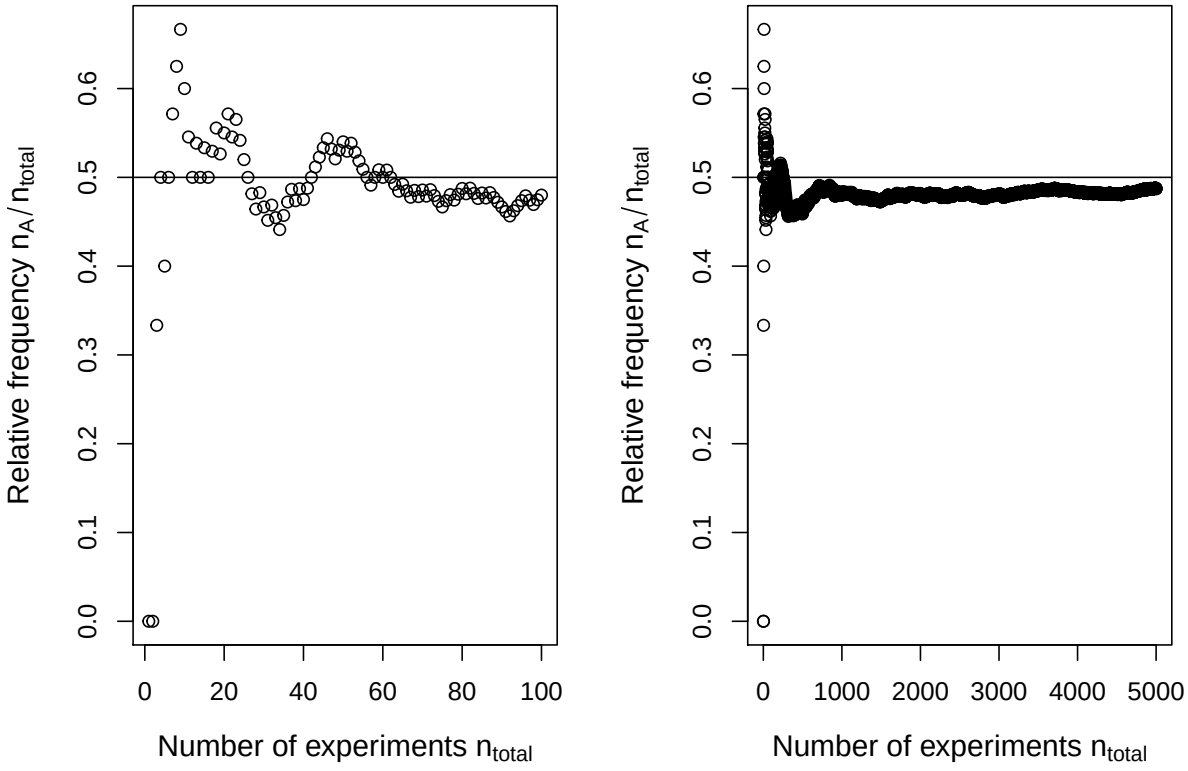


Figure 2.5: Relative frequencies of heads when repeatedly tossing a fair coin.

2.3.3 Subjective

For an event A, we can assign the probability $\mathrm{P}(A)$ according to our personal "degree of belief". This could be done according to historical information or local knowledge. This probability will not need to be

[2]The pictures of the one pound coin are attributed to Sir Magnus Fluffbrains [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)] https://commons.wikimedia.org/wiki/File:Pound_coin_front.png and https://commons.wikimedia.org/wiki/File:Pound_coin_back.png.

the same for each individual. The subjective approach may be difficult to implement in practice, but is a valid and universal interpretation of probability.

Remark 2.3.5. *It is important to remember that all three interpretations of probability depend on assumptions about experimental conditions.*

Chapter 3

Counting

The material of this chapter is based on Blitzstein & Hwang (2019), p.8-28, Anderson et al. (2018), p.4-11.

In order to compute (naive) probabilities we need to be able to count events in possibly large (but finite) sample spaces. The area of mathematics which deals with counting is called *Combinatorics*. We will study some key ideas from combinatorics and show their interplay with probability theory.

3.1 The multiplication principle

Theorem 3.1.1 (The multiplication principle)**.** *Consider two experiments: Experiment A has a possible outcomes, and Experiment B has b possible outcomes. Then the compound experiment of performing Experiment A and B (in any order) has ab possible outcomes.*

Proof. Without loss of generality we assume that we conduct Experiment A first. We draw a tree diagram consisting of a branches with one branch for each possible outcome of Experiment A. For each of theses branches we then generate b branches for each possible outcome of Experiment B. We can then directly read off that there are $\underbrace{b + \cdots + b}_{a} = ab$ possibilities. □

We illustrate the proof of the multiplication principle in Figure 3.1 in the case when the outcomes of Experiment A are labelled as A_1, A_2 (for $a = 2$) and the outcomes of Experiment B are labelled as B_1, B_2, B_3 (for $b = 3$).

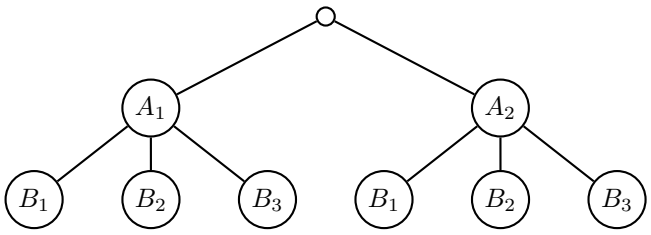


Figure 3.1: Illustration of the proof of the multiplication principle in the case when the outcomes of Experiment A are labelled as A_1, A_2 (for $a = 2$) and the outcomes of Experiment B are labelled as B_1, B_2, B_3 (for $b = 3$).

End of lecture 2. .....

3.2 Power sets

Exercise 3.2.1. *Consider a set Ω with $\text{card}(\Omega) = n \in \mathbb{N}$ elements. Use the multiplication principle to show that there are 2^n possible subsets of Ω if you include the empty set $\emptyset$ and Ω.*

Solution: For each element in Ω, you can choose whether or not to include it in the subset, so you have two options each. Hence you have $\underbrace{2 \cdots 2}_{n} = 2^n$ possible outcomes.

Definition 3.2.2 (Power set)**.** *A* power set *of a set A, denoted as $\mathcal{P}(A)$ is defined as the set of all possible subsets of A including $\emptyset$ and A.*

We have already proven the following result:

Theorem 3.2.3 (Cardinality of the power set)**.** *Consider a sample space Ω with $\text{card}(\Omega) < \infty$. Then $\text{card}(\mathcal{P}(\Omega)) = 2^{\text{card}(\Omega)}$.*

Example 3.2.4. *Let $\Omega = \{A, B, C\}$. We want to find the corresponding power set $\mathcal{P}(\Omega)$, which consists of $2^3 = 8$ elements. We have*

$$\mathcal{P}(\Omega) = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \Omega\}$$

3.3 Sampling with and without replacement

We can use the multiplication principle to derive the number of outcomes when sampling with or without replacement. In the following, we will state the results in the context when we have an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, 2, \ldots, n\}$ and we draw $k \in \mathbb{N}$ balls from the urn.

Figure 3.2: Consider an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, \ldots, n\}$. How many possible ways are there to draw $k \in \mathbb{N}$ balls with or without replacement?

3.3.1 Sampling with replacement – ordered

Consider an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, \ldots, n\}$. Let $k \in \mathbb{N}$. Suppose that you take a ball out of the urn and write down its number. Then you put it back into the urn (i.e. you replace it). You do this k times in total and write down the labels in the order in which they appear. The sample space Ω of this experiment can be expressed in the following way: We denote by $S = \{1, \ldots, n\}$ the labels of the balls in the urn. A possible outcome of the experiment can be written as $\omega = (s_1, \ldots, s_k)$, where s_i denotes the number of the ith ball for $i \in \{1, \ldots, n\}$. Hence

$$\Omega = \underbrace{S \times \cdots \times S}_{k \text{ times}} = S^k = \{(s_1, \ldots, s_k) : s_i \in S \text{ for } i = 1, \ldots, k\}.$$

Theorem 3.3.1 (Sampling with replacement)**.** *In the case of sampling k balls with replacement from an urn containing n balls as described above, there are $\mathrm{card}(\Omega) = n^k$ possible outcomes when the order of the objects matters.*

Proof. The result is a direct consequence of the multiplication principle: Each time we draw a ball, there are n possible outcomes. We carry out this experiment k times, so there are n^k ways of obtaining a sample consisting of k balls. □

3.3.2 Sampling without replacement – ordered

Again we consider an urn with $n \in \mathbb{N}$ balls which are labelled $\{1, \dots, n\}$. Let $k \in \mathbb{N}$. Suppose that you take a ball out of the urn and write down its number. Then you remove the ball and do not put it back into the urn (i.e. you do not replace it). You do this k times in total. Since you are removing balls from the urn permanently, k cannot be larger than n.

The sample space Ω of this experiment can be expressed in the following way: We denote by $S = \{1, \dots, n\}$ the labels of the balls in the urn. Then

$$\Omega = \{(s_1, \dots, s_k) : s_i \in S \text{ for } i = 1, \dots, k, \text{ and } s_i \neq s_j \text{ if } i \neq j\}.$$

Theorem 3.3.2 (Sampling without replacement)**.** *In the case of sampling k balls without replacement from an urn containing n balls as described above, there are $\mathrm{card}(\Omega) = n(n-1)\cdots(n-(k-1)) = (n)_k$ possible outcomes when the order of the objects matters.*

Note that we use the convention that $(n)_1 = n$.

Proof. Also this result is a direct consequence of the multiplication principle: The first time, we draw a ball, there are n possible outcomes, the second time, there are $n-1$ possible outcomes and so on, and when we draw the kth ball, there are $n-(k-1)$ possible labels left for the kth ball. Multiplying the number of possible outcomes for each sub-experiment together leads to the stated result. □

Definition 3.3.3 (Factorial)**.** *Let $n \in \mathbb{N}$. The* factorial *of n, denoted by $n!$, is defined as the product of all natural numbers less than or equal to n, i.e. $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 = \prod_{i=1}^{n} i$. We define $0! = 1$.*

Definition 3.3.4 (Descending factorial)**.** *For $k, n \in N$ with $k \leq n$, we define the* descending factorial*, denoted by $(n)_k$ as $(n)_k = n(n-1)\cdots(n-k+1) = \prod_{i=0}^{k-1}(n-i) = \prod_{j=n-k+1}^{n} j$ with the convention that $(n)_1 = n$. We note that the descending factorial can be expressed as $(n)_k = \frac{n!}{(n-k)!}$.*

The factorial arises naturally in the context of so-called *permutations*. Consider the set of numbers $\{1, 2, \dots, n\}$. A permutation brings these numbers into a certain order. As a consequence of Theorem 3.3.2 with $k = n$, we deduce that the numbers in the set $\{1, 2, \dots, n\}$ can be arranged in exactly $n!$ possible ways.

Example 3.3.5. *Consider the set $\{1, 2, 3\}$. How many permutations (i.e. possible orderings) are there? We can write $(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)$, so we have $3! = 6$ possible permutations.*

3.3.3 The birthday problem

Let us now study the famous birthday problem:

Example 3.3.6. *Assume there are $k \in \mathbb{N}$ people in a room and assume that each person's birthday is equally likely to be any of the 365 days of the year (with the 29th February excluded). What is the probability that at least two people in the room have the same birthday?*

- *Due to our assumption, the naive probability definition is applicable here. First we count how many possible ways there are to assign birthdays to the k people in the room.*

- *This problem can be viewed as sampling with replacement, so we have* 365^k *possible birthday combinations.*
- *Next, we need to count how many scenarios there are such that at least two people have the same birthday. It appears that this is rather challenging...*
- *What is easier to compute is the complement, i.e. the number of scenarios such that no two people share the same birthday. This number can be computed using sampling without replacement, which leads to* $(365)_k$ *possible outcomes.*

Combining the results, we get

$$\begin{aligned}&\mathrm{P}_{\mathrm{Naive}}(\text{At least two people in the room have the same birthday})\\&= 1 - \mathrm{P}_{\mathrm{Naive}}(\text{All people in the room have distinct birthdays})\\&= 1 - \frac{(365)_k}{365^k} = 1 - \frac{365}{365}\frac{364}{365}\cdots\frac{365-(k-1)}{365} =: f(k).\end{aligned}$$

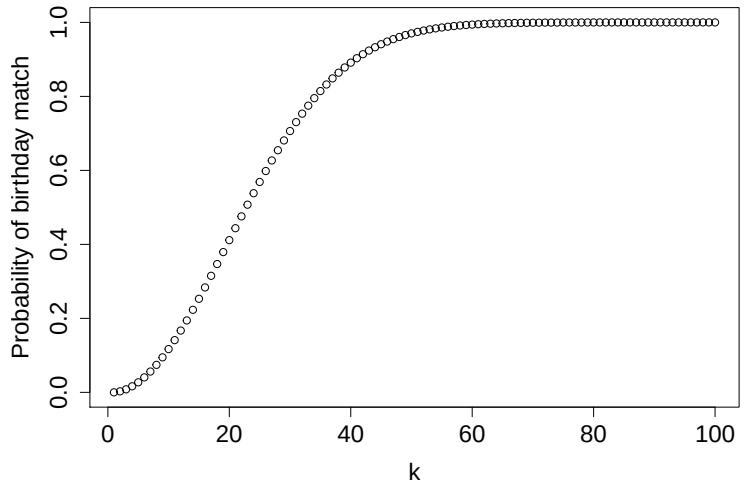


Figure 3.3: We compute the probability $f(k)$ that, out of k people in a room, at least two share the same birthday.

We plot the probabilities $f(k)$ *for* $k = 1, \ldots, 100$ *in Figure 3.3. Note that* $f(22) \approx 0.476$ *and* $f(23) \approx 0.507$*, so you need to have at least 23 people in the room to have a probability of at least* 50% *such that at least two people share the same birthday.*

End of lecture 3. .....

3.3.4 Sampling without replacement – unordered

Consider the case of an urn with $n \in \mathbb{N}$ balls, where we take out $k \in \mathbb{N}$ ($k \leq n$) balls and write down their labels, which are distinct numbers in $\{1, \ldots, n\}$. We do not care about the order in which the balls are collected. (For instance, think of drawing the winning numbers in the lottery!) Note that you could view this experiment as drawing k balls at once rather than one at a time. Hence the outcome of the experiment is a subset of size k from $S = \{1, \ldots, n\}$. Hence we can write $\Omega = \{\omega \subseteq S : \mathrm{card}(\omega) = k\}$.

Definition 3.3.7 (Binomial coefficient)**.** *For any $k, n \in \mathbb{N} \cup \{0\}$, the* binomial coefficient *is defined as the number of subsets of size k for a set of size n. It is denoted by $\binom{n}{k}$ and we say "n choose k".*

Theorem 3.3.8 (Binomial coefficient)**.** *For any $k, n \in \mathbb{N} \cup \{0\}$, we have*

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-(k-1))}{k!} = \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}.$$

for $k \leq n$, and $\binom{n}{k} = 0$ for $k > n$.

Proof. Consider the setting of the urn with n balls, where we draw k balls at once as described above. Clearly, if $k > n$, then $\binom{n}{k} = 0$. Suppose now that $k \leq n$. By Theorem 3.3.2, we know that there are $(n)_k$ possible choices if we draw k balls without replacement and care about the ordering. Now we need to make an adjustment for the overcounting since we do not care about the order any more. For each subset of size k, we have $k!$ permutations. So we conclude that when we divide $(n)_k$ by $k!$ we have adjusted for the overcounting and obtain the result. □

Example 3.3.9. *Recall the binomial theorem, which states that for any $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$ we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

The theorem can be proven as follows: We expand $(x+y)^n = (x+y)\cdots(x+y)$ in n factors and then pick either the x or the y from the first factor, multiply this to either the x or the y of the second factor and so on. There are $\binom{n}{k}$ ways of picking exactly k x's and each of these choices leads to a term of the form $x^k y^{n-k}$. Summing over all possible k leads the result.

3.3.5 Sampling with replacement – unordered

As before, let $k, n \in \mathbb{N}$. Let us consider the case that we have an urn with n balls with labels in $\{1, \ldots, n\}$, and we want to choose k balls one after the other with replacement. Assuming the order of the balls does not matter, how many possible outcomes are there?The sample space for our experiment is given by $\Omega = \{\omega : \omega \text{ is a } k - \text{element subset of} \{1, \ldots, n\}\}$.

Theorem 3.3.10 (Sampling with replacement when the order does not matter)**.** *In the sampling with replacement problem described above and assuming that the order of the balls does not matter, we have $\text{card}(\Omega) = \binom{n+k-1}{k}$ possibilities.*

The proof of the theorem relies on the so-called *stars and bars* argument presented in Feller (1957), which is a graphical tool for counting. Note that since we sample with replacement, we can have that $k > n$.

Proof. Consider n distinguishable boxes representing the n distinct labels of the balls in the urn. We can draw these n boxes using $n+1$ bars which represent the walls of the boxes, see Figure 3.4 for an illustration. Next, we have k indistinguishable balls, which we now draw as stars and place them into the boxes, i.e. between the bars.

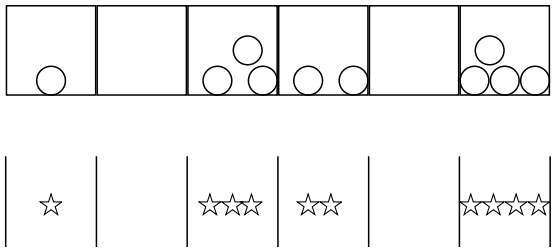


Figure 3.4: Illustration of the stars and bars method in the case when $n = 6$ and $k = 10$. Here we have $\binom{n+k-1}{k} = \binom{15}{10} = 3003$ possible outcomes.

I.e. we can view the stars as "check marks" which count how often a particular label gets selected. Between the two outer walls (i.e. the first and the last bar), which are fixed, there are $n-1$ bars and k stars, i.e. $n+k-1$ symbols which can be arranged in any possible order. Out of the $n+k-1$ possible positions, we choose k positions for the stars and fill the remaining positions with bars. Hence $\text{card}(\Omega) = \binom{n+k-1}{k}$. Note that we could have picked the $n-1$ bars instead and then filled in the remaining stars, which leads to the identity

$$\text{card}(\Omega) = \binom{n+k-1}{k} = \binom{n+k-1}{n-1}.$$

□

It is important to remember that you should not use the above result in connection with the naive probability since the unordered samples are typically not equally likely.

3.3.6 Summary table

We can summarise the preceding discussion as follows: Consider an urn with $n \in \mathbb{N}$ balls, where you draw $k \in \mathbb{N}$ balls. The number of possible outcomes is then given as follows:

	Ordered	Unordered
With replacement	n^k	$\binom{n+k-1}{k}$
Without replacement (for $k \leq n$ only)	$(n)_k$	$\binom{n}{k}$

End of lecture 4. .....

Chapter 4

Axiomatic definition of probability

The material of this chapter is based on Blitzstein & Hwang (2019), p.21-26, Anderson et al. (2018), p.1-21, Grimmett & Welsh (1986), p.3-9.

In this chapter we will focus on an axiomatic definition of probability and derive some of the key properties of the probability measure.

4.1 The event space $\mathcal{F}$

Recall Definition 2.2.1, which states that the *sample space* Ω is defined as the set of all possible outcomes of an experiment. In our previous discussion we assigned (naive) probabilities to *events* which were subsets of Ω. We typically denote by $\mathcal{F}$ the *event space*, which contains the events we are allowed to consider. This is a rather vague statement, which we will need to make precise! In fact, in probability theory, we always require that the event space $\mathcal{F}$ is a so-called σ-algebra (which is the same as a σ-field).

Definition 4.1.1 (Algebra and σ-algebra)**.** *A collection of subsets of Ω denoted by $\mathcal{F}$ is called*

1. *an* algebra *(or a* field*) if*
   (a) $\emptyset \in \mathcal{F}$,
   (b) $\mathcal{F}$ *is closed under complements, i.e.* $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, *and*
   (c) $\mathcal{F}$ *is closed under unions of pairs of members, i.e.* $A_1, A_2 \in \mathcal{F} \Rightarrow A_1 \cup A_2 \in \mathcal{F}$.
2. *an* σ-algebra *(or a* σ-field*) if*
   (a) $\emptyset \in \mathcal{F}$,
   (b) $\mathcal{F}$ *is closed under complements, i.e.* $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, *and*
   (c) $\mathcal{F}$ *is closed under countable union, i.e.* $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Remark 4.1.2.
1. *Any algebra is closed under finite unions and finite intersections.*
2. *Any σ-algebra is closed under countable intersections since $\cap_{i=1}^{\infty} A_i = (\cup_{i=1}^{\infty} A_i^c)^c$ and each $A_i^c \in \mathcal{F}$.*
3. *Any (σ-)algebra contains $\Omega = \emptyset^c$.*

Example 4.1.3. *Consider any sample space Ω. Then the so-called* trivial *σ-algebra is defined as $\mathcal{F}_{trivial} = \{\emptyset, \Omega\}$ and the* total *or* power *σ-field is defined as $\mathcal{F} = \mathcal{P}(\Omega) = \{\text{all subsets of } \Omega\}$.*

Example 4.1.4. *Consider a sample space Ω with $A \subseteq \Omega$. Then $\{\emptyset, \Omega, A, A^c\}$ is a σ-algebra (in fact the smallest σ-algebra including A).*

Throughout the course, we shall assume that $\mathcal{F}$ is a σ-algebra. This will allow us to consider countable infinite rather than finite unions. Note that this is clearly more restrictive than assuming that $\mathcal{F}$ is an algebra.

So why do we care about algebras at all? In probability we typically define a probability measure first on an algebra and extend it to a σ-algebra. The details behind this construction are beyond the scope of this introductory course, but you can learn more about this in our measure theory course.

4.2 Definition of probabilities and basic properties

4.2.1 Probability measure and probability space

Definition 4.2.1 (Probability measure). *A mapping $\mathrm{P} : \mathcal{F} \to \mathbb{R}$ is called a* probability measure *on $(\Omega, \mathcal{F})$ if it satisfies three conditions:*

(i) $\mathrm{P}(A) \geq 0$ *for all events $A \in \mathcal{F}$,*

(ii) $\mathrm{P}(\Omega) = 1$,

(iii) For any sequence of disjoint events $A_1, A_2, A_3, \cdots \in \mathcal{F}$ we have

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

[Note that by "disjoint events" we mean that $A_i \cap A_j = \emptyset$ for all $i \neq j$.]

Definition 4.2.2 (Probability space). *We define a probability space as the triplet $(\Omega, \mathcal{F}, \mathrm{P})$, where Ω is a set (the sample space), $\mathcal{F}$ is a σ-algebra (the event space) consisting of subsets of Ω and P is a probability measure on $(\Omega, \mathcal{F})$.*

4.2.2 Basic properties of the probability measure

Theorem 4.2.3. *Consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Then, for any events $A, B \in \mathcal{F}$, we have*

1. $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.
2. *If $A \subseteq B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$.*
3. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$.

Proof.

1. Show that $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$: Since A and A^c are disjoint and $\Omega = A \cup A^c$, the second axiom (ii) leads to $1 = \mathrm{P}(\Omega) = \mathrm{P}(A \cup A^c)$ and the third axiom (iii) leads to $\mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c)$. Altogether, we have $\mathrm{P}(A) + \mathrm{P}(A^c) = 1$.
2. Show that, if $A \subseteq B$, then $\mathrm{P}(A) \leq \mathrm{P}(B)$: We can express B as a union of two disjoint sets: $B = (B \cap A) \cup (B \cap A^c)$. Since $A \subseteq B$, we have that $B \cap A = A$. So, using the axiom (iii), we have that $\mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c) = \mathrm{P}(A) + \mathrm{P}(B \cap A^c)$. Using the fact that the probability measure is nonnegative (axiom (i)), we conclude that $\mathrm{P}(B \cap A^c) \geq 0$ and hence $\mathrm{P}(B) = \mathrm{P}(A) + \mathrm{P}(B \cap A^c) \geq \mathrm{P}(A)$.
3. Show $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$: We express A and B in terms of disjoint unions:

$$A = (A \cap B) \cup (A \cap B^c), \qquad B = (B \cap A) \cup (B \cap A^c).$$

By axiom (iii), we have that

$$\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c), \qquad \mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c).$$

Hence

$$\mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) + \mathrm{P}(B \cap A^c).$$

Also the union $A \cup B$ can be expressed as a union of disjoint sets

$$A \cup B = (A \cap B) \cup (A \cap B^c) \cup (B \cap A^c).$$

By axiom (iii), we have that

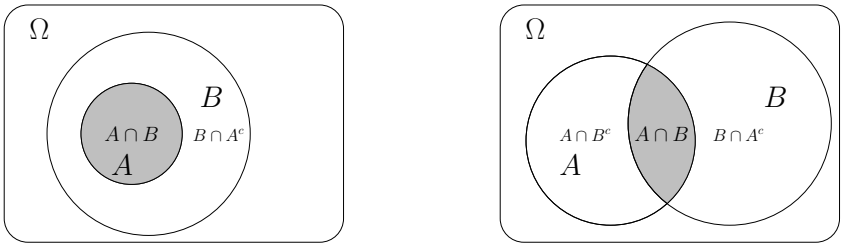
$$\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) + \mathrm{P}(B \cap A^c).$$

So, indeed,

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).$$

□

Some graphical illustrations for the arguments presented in the above proof are given in Figure 4.1.



(a) (b)

Figure 4.1: We consider a sample space Ω with subsets $A, B \subseteq \Omega$. Figures 4.1a and 4.1b depict the settings we are considering in the proofs of Theorem 4.2.3 part 2 and 3, respectively.

Remark 4.2.4. *The above theorem implies that* $\mathrm{P}(\emptyset) = 0$. *To see this, note that* Ω *and* $\emptyset$ *are disjoint since* $\Omega \cap \emptyset = \emptyset$. *Also,* $\Omega = \Omega \cup \emptyset$. *So, altogether we have* $1 = \mathrm{P}(\Omega) = \mathrm{P}(\Omega \cup \emptyset) = \mathrm{P}(\Omega) + \mathrm{P}(\emptyset) = 1 + \mathrm{P}(\emptyset)$. *Hence,* $\mathrm{P}(\emptyset) = 0$.

4.2.3 Examples

Example 4.2.5. *We continue with the classical example of flipping a fair coin.*

Figure 4.2: Flipping a fair coin.

We write H *for heads and* T *for tail. The sample space is given by* $\Omega = \{H, T\}$. *The event space can be taken as* $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$ *which is the collection of all subsets of* Ω. *Since we are considering a "fair" coin, we have that* $\mathrm{P}(\{H\}) = \mathrm{P}(\{T\}) = \frac{1}{2}$, *where we typically shorten the notation to:*

$$\mathrm{P}(H) = \mathrm{P}(T) = \frac{1}{2}.$$

Moreover, we have $\mathrm{P}(\emptyset) = 0$ *and* $\mathrm{P}(\Omega) = 1$.

End of lecture 5. .....

Chapter 5

Conditional probabilities

The material of this chapter is based on Blitzstein & Hwang (2019), p.45-63, Anderson et al. (2018), p.43-56, Grimmett & Welsh (1986), p.11-12.

After having introduced the axiomatic definition of a probability measure, we will now turn our attention to so-called conditional probabilities. We will learn how probabilities can be computed based on some given evidence. Conditional probabilities play a key role in almost all subsequent probability and statistics course und you will learn that they constitute a powerful concept for computing unconditional probabilities as well.

5.1 Definition

Definition 5.1.1 (Conditional probability)**.** *Consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Consider events $A, B \in \mathcal{F}$ with $\mathrm{P}(B) > 0$. Then the* conditional probability *of A given B, denoted by $\mathrm{P}(A|B)$, is defined as*

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

Remark 5.1.2. *Interpretation: You could call $\mathrm{P}(A)$ the* prior *probability of event A and $\mathrm{P}(A|B)$ the* posterior *probability of A. Here we view B as additional evidence which becomes available, and the prior probability is formulated without knowledge of the additional evidence and the posterior probability describes the updated probability based on the additional evidence.*

Let us now show that the conditional probability measure does indeed satisfy the axioms of a probability measure:

Theorem 5.1.3 (Conditional probability)**.** *Let $B \in \mathcal{F}$ with $\mathrm{P}(B) > 0$ and define $\mathrm{Q} : \mathcal{F} \to \mathbb{R}$ by $\mathrm{Q}(A) = \mathrm{P}(A|B)$. Then $(\Omega, \mathcal{F}, \mathrm{Q})$ is a probability space.*

Proof. The only thing we need to show is that Q satisfies the axioms of a probability measure on $(\Omega, \mathcal{F})$.

Axiom (i): Since P is a probability measure satisfying axiom (i), we deduce that, for any $A \in \mathcal{F}$, $\mathrm{Q}(A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \geq 0$. **Axiom (ii):**

$$\mathrm{Q}(\Omega) = \frac{\mathrm{P}(\Omega \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B)}{\mathrm{P}(B)} = 1.$$

Axiom (iii): Consider disjoint events $A_1, A_2, \cdots \in \mathcal{F}$. Then

$$\mathrm{Q}\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{\mathrm{P}((\bigcup_{i=1}^{\infty} A_i) \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(\bigcup_{i=1}^{\infty} (A_i \cap B))}{\mathrm{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathrm{P}(A_i \cap B)}{\mathrm{P}(B)}$$
$$= \sum_{i=1}^{\infty} \frac{\mathrm{P}(A_i \cap B)}{\mathrm{P}(B)} = \sum_{i=1}^{\infty} \mathrm{Q}(A_i),$$

where for the third equality we used the fact that the events $A_i \cap B \in \mathcal{F}$ are disjoint and that P satisfies axiom (iii). □

5.2 Examples

Example 1. We roll a single fair die. Hence $\Omega = \{1, 2, 3, 4, 5, 6\}$. What is the probability that the score is greater than 3 given that the score is even? We define the events $B = \{\omega \in \Omega : \omega \text{ is even}\} = \{2, 4, 6\}$, $A = \{\omega \in \Omega : \omega > 3\} = \{4, 5, 6\}$. Then, $A \cap B = \{4, 6\}$. We have $\mathrm{P}(A) = 1/2, \mathrm{P}(B) = 1/2$ and $\mathrm{P}(A \cap B) = 2/6 = 1/3$.

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Example 2. A family has two children. Assume that Female (F)/Male (M) are equally likely and successive births are independent. We write $\Omega = \{FF, FM, MF, MM\}$, where e.g. FM stands for the event that the first child is Female and the second child is Male. Note that all four outcomes are equally likely, so $\mathrm{P}(\omega) = 1/4$ for all $\omega \in \Omega$. Questions:

1. If one child is a boy, what is the probability that the other child is a boy?
2. If the eldest is a boy, what is the probability that the other child is a boy?

Let $A = \{MM\}$ both male, $B = \{MM, MF, FM\}$ at least one male, $C = \{MM, FM\}$ eldest is male. Then

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{1/4}{3/4} = \frac{1}{3},$$

and

$$\mathrm{P}(A|C) = \frac{\mathrm{P}(A \cap C)}{\mathrm{P}(C)} = \frac{1/4}{2/4} = \frac{1}{2}.$$

Remark 5.2.1. *Consider the case of a finite state space Ω where all events are equally likely and hence the classical interpretation of probability can be used. Then for two events $A, B \subseteq \Omega$, we have that*

$$\mathrm{P}(A|B) = \frac{\mathrm{card}(A \cap B)}{\mathrm{card}(B)} = \frac{\frac{\mathrm{card}(A \cap B)}{\mathrm{card}(\Omega)}}{\frac{\mathrm{card}(B)}{\mathrm{card}(\Omega)}} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

5.3 Multiplication rule

Suppose that $A, B \in \mathcal{F}$ with $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$. Then

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

Similarly,

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

Let us extend the above result to three events: Let $C \in \mathcal{F}$ with $\mathrm{P}(C) > 0$. Then

$$\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A|B \cap C)\mathrm{P}(B \cap C) = \mathrm{P}(A|B \cap C)\mathrm{P}(B|C)\mathrm{P}(C).$$

Repeating the arguments above, we obtain the following result:

Theorem 5.3.1 (Multiplication rule). *Let $n \in \mathbb{N}$, then for any events $A_1, \ldots, A_n$ with $\mathrm{P}(A_2 \cap \cdots \cap A_n) > 0$, we have*

$$\mathrm{P}(A_1 \cap \cdots \cap A_n) = \mathrm{P}(A_1|A_2 \cap \cdots \cap A_n)\mathrm{P}(A_2|A_3 \cap \cdots \cap A_n) \cdots \\ \mathrm{P}(A_{n-2}|A_{n-1} \cap A_n)\mathrm{P}(A_{n-1}|A_n)\mathrm{P}(A_n),$$

where the right hand side is a product of n terms.

Clearly, the ordering of the events in the theorem above can be changed, so there are $n!$ possible formulations of the above theorem!

End of lecture 6. .....

5.4 Bayes' rule and law of total probability

5.4.1 Bayes' rule

Bayes' rule is a famous and extremely useful result for computing conditional probabilities:

Theorem 5.4.1. *Let $A, B \in \mathcal{F}$ with $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$. Then*

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$$

Proof. This is an immediate consequence of the definition of conditional probability and the multiplication rule. □

5.4.2 Law of total probability

Next we study the so-called *law of total probability*, which is an extremely useful tool for computing complicated probabilities in terms of simpler pieces based on conditional probabilities.

Definition 5.4.2 (Partition). *A* partition *of the sample space Ω is a collection $\{B_i : i \in \mathcal{I}\}$ (for a countable index set I) of disjoint events (meaning that $B_i \in \mathcal{F}$ and $B_i \cap B_j = \emptyset$ for $i \neq j$) such that $\Omega = \bigcup_{i \in \mathcal{I}} B_i$.*

Remark 5.4.3. *We note that a partition of the sample space is often not unique and the choice of the particular partition typically very much depends on the problem we want to solve!*

Theorem 5.4.4 (Law of total probability). *Let $\{B_i : i \in \mathcal{I}\}$ denote a partition of Ω, with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$. Then*

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i) = \sum_{i \in \mathcal{I}} \mathrm{P}(A|B_i)\mathrm{P}(B_i).$$

Proof. Using the properties of a partition, we deduce that

$$A = A \cap \Omega = A \cap \left(\bigcup_{i \in \mathcal{I}} B_i \right) = \bigcup_{i \in \mathcal{I}} (A \cap B_i),$$

where the events $A \cap B_i$ are disjoint. Using axiom (iii) of the definition of the probability measure leads to

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i),$$

and using the multiplication formula leads to

$$\mathrm{P}(A) = \sum_{i \in \mathcal{I}} \mathrm{P}(A \cap B_i) = \sum_{i \in \mathcal{I}} \mathrm{P}(A|B_i)\mathrm{P}(B_i).$$

□

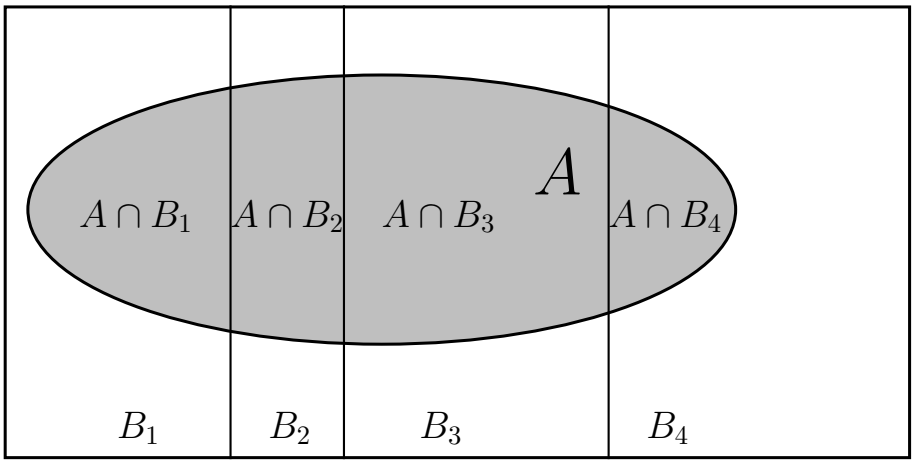


Figure 5.1: Illustration of the law of total probability. Here we have that $\mathrm{P}(A) = \sum_{i=1}^{4} \mathrm{P}(A \cap B_i)$.

5.4.3 General Bayes' rule

When we combined Bayes' rule with the law of total probability, we get the following useful result:

Theorem 5.4.5. *Consider a partition $\{B_i : i \in \mathcal{I}\}$ of Ω with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$, then for any event $A \in \mathcal{F}$ with $\mathrm{P}(A) > 0$, we have*

$$\mathrm{P}(B_i|A) = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B_i)\mathrm{P}(B_i)}{\sum_{k\in\mathcal{I}} \mathrm{P}(A|B_k)\mathrm{P}(B_k)}.$$

5.5 Example: Testing for a rare disease

Example 3. Let us look in more detail into visualisation of conditional probabilities, which often arise in medical screening tests. We quote the example from the article (Spiegelhalter et al. 2011, p. 1396).

Consider a

> 'mammography test on a population with a 1% prevalence of breast cancer. The test is positive for around 90% of women with cancer, but it is also positive for around 10% of women without cancer'. ((Spiegelhalter et al. 2011, p. 1396))

What is the probability that a woman whose mammography test is positive has breast cancer?

Define the events: B := breast cancer present; TP = test is positive. We know that $\mathrm{P}(B) = 0.01$ and $\mathrm{P}(TP|B) = 0.9$ and $\mathrm{P}(TP|B^c) = 0.1$. We can derive that $\mathrm{P}(B^c) = 1 - \mathrm{P}(B) = 0.99$. Further, by the law of total probability,

$$\begin{aligned}\mathrm{P}(TP) &= \mathrm{P}(TP|B)\mathrm{P}(B) + \mathrm{P}(TP|B^c)\mathrm{P}(B^c)\\ &= 0.9 \cdot 0.01 + 0.1 \cdot 0.99 = 0.108.\end{aligned}$$

Using Bayes' Theorem, we get

$$\mathrm{P}(B|TP) = \frac{\mathrm{P}(B \text{ and } TP)}{\mathrm{P}(TP)} = \frac{\mathrm{P}(TP|B)\mathrm{P}(B)}{\mathrm{P}(TP)} = \frac{0.9 \cdot 0.01}{0.108} \approx 8\%$$

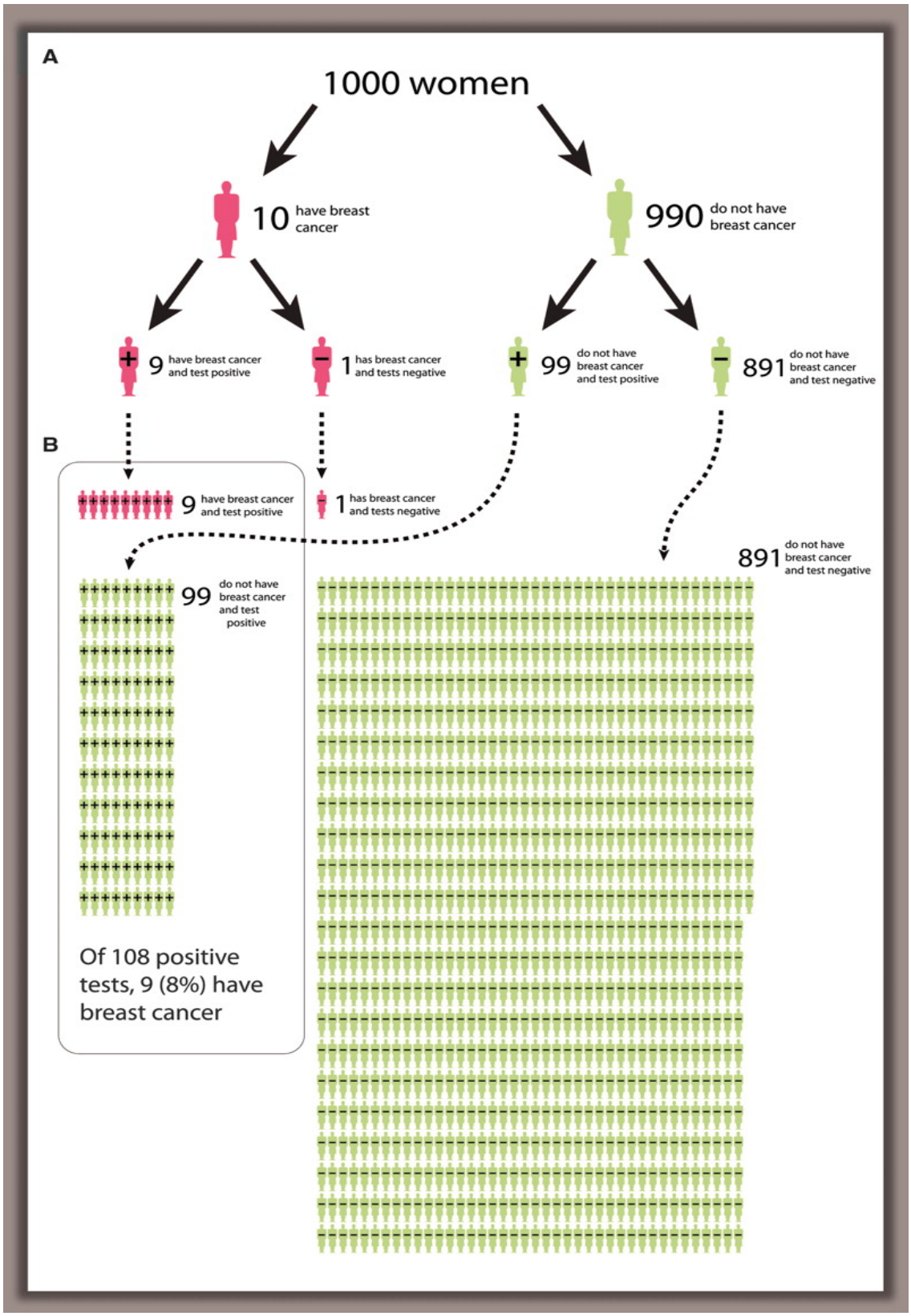


Figure 5.2: Visualising conditional probabilities in the case of testing for a rare disease. This picture is a copy of Figure 4 in the article Spiegelhalter et al. (2011).

5.6 Strategy: Condition on the missing information

As a general strategy, we typically use conditioning on the information we wish we had to make our probability computations easier. To illustrate this strategy, we consider a famous example.

5.6.1 Example: Monty Hall

In the TV Game show *Let's make a deal*, hosted by Monty Hall, a contestant selects one of three doors; behind one of the doors there is a prize (a car), and behind the other two there are no prizes (in fact, there are goats!). After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether they want to SWITCH their choice to the other unopened door, or STICK to their original choice. Is it probabilistically advantageous for the contestant to SWITCH doors, or is the probability of winning the prize the same whether they STICK or SWITCH ? (Assume that the host selects a door to open, from those available, with equal probability).

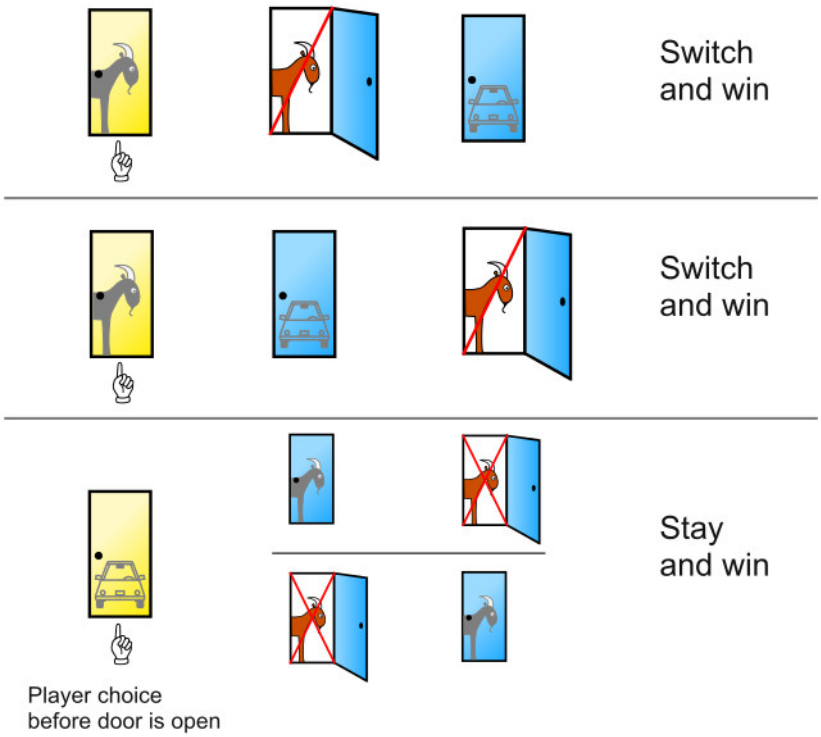


Figure 5.3: In the Monty Hall example it is advantageous to SWITCH!

Label the doors 1, 2 and 3 and assume without loss of generality that the candidate selects door 1. Monty Hall then opens a door and reveals a goat. When deciding whether or not to switch, which information would the contestant like to have? She would like to know the location of the car.

So let us consider the partition: $\{C_i, i = 1, 2, 3\}$ where C_i is the event that the car is behind the ith door (for $i = 1, 2, 3$). Also, we denote by H_2 the event that Monty Hall opens door 2. Then $\mathrm{P}(C_1) = \mathrm{P}(C_2) = \mathrm{P}(C_3) = 1/3$ and $\mathrm{P}(H_2|C_1) = 1/2$, $\mathrm{P}(H_2|C_2) = 0$ and $\mathrm{P}(H_2|C_3) = 1$. We want to compare the probabilities of $\mathrm{P}(C_1|H_2)$ (STICK) with $\mathrm{P}(C_3|H_2)$ (SWITCH). Using the law of total probability, we have

$$\begin{aligned}\mathrm{P}(H_2) &= \mathrm{P}(H_2|C_1)\mathrm{P}(C_1) + \mathrm{P}(H_2|C_2)\mathrm{P}(C_2) + \mathrm{P}(H_2|C_3)\mathrm{P}(C_3) \\ &= \frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3} = \frac{1}{2}.\end{aligned}$$

Then, using the (general) Bayes' rule implies that

$$\mathrm{P}(C_1|H_2) = \frac{\mathrm{P}(H_2|C_1)\mathrm{P}(C_1)}{\mathrm{P}(H_2)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

Similarly

$$\mathrm{P}(C_3|H_2) = \frac{\mathrm{P}(H_2|C_3)\mathrm{P}(C_3)}{\mathrm{P}(H_2)} = \frac{1\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

So it is better to SWITCH, see Figure 5.3.

End of lecture 7. .....

Chapter 6

Independence

The material of this chapter is based on Blitzstein & Hwang (2019), p.63-65, Anderson et al. (2018), p.51-56, Grimmett & Welsh (1986), p.12-16.

6.1 Independence of events

We will call two events $A, B \in \mathcal{F}$ *independent* if the occurrence of one of them does not affect the probability that the other one occurs, meaning that, if $\mathrm{P}(A) > 0, \mathrm{P}(B) > 0$,

$$\mathrm{P}(A|B) = \mathrm{P}(A), \text{ and } \mathrm{P}(B|A) = \mathrm{P}(B). \tag{6.1.1}$$

Now, recall the definition of the conditional probability as

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

Hence, the following definition appears suitable:

Definition 6.1.1 (Independent events)**.** *The events A, B are called* independent *if*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B), \tag{6.1.2}$$

and dependent *otherwise.*

Remark 6.1.2. *The definition given in equation 6.1.2 is more general than the one in equation 6.1.1 since it does not require that A and B have nonzero probabilities.*

Theorem 6.1.3. *If the events A and B are independent, then the same is true for each of the pairs A^c and B, A and B^c, and A^c and B^c.*

Proof. We only prove that $\mathrm{P}(A^c \cap B) = \mathrm{P}(A^c)\mathrm{P}(B)$ (the proof for the remaining pairs follows the same arguments). From the law of total probability, we have

$$\mathrm{P}(B) = \mathrm{P}(B \cap A) + \mathrm{P}(B \cap A^c) = \mathrm{P}(B)\mathrm{P}(A) + \mathrm{P}(B \cap A^c),$$

where we used that A and B are independent. Rearranging the terms leads to

$$\mathrm{P}(B \cap A^c) = \mathrm{P}(B)(1 - \mathrm{P}(A)) = \mathrm{P}(B)\mathrm{P}(A^c).$$

□

Let us now generalise the definition of independence to more than two events.

Definition 6.1.4 (Independence of events (general case)). *1. A finite collection of events $A_1, \ldots, A_n$ is defined to be independent if*

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k})$$

for every subcollection $\{i_1, \ldots, i_k\}$, $k = 1, \ldots, n$.

2. A countable or uncountably infinite collection of events is defined to be independent if each finite subcollection is independent.

Remark 6.1.5. *Note that pairwise independence of (A_i, A_j) is in general not sufficient to conclude the independence of $(A_1, \ldots, A_n)$.*

Example 6.1.6. *The three events A_1, A_2, A_3 are independent if and only if*

$$\begin{aligned} \mathrm{P}(A_1 \cap A_2 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_2)\mathrm{P}(A_3), \\ \mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_1)\mathrm{P}(A_2), \\ \mathrm{P}(A_1 \cap A_3) &= \mathrm{P}(A_1)\mathrm{P}(A_3), \\ \mathrm{P}(A_2 \cap A_3) &= \mathrm{P}(A_2)\mathrm{P}(A_3), \end{aligned}$$

Example 6.1.7. *We roll two fair dice and write the sample space as $\Omega = \{(i, j) : i, j = 1, \ldots, 6\}$. We note that $\mathrm{card}(\Omega) = 6^2 = 36$ and all outcomes are equally likely. We define three events: $A_1 =$ first roll is odd, $A_2 =$ second roll is odd, $A_3 =$ sum is odd. Then A_1, A_2, A_3 are pairwise independent, but they are not independent since $\mathrm{P}(A_1 \cap A_2 \cap A_3) = 0$.*

6.1.1 Conditional independence of events

Definition 6.1.8 (Conditional independence of events). *Consider events $A, B, C \in \mathcal{F}$ with $\mathrm{P}(C) > 0$. Then we say that A and B are* conditionally independent given *C if*

$$\mathrm{P}(A|B \cap C) = \mathrm{P}(A|C),$$

or, equivalently,

$$\mathrm{P}(A \cap B|C) = \mathrm{P}(A|C)\mathrm{P}(B|C).$$

6.1.2 Product rule and continuity of the probability measure

We can now formulate the so-called product rule for countable number of independent sets:

Theorem 6.1.9. *If $A_1, A_2, \ldots$ is a countably infinite set of independent events, then*

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \prod_{i=1}^{\infty} \mathbb{P}(A_i).$$

The proof of the above theorem relies on the so-called *continuity property* of the probability measure, which we will introduce next.

Lemma 6.1.10. *Any countable union can be written as a countable union of disjoint sets. I.e. let $A_1, A_2, \cdots \in \mathcal{F}$ and define $D_1 = A_1, D_2 = A_2 \setminus A_1, D_3 = A_3 \setminus (A_1 \cup A_2), \ldots$. Then $\{D_i\}$ is a collection of disjoint sets and $\cup_{i=1}^n A_i = \cup_{i=1}^n D_i$ for n being any positive integer or ∞.*

The proof of the Lemma is left as an exercise, see Exercise 3- 5.

Definition 6.1.11 (Increasing and decreasing sets). *A sequence of sets $(A_i)_{i=1}^{\infty}$ is said to* increase *to A, i.e. $A_i \uparrow A$, if $A_1 \subset A_2 \subset \cdots$ and $\cup_{i=1}^{\infty} A_i = A$. Similarly, a sequence of sets $(A_i)_{i=1}^{\infty}$ is said to* decrease *to A, i.e. $A_i \downarrow A$, if $A_1 \supset A_2 \supset \cdots$ and $\cap_{i=1}^{\infty} A_i = A$.*

Next we will state and prove the continuity property of the probability measure[1].

Theorem 6.1.12. *If $A_1, A_2, \cdots \in \mathcal{F}$ and $A_i \uparrow A$ or $A_i \downarrow A$, then $\mathbb{P}(A_i) \to \mathbb{P}(A)$ as $i \to \infty$.*

Proof. Suppose that $A_i \uparrow A$. Then using Lemma 6.1.10, we write $A = \cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} D_i$, where the $D_i = A_i \setminus (\cup_{k=1}^{i-1} A_k)$ are disjoint. By axiom (iii) of the definition of the probability measure, we deduce that

$$\begin{aligned}\mathbb{P}(A) &= \sum_{i=1}^{\infty} \mathbb{P}(D_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(D_i) \\ &= \lim_{n\to\infty} \mathbb{P}(\cup_{i=1}^{n} D_i) = \lim_{n\to\infty} \mathbb{P}(\cup_{i=1}^{n} A_i) = \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned}$$

Now let $A_i \downarrow A$. Then $F_i = A_i^c \uparrow F = A^c$. Then, we deduce from the first part of the proof that $\mathbb{P}(F_i) \to \mathbb{P}(F)$. Using the properties of a probability measure, since $\mathbb{P}(F_i) = 1 - \mathbb{P}(A_i)$ and $\mathbb{P}(F) = 1 - \mathbb{P}(A)$, we deduce that $\mathbb{P}(A_i) \to \mathbb{P}(A)$. □

End of lecture 8.

Proof of Theorem 6.1.9. Let $B_n = \cap_{i=1}^{n} A_i$. Then $B_n \downarrow B = \cap_{i=1}^{\infty} A_i$, so by the continuity property of the probability measure, see Theorem 6.1.12, we deduce that

$$\begin{aligned}\mathrm{P}(\cap_{i=1}^{\infty} A_i) = \mathrm{P}(B) &= \lim_{n\to\infty} \mathrm{P}(B_n) \\ &= \lim_{n\to\infty} \mathrm{P}(\cap_{i=1}^{n} A_i) = \lim_{n\to\infty} \prod_{i=1}^{n} \mathrm{P}(A_i) = \prod_{i=1}^{\infty} \mathrm{P}(A_i).\end{aligned}$$

□

[1] Recall that a sequence of real numbers (x_n) is said to converge to a real number x if for all $\epsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $|x_n - x_0| < \epsilon$.

Chapter 7

Discrete random variables

The material of this chapter is based on Blitzstein & Hwang (2019), p.103-120, Grimmett & Welsh (1986), p.24-28.

In this and the following chapter, we will be introducing discrete and continuous random variables and their distributions.

7.1 Random variables

We consider a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and we will think of a *random variable* (r.v.) as a function from the sample space to the real numbers $\mathbb{R}$, i.e.

$$X : \Omega \to \mathbb{R}.$$

The function needs to satisfy some properties, which we introduce in the formal definition below. Note that

- Despite the name, a random variable is a *function* and not a variable.
- We typically use capital letters such as X, Y, Z to denote random variables.
- The value of the random variable X at the sample point ω is given by $X(\omega)$ and is called a *realisation* of X.
- The randomness stems from $\omega \in \Omega$ (we don't know which outcome ω appears in the random experiment), the mapping itself given by X is deterministic.

7.2 Discrete random variables and probability distributions

Definition 7.2.1 (Discrete random variable)**.** *A* discrete random variable *on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is defined as a mapping $X : \Omega \to \mathbb{R}$ such that*

(i) the set $\{X(\omega) : \omega \in \Omega\} (=: \mathrm{Im}X)$ (called the image/range of Ω under X*) is a countable subset of $\mathbb{R}$,*

(ii) $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.

Remark 7.2.2. *The name* discrete *stems from the first condition in the above definition, which says that the random variable can only take countably many values in $\mathbb{R}$. In most applications, we deal with discrete random variables taking values in (a subset of) $\mathbb{N}$ or $\mathbb{Z}$.*

Remark 7.2.3. *Let us clarify the second condition in the above definition: We note that the set appearing there is the so-called pre-image of x defined as*

$$X^{-1}(x) = \{\omega \in \Omega : X(\omega) = x\},$$

i.e. the set of all ω which X maps to x. We require that this set is an event in $\mathcal{F}$ (for all possible x) so that we can later assign probabilities to these events.

Definition 7.2.4 (Probability mass function)**.** *The* probability mass function *(pmf) of the discrete random variable X is defined as the function $p_X : \mathbb{R} \to [0,1]$ given by*

$$p_X(x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}). \tag{7.2.1}$$

We typically shorten the notation significantly and write $p_X(x) = \mathrm{P}(X = x)$. Keep in mind, that this is short hand notation for equation (7.2.1).

Note that the definition of the pmf implies the following properties:

$$p_X(x) = 0 \quad \text{if } x \notin \mathrm{Im}X.$$

Using axiom (iii) in the definition of the probability measure, we have

$$\sum_{x \in \mathrm{Im}X} p_X(x) = \mathrm{P}\left(\bigcup_{x \in \mathrm{Im}X} \{\omega \in \Omega : X(\omega) = x\} \right) = \mathrm{P}(\Omega) = 1.$$

The above equation is often written as

$$\sum_{x \in \mathbb{R}} p_X(x) = 1,$$

since only countably many values of x result in non-zero values for the pmf and hence non-zero contributions to the sum.

Theorem 7.2.5. *Let $\mathcal{I}$ denote a countable (index) set. Suppose that $S = \{s_i : i \in \mathcal{I}\}$ is a countable set of distinct real numbers and $\{\pi_i : i \in \mathcal{I}\}$ is a collection of numbers satisfying*

$$\pi_i \geq 0 \text{ for all } i \in \mathcal{I}, \text{and } \sum_{i \in \mathcal{I}} \pi_i = 1,$$

then there exists a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and a discrete random variable X on that probability space such that its probability mass function is given by

$$\begin{aligned} p_X(s_i) &= \pi_i, \qquad \text{for all } i \in \mathcal{I} \\ p_X(s) &= 0, \qquad \text{if } s \notin S. \end{aligned}$$

Proof. This is a constructive proof: Take $\Omega = S$, let $\mathcal{F} = \mathcal{P}(\Omega)$ be the power set (i.e. the set of all subsets of Ω) and set

$$\mathrm{P}(A) = \sum_{i : s_i \in A} \pi_i \qquad \text{for all } A \in \mathcal{F}.$$

The discrete random variable $X : \Omega \to \mathbb{R}$ is then defined as $X(\omega) = \omega$ for all $\omega \in \Omega$. □

The above theorem is incredibly useful for our further study of discrete random variables. It implies that we do not need to worry about sample spaces, event spaces and probability measures too much. Instead, we can just say that we study a random variable X taking the value s_i with probability π_i for $i \in \mathcal{I}$ and we know that such a random variable actually exists!

End of lecture 9. .....

7.3 Common discrete distributions

In this section, we will introduce some widely used discrete distributions.

7.3.1 Bernoulli distribution

Definition 7.3.1 (Bernoulli distribution)**.** *A discrete random variable X is said to have* Bernoulli distribution *with parameter $p \in (0,1)$, if X can only take two possible values, 0 and 1, i.e. $\mathrm{Im}X = \{0,1\}$ and*

$$p_X(1) = \mathrm{P}(X=1) = p, \quad p_X(0) = \mathrm{P}(X=0) = 1-p, \quad p_X(x) = 0 \text{ if } x \notin 0,1.$$

We write $X \sim \mathrm{Bern}(p)$.

Note that for *any* event there is a natural way of associating a Bernoulli random variable with it: We can define the so-called indicator variable of the event:

Definition 7.3.2 (Indicator variable)**.** *Consider an event $A \in \mathcal{F}$, we denote by*

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A, \end{cases}$$

the indicator variable of the event *A.*

Note that the random variable $\mathbb{I}_A \sim \mathrm{Bern}(p)$ with $p = \mathrm{P}(A)$, since

$$\mathrm{P}(\mathbb{I}_A = 1) = \mathrm{P}(A), \quad \mathrm{P}(\mathbb{I}_A = 0) = \mathrm{P}(A^c) = 1 - \mathrm{P}(A), \quad \mathrm{P}(\mathbb{I}_A = x) = 0 \text{ for } x \notin \{0,1\}.$$

Background: Think of an experiment with two possible outcomes "success" or "failure" (but not both). We call such an experiment a *Bernoulli trial*. We can think of a Bernoulli random variable as an indicator of success, where an outcome of 1 represents success and an outcome of 0 represents failure. Hence we often call the parameter p in the Bernoulli distribution the *success probability*.

7.3.2 Binomial distribution

Consider a sequence of $n \in \mathbb{N}$ independent and identical Bernoulli trials with success probability $p \in (0,1)$ and count the number of successes and denote it by the random variable X [e.g. count the number of heads when tossing a coin repeatedly]. Then X can take the values $\mathrm{Im}X = \{0,1,\dots,n\}$. Let $x \in \mathrm{Im}X$, and suppose we have x successes and $n-x$ failures. Since the trials are independent, the probability of any sequence with x successes is $p^x(1-p)^{n-x}$. In total, there are $\binom{n}{x}$ possible sequences with x successes and $n-x$ failures, hence

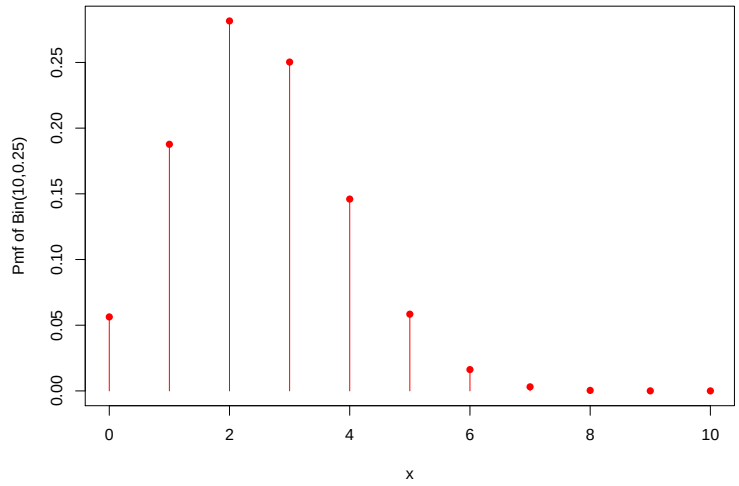
$$\mathrm{P}(X=x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Definition 7.3.3 (Binomial distribution)**.** *A discrete random variable X is said to follow the* binomial distribution *with parameters $n \in \mathbb{N}$ and $p \in (0,1)$ if $\mathrm{Im}X = \{0,1,\dots,n\}$ and*

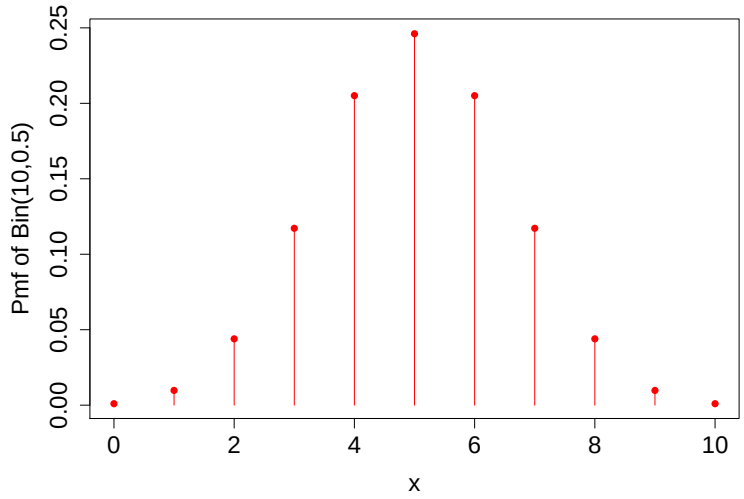
$$\mathrm{P}(X=x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x \in \{0,1,\dots,n\},$$

and $\mathrm{P}(X=x) = 0$ otherwise. We write $X \sim \mathrm{Bin}(n,p)$.

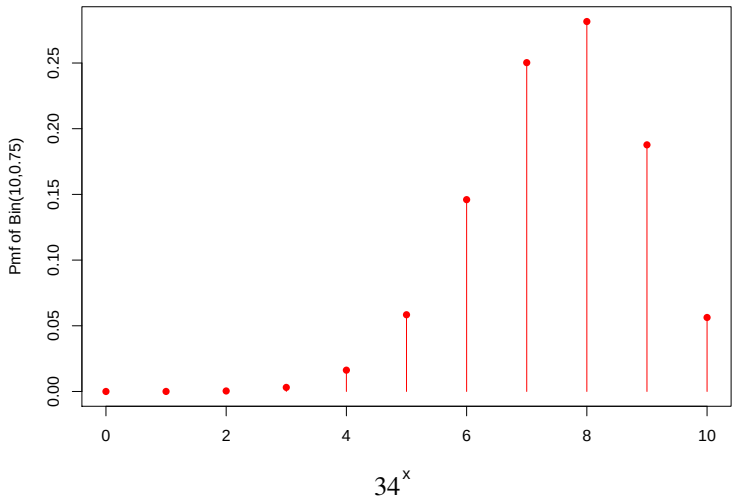
We depict the probability mass function for three random variables with binomial distribution and parameters $n = 10$ and $p \in \{0.25, 0.5, 0.75\}$ in Figure 7.1. We observe that for $p = 0.5$ the pmf is symmetric about $n/2 = 5$ and skewed when $p \neq 0.5$. We will show prove this finding on the problem sheet.



(a) P.m.f. of $X \sim \text{Bin}(10, 0.25)$



(b) P.m.f. of $X \sim \text{Bin}(10, 0.5)$



(c) P.m.f. of $X \sim \text{Bin}(10, 0.75)$

Figure 7.1: We depict the probability mass function for three random variables with binomial distribution and parameters $n = 10$ and $p \in \{0.25, 0.5, 0.75\}$. Note that for $p = 0.5$ the pmf is symmetric about 5 and skewed when $p \neq 0.5$.

7.3.3 Hypergeometric distribution

Consider an urn filled with N balls, with $K \in \mathbb{N}$ being white balls and $N - K$ being black. When we draw $n \in N$ balls *with replacement*, we obtain a $\text{Bin}(n, K/N)$ distribution for the number of white balls drawn. Suppose now we draw *without replacement*, then the number of white balls follows the so-called *hypergeometric distribution*.

Definition 7.3.4 (Hypergeometric distribution)**.** *A discrete random variable X is said to follow the* hypergeometric distribution *with the three parameters $N \in \mathbb{N} \cup \{0\}$, $K, n \in \{0, 1, \ldots, N\}$ if $\text{Im}X = \{0, 1, \ldots, \min(n, K)\}$ and*

$$\text{P}(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, \text{ for } x \in \{0, 1, \ldots, K\} \text{ and } n - x \in \{0, 1, \ldots, N - K\},$$

and $\text{P}(X = x) = 0$ otherwise. We write $X \sim \text{HGeom}(N, K, n)$.

Remark 7.3.5. *We think of N as the size of the population, K the number of success states in the population (e.g. number of white balls), n the number of draws and x is the number of observed successes.*

In Figure 7.2 we show how the pmf of the hypergeometric distributions with $N = 500$ and $K = 200$ shifts when we increase the number of draws from $n = 10$ to 30 and then 50.

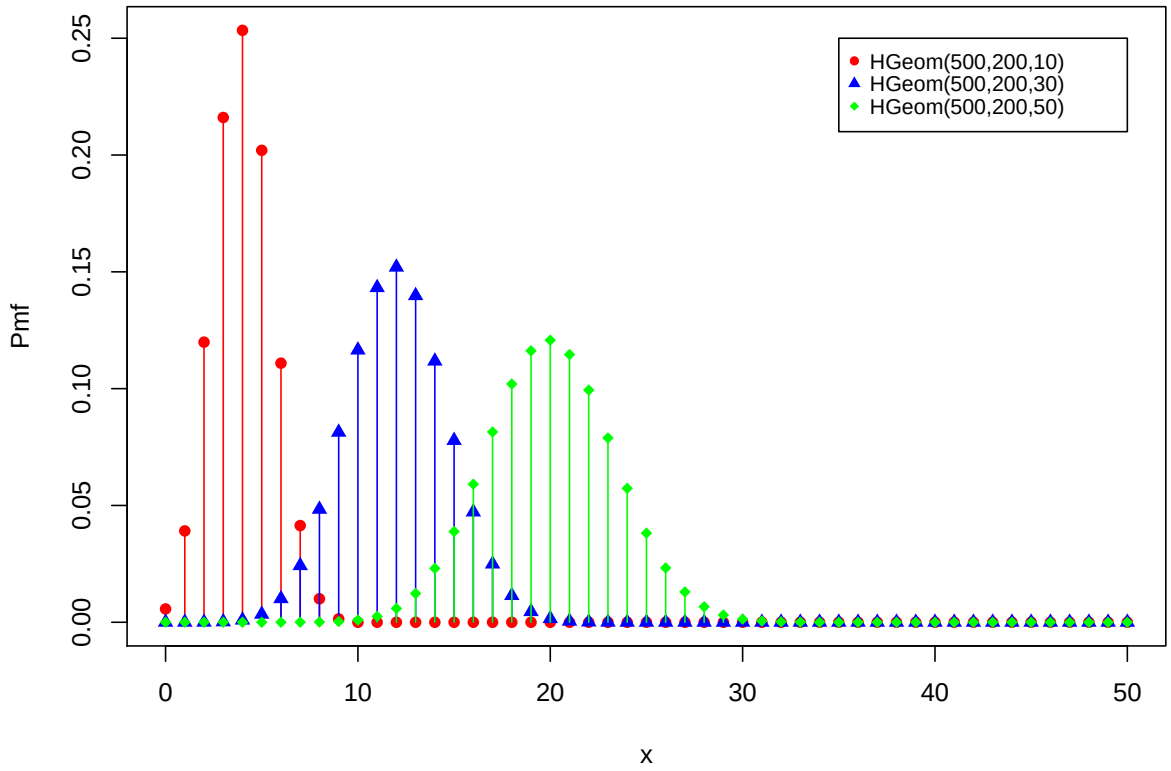


Figure 7.2: This graph shows the probability mass function of the hypergeometric distribution with parameters $N = 500$, $K = 200$ and $n \in \{10, 30, 50\}$.

7.3.4 Discrete uniform distribution

Definition 7.3.6 (Discrete uniform distribution)**.** *Let C denote a finite nonempty set of numbers. We say that a discrete random variable X follows the* discrete uniform distribution on C*, i.e. $X \sim \mathrm{DUnif}(C)$, if $\mathrm{Im}X = C$ and*

$$\mathrm{P}(X = x) = \frac{1}{\mathrm{card}(C)},$$

for $x \in C$ and $\mathrm{P}(X = x) = 0$ otherwise.

Example 7.3.7. *Let $C = \{1, \ldots, n\}$. If $X \sim \mathrm{DUnif}(C)$, then $\mathrm{P}(X = x) = 1/n$ for all $x \in \{1, \ldots, n\}$ and 0 otherwise.*

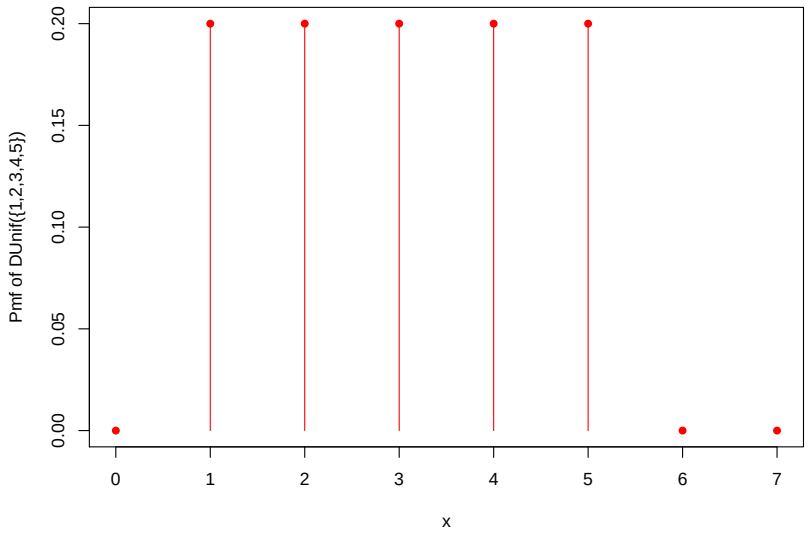


Figure 7.3: This graph shows the probability mass function of the discrete uniform distribution on the set $C = \{1, 2, 3, 4, 5\}$.

7.3.5 Poisson distribution

We will now introduce the Poisson distribution which is widely used for counting the number of events/-successes in a certain time period, e.g. the number of earthquakes in some region in the world.

Definition 7.3.8 (Poisson distribution)**.** *A discrete random variable X is said to follow the* Poisson distribution *with parameter $\lambda > 0$, i.e. $X \sim \mathrm{Poi}(\lambda)$, if $\mathrm{Im}X = \{0, 1, 2, \ldots\} = \mathbb{N} \cup \{0\}$ and*

$$\mathrm{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad \textit{for } x = 0, 1, 2, \ldots.$$

We typically call the parameter λ in the Poisson distribution the *rate* or *intensity* [of the occurrence of (rare) events].

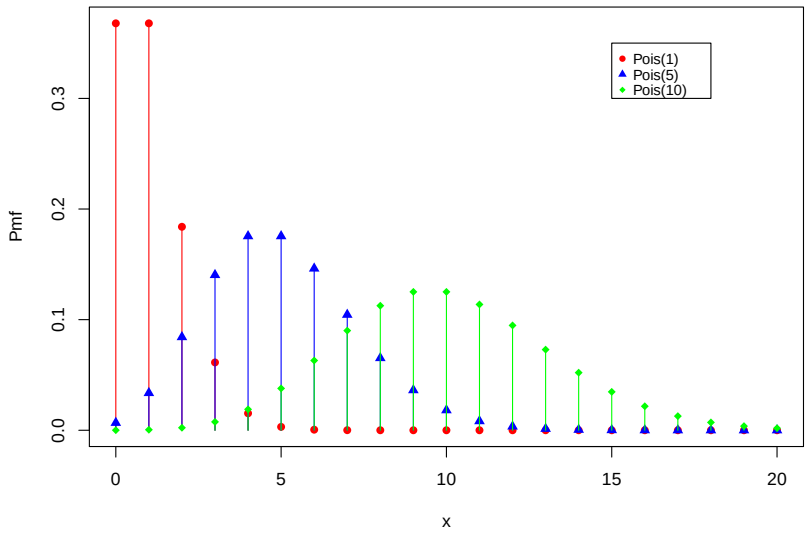


Figure 7.4: This graph shows the probability mass function of the Poisson distribution with three different rate parameters: $\lambda \in \{1, 5, 10\}$.

7.3.6 Geometric distribution

Definition 7.3.9 (Geometric distribution)**.** *A discrete random variable X is said to follow the* geometric distribution *with parameter $p \in (0, 1)$, i.e. $X \sim \text{Geom}(p)$ if $\text{Im}X = \mathbb{N}$ and*

$$\text{P}(X = x) = (1 - p)^{x-1}p, \quad \textit{for } x = 1, 2, \ldots.$$

We can think of an experiment where we carry out repeated (independent) Bernoulli trials with success probability p. We stop the experiment after the first success. We denote by X the number of trials to obtain the first success. Then we obtain that $X \sim \text{Geom}(p)$. **Warning:** If we set Y to be the *number of failures until first success* we obtain a slightly different definition of the geometric distribution. Here we have that $\text{Im}Y = \mathbb{N} \cup \{0\}$ and

$$\text{P}(Y = x) = (1 - p)^{x}p, \quad \text{for } x = 0, 1, 2, \ldots.$$

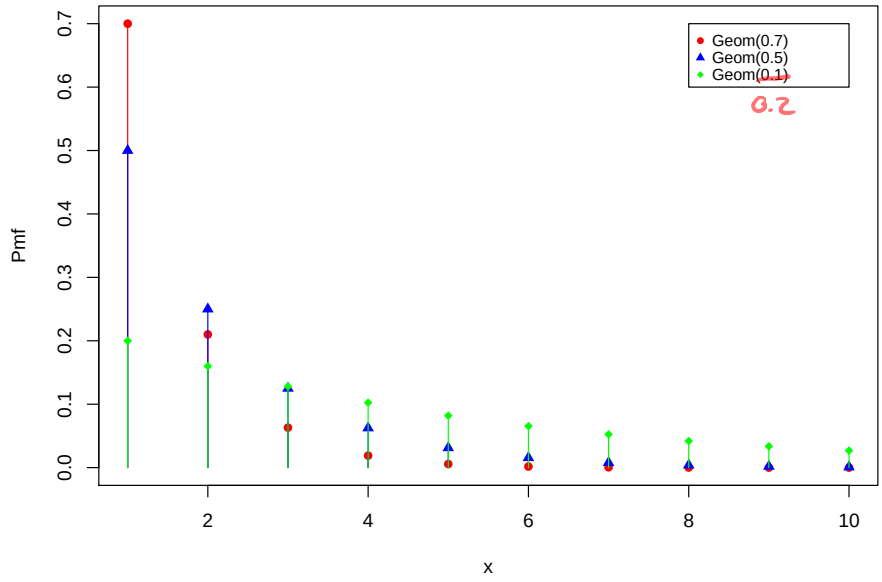


Figure 7.5: This graph shows the probability mass function of the Geometric distribution with three different success probabilities: $p \in \{0.7, 0.5, 0.2\}$.

7.3.7 Negative binomial distribution

Definition 7.3.10 (Negative binomial distribution)**.** *A discrete random variable X is said to follow the* negative binomial distribution *with parameters $r \in \mathbb{N}$ and $p \in (0,1)$, written $X \sim \mathrm{NBin}(r,p)$, if $\mathrm{Im}X = \mathbb{N} \cup \{0\}$ and*

$$\mathrm{P}(X = x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \quad \text{for } x = 0, 1, \ldots. \tag{7.3.1}$$

The negative binomial distribution arises as the distribution of the number of failures in a sequence of independent Bernoulli trials with success parameter p before r successes have occurred. To see this, let us consider strings of "0" (for failure) and "1" (for success). Each string of r "1"s and x "0"s has probability $p^r(1-p)^x$. Now we need to find the number of such strings: We stop when we reach the rth success, so the last element in the string will always be a "1". This leaves us with $r+x-1$ positions, to which we need to assign the remaining $r-1$ "1"s. Hence we obtain equation (7.3.1).

Remark 7.3.11. *Recall that in the case of a $\mathrm{Bin}(n,p)$ distribution, we also consider a sequence of independent Bernoulli trials, but we fix the number of trials n and count the number of successes.*

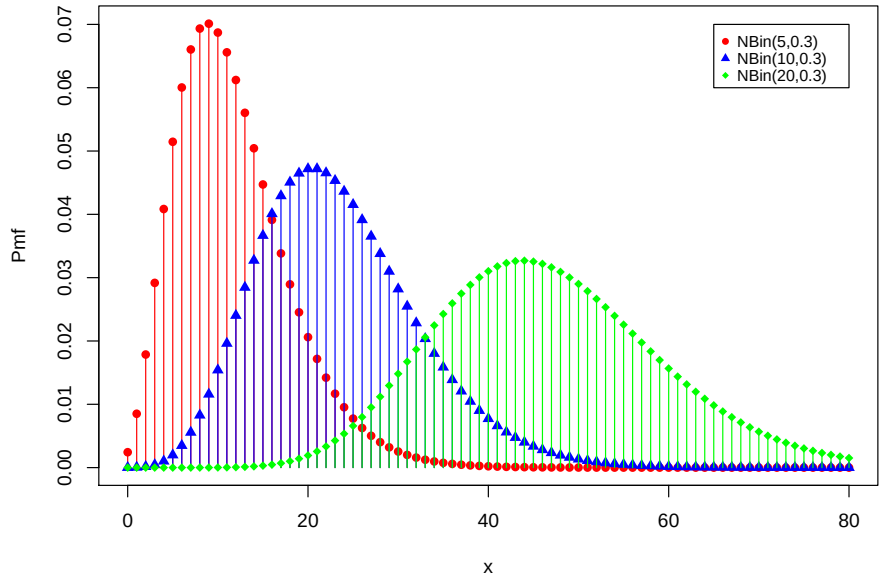


Figure 7.6: This graph shows the probability mass function of the negative binomial distribution with parameter $p = 0.3$ and $r \in \{5, 10, 20\}$.

7.3.8 Exercise

Exercise 7.3.12. *Verify that all the probability mass functions listed above are valid in the sense that* $p_X(x) \geq 0$ *for all* x *and* $\sum_x p_X(x) = 1$.

End of lecture 10.

Chapter 8

Continuous random variables

The material of this chapter is based on Blitzstein & Hwang (2019), p.121-123, 213-244, Grimmett & Welsh (1986), p.56-65.

8.1 Random variables and their distributions

So far, we have only considered discrete random variables which can take at most countably many values. Now we will give a more general definition which is also suitable for broader applications.

Definition 8.1.1 (Random variable)**.** *A* random variable *of the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ is defined as the mapping $X : \Omega \to \mathbb{R}$ which satisfies*

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \qquad \textit{for all } x \in \mathbb{R}. \tag{8.1.1}$$

Note that a discrete random variable (see Definition 7.2.1) satisfies this more general definition as well. We can write

$$\{\omega \in \Omega : X(\omega) \leq x\} = \bigcup_{y \in \mathrm{Im}X : y \leq x} \{\omega : X(\omega) = y\}.$$

The right hand side is a countable union of elements of $\mathcal{F}$ and (according the the definition of the sigma-algebra) hence also an element of $\mathcal{F}$.

Remark 8.1.2. *Note that, similarly to our previous definition, we call the set $X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \leq x\}$ the pre-image of $(-\infty, x]$. We can only make probability statements about the set $X^{-1}((-\infty, x])$ if it is an element of the event space $\mathcal{F}$ which motivates our definition of a random variable.*

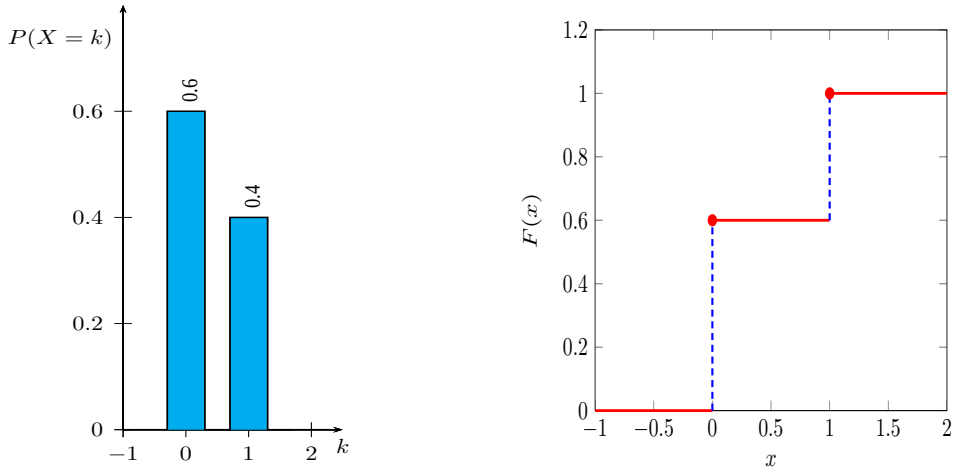
Definition 8.1.3 (Cumulative distribution function (c.d.f.))**.** *Suppose that X is a random variable on $(\Omega, \mathcal{F}, \mathrm{P})$, then the* cumulative distribution function *(c.d.f.) of X is defined as the mapping $F_X : \mathbb{R} \to [0, 1]$ given by*

$$F_X(x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x\}),$$

which is typically abbreviated to $F_X(x) = \mathrm{P}(X \leq x)$.

Example 8.1.4. *Consider a Bernoulli random variable $X \sim Ber(p)$. The c.d.f. of a Bernoulli random variable is given by*

$$F_X(x) = \mathrm{P}(X \leq x) = \sum_{k \leq x} \mathrm{P}(X = k) = \begin{cases} 0, & \textit{for } x < 0, \\ 1 - p, & \textit{for } x \in [0, 1), \\ 1, & \textit{for } x \geq 1. \end{cases}$$



(a) P.m.f. of $X \sim \text{Ber}(0.4)$

(b) C.d.f. of $X \sim \text{Ber}(0.4)$

Figure 8.1: Consider a Bernoulli random variable X with parameter $p = 0.4$. Its probability mass function is depicted in Figure 8.1a and its cumulative distribution function in Figure 8.1b.

Let us now derive important properties of the c.d.f..

Theorem 8.1.5 (Properties of the c.d.f.)**.** *1. F_X is monotonic non-decreasing, i.e. for all $x \leq y$ we have $F_X(x) \leq F_X(y)$.*

2. F_X is right-continuous, i.e. for all sequences $(x_n)_{n\in\mathbb{N}}$, with $\lim_{n\to\infty} x_n = x$ and $x_n \geq x$, we have $\lim_{n\to\infty} F_X(x_n) = F_X(x)$.

3. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.

Proof. 1. Monotonicity: Let $x \leq y$. Then

$$\{\omega \in \Omega : X(\omega) \leq x\} \subseteq \{\omega \in \Omega : X(\omega) \leq y\}.$$

Then the result follows from the monotonicity of the probability measure, see the second statement in Theorem 4.2.3.

2. Right continuity: We prove that if $x_n \downarrow x$ (i.e.$(x_n)_{n\in\mathbb{N}}$, with $\lim_{n\to\infty} x_n = x$ and $x_n \geq x$), then $F_X(x_n) \to F_X(x)$ as $n \to \infty$. Define events

$$E_n = \{\omega : X(\omega) \leq x_n\} \downarrow E = \{\omega : X(\omega) \leq x\}.$$

Using the continuity of the probability measure, see Theorem 6.1.12, $\mathbb{P}(E_n) \to \mathbb{P}(E)$. Since $\mathbb{P}(E_n) = F_X(x_n), \mathbb{P}(E) = F_X(x)$, we have that $F_X(x_n) \to F_X(x)$ as $n \to \infty$.

3. Limit behaviour at $\pm\infty$: Define for an x_n

$$E_n = \{\omega : X(\omega) \leq x_n\}.$$

Then $E_n \downarrow \emptyset$ as $x_n \downarrow -\infty$ and $E_n \uparrow \Omega$ as $x_n \uparrow \infty$. Using the continuity property of $\mathbb{P}$ again, we deduce that $F(x_n) = \mathbb{P}(E_n) \to \mathbb{P}(\emptyset) = 0$ as $x_n \to -\infty$, and $F(x_n) = \mathbb{P}(E_n) \to \mathbb{P}(\Omega) = 1$ as $x_n \to \infty$.

□

Remark 8.1.6. *One can show that for any function F which satisfies the three conditions stated in Theorem 8.1.5, there exists a probability space and a random variable on that space which has F as its c.d.f..*

Note that in applications, we often use the following result:

Theorem 8.1.7. *For $a < b$, we have $\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a)$.*

Proof. Note that for $a < b$, we have

$$\{\omega \in \Omega : X(\omega) \leq b\} = \{\omega \in \Omega : X(\omega) \leq a\} \cup \{\omega \in \Omega : a < X(\omega) \leq b\},$$

where the two events on the right hand side are disjoint. Hence

$$\mathrm{P}(\{\omega \in \Omega : X(\omega) \leq b\}) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq a\}) + \mathrm{P}(\{\omega \in \Omega : a < X(\omega) \leq b\}),$$

which implies that

$$\mathrm{P}(\{\omega \in \Omega : a < X(\omega) \leq b\}) = F_X(b) - F_X(a).$$

□

End of lecture 11.

8.2 Continuous random variables and probability density function

When looking at the c.d.f. of a Bernoulli random variable, see Figure 8.1b, we noted that the c.d.f. looks like a step function. Indeed, all discrete random variables have c.d.f.s which are right-continuous step functions (with possibly (many) more steps than in the Bernoulli case). In the remainder of this chapter we will now focus on random variables with a smooth c.d.f.:

Definition 8.2.1 (Continuous random variable and probability density function)**.** *A random variable X is called* continuous *if its c.d.f. can be written as*

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du, \qquad \text{for all } x \in \mathbb{R}, \tag{8.2.1}$$

where the function $f_X : \mathbb{R} \to \mathbb{R}$ satisfies

(i) $f_X(u) \geq 0$ *for all* $u \in \mathbb{R}$,

(ii) $\int_{-\infty}^{\infty} f_X(u)du = 1$.

We call f_X the probability density function *(p.d.f.) of X (or just the* density*).*[1]

The so-called Fundamental Theorem of Calculus guarantees that a function F_X given as in Definition 8.2.1 is differentiable at every point x where f is continuous with $F_X'(x) = f_X(x)$.

Remark 8.2.2. *Note that $f_X(x)$ is* not *a probability and while f_X is non-negative it is not restricted to be smaller than 1.*

We compare properties of the p.m.f. and the c.d.f. in the following table:

Discrete random variable	Continuous random variable
$p_X(x) \geq 0$, for all $x \in \mathbb{R}$	$f_X(x) \geq 0$, for all $x \in \mathbb{R}$
$\sum_{x \in \mathrm{Im}X} p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x) = 1$
$F_X(x) = \sum_{u \in \mathrm{Im}X : u \leq x} p_X(u)$	$F_X(x) = \int_{-\infty}^{x} f_X(u)du$

Table 8.1: Comparing discrete and random variables with p.m.f. p_X and p.d.f. f_X, respectively.

[1] In a later analysis/measure course we will say that equation 8.2.1 means that the "c.d.f. of a continuous random variable is *absolutely continuous* with respect to the Lebesgue measure".

It turns out that, although $f_X(x)$ is not a probability, it can be linked to a probability when we scale it appropriately. Consider a small quantity which we shall denote by $dx > 0$. Then the probability that X is close to x can be written as

$$\mathrm{P}(x < X \leq x + dx) = F_X(x + dx) - F_X(x) = \int_x^{x+dx} f_X(u)du \approx f_X(x)dx.$$

So, we can view the quantity $f_X(x)dx$ as the continuous analogue to a probability mass function $p_X(x)$.

The reason why we typically do not consider point probabilities for continuous random variables becomes clear in the next theorem.

Theorem 8.2.3. *For a continuous random variable X with density f_X, we have*

$$\mathrm{P}(X = x) = 0, \qquad \textit{for all } x \in \mathbb{R}, \tag{8.2.2}$$

and

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f_X(u)du, \qquad \textit{for all } a, b \in \mathbb{R} \textit{ with } a \leq b. \tag{8.2.3}$$

Proof. Consider any $x \in \mathbb{R}$ with a sequence $x_n \uparrow x$ (i.e.$(x_n)_{n\in\mathbb{N}}$, with $\lim_{n\to\infty} x_n = x$ and $x_n \leq x$), and define events

$$E_n = \{\omega : x_n < X(\omega) \leq x\} \downarrow E = \{\omega : X(\omega) = x\}.$$

Using the continuity of the probability measure, see Theorem 6.1.12, $\mathbb{P}(E_n) \to \mathbb{P}(E)$. Hence we can write

$$\begin{aligned}\mathrm{P}(X = x) &= \lim_{n\to\infty} \mathrm{P}(E_n) = \lim_{n\to\infty} \mathrm{P}(\{\omega : x_n < X(\omega) \leq x\}) \\ &= \lim_{n\to\infty} (F_X(x) - F_X(x_n)) \\ &= \lim_{n\to\infty} \int_{x_n}^{x} f_X(u)du = 0.\end{aligned}$$

Now, let $a \leq b$, then we know from the above that $\mathrm{P}(X = a) = 0$, hence

$$\mathrm{P}(a \leq X \leq b) = \mathrm{P}(a < X \leq b) = F_X(b) - F_X(a),$$

where we used Theorem 8.1.7.

□

Remark 8.2.4. *We note that the c.d.f. of a continuous random variable is continuous. It is important to remember that the definition of the continuous random variable guarantees the existence of the density and then the continuity of the associated c.d.f. follows from the properties of the (improper) Riemann integral. Note that if you only assumed that a random variable X has a continuous c.d.f. with, in particular, $\mathrm{P}(X = x) = 0$ for all x, then the existence of a density function is not guaranteed. A notorious example of such a case is the so-called* Cantor function *which you might study in a later analysis/measure course.*

8.3 Common continuous distributions [Reading material]

8.3.1 Uniform

Definition 8.3.1 (Uniform distribution)**.** *A continuous random variable X is said to have the* uniform distribution *on the interval (a, b) for $a < b$, i.e. $X \sim \mathrm{U}(a, b)$, if its density function is given by*

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \textit{if } a < x < b, \\ 0, & \textit{otherwise.} \end{cases}$$

Its cumulative distribution function is given by

$$F_X(x) = \begin{cases} 0, & \textit{if } x \leq a, \\ \frac{x-a}{b-a}, & \textit{if } a < x < b, \\ 1, & \textit{if } x \geq b. \end{cases}$$

8.3.2 Exponential

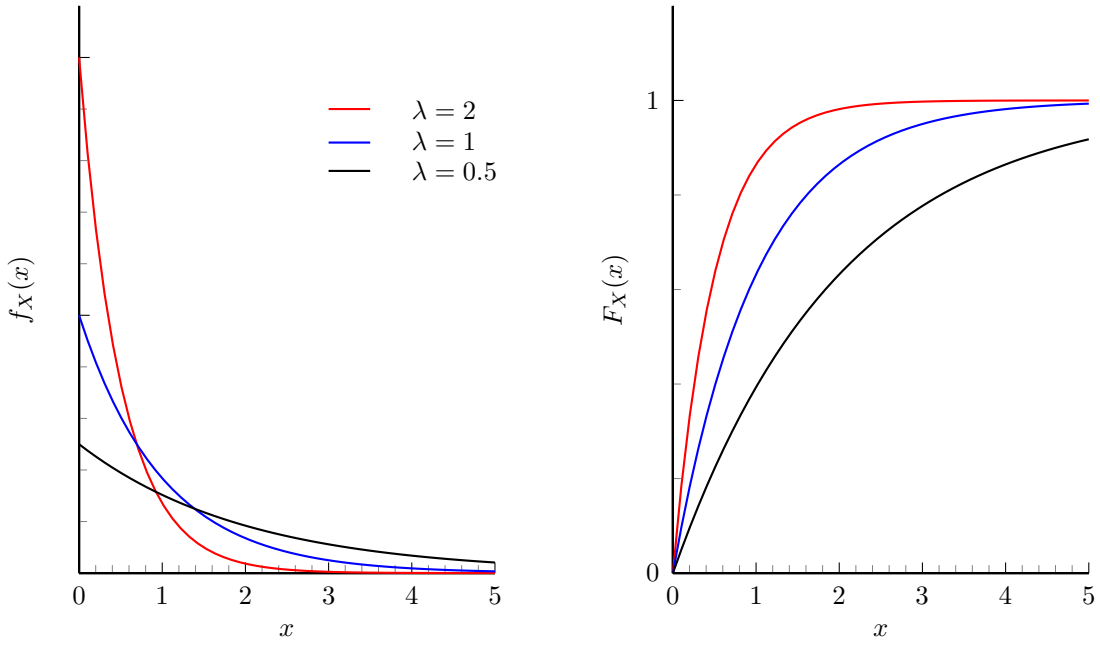


Figure 8.2: Plot of the p.d.f. (left) and the c.d.f. (right) of an $\text{Exp}(\lambda)$ random variable for $\lambda \in \{0.5, 1, 2\}$.

Definition 8.3.2 (Exponential distribution)**.** *A continuous random variable X is said to have the* exponential distribution *with parameter $\lambda > 0$, i.e. $X \sim \text{Exp}(\lambda)$, if its density function is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Its cumulative distribution function is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1 - e^{-\lambda x}, & \text{if } x > 0. \end{cases}$$

8.3.3 Gamma distribution

The Gamma distribution is – as the exponential distribution – also supported on the positive real line only and extends the exponential distribution discussed above.

For $t > 0$ we define the *Gamma function* by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx,$$

which has the following properties:

$$\Gamma(t) = (t-1)\Gamma(t-1), \text{ for } t > 1,$$

and in the case when $t \in \mathbb{N}$ we have $\Gamma(t) = (t-1)!$.

Definition 8.3.3 (Gamma distribution)**.** *A continuous random variable X is said to have the* Gamma distribution *with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, i.e. $X \sim \text{Gamma}(\alpha, \beta)$, if its density function is given by*

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Its cumulative distribution function is not available in closed form.

In the case when $\alpha = n \in \mathbb{N}$, we often call the Gamma distribution the *Erlang distribution* which has density

$$f_X(x) = \begin{cases} \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

8.3.4 Chi-squared distribution

Definition 8.3.4 (Chi-squared distribution)**.** *A continuous random variable X is said to have the* chi-squared distribution *with $n \in \mathbb{N}$ degrees of freedom, i.e. $X \sim \chi^2(n)$ (or also $X \sim \chi^2_n$) , if its density function is given by*

$$f_X(x) = \begin{cases} \frac{1}{2\Gamma(n/2)} \left(\frac{x}{2}\right)^{n/2-1} e^{-x/2}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Its cumulative distribution function is not available in closed form.

We note that the $\chi^2(n)$ distribution is the same as the $\text{Gamma}(n/2, 1/2)$ distribution.

8.3.5 F-distribution

Definition 8.3.5 (F-distribution)**.** *A continuous random variable X is said to have the* F-distribution *with $d_1, d_2 > 0$ degrees of freedom, i.e. $X \sim F(d_1, d_2)$ (or also $X \sim F_{d_1,d_2}$) , if its density function is given by*

$$f_X(x) = \begin{cases} \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)\left(1+\frac{d_1}{d_2}x\right)^{(d_1+d_2)/2}}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Its cumulative distribution function is not available in closed form.

We note that the positive parameters d_1, d_2 are not restricted to be integer-valued.
Note that if we have independent random variables $X_1 \sim \chi^2_n$ and $X_2 \sim \chi^2_m$, then the random variable

$$X = \frac{X_1/n}{X_2/m} \sim F_{n,m}.$$

8.3.6 Beta distribution

For $\alpha, \beta > 0$ denote by

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

the so-called Beta function.

Definition 8.3.6 (Beta distribution)**.** *A continuous random variable X is said to have the* Beta distribution *with parameters $\alpha, \beta > 0$, i.e. $X \sim \text{Beta}(\alpha, \beta)$, if its density function is given by*

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Its cumulative distribution function is not available in closed form.

8.3.7 Normal distribution

Definition 8.3.7 (Standard normal distribution)**.** *A random variable X has the* standard normal/standard Gaussian *distribution if it has density function $f(x) = \phi(x)$ with*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \textit{for } x \in \mathbb{R}.$$

Note that we typically write $X \sim \mathrm{N}(0,1)$ since a standard normal random variable has mean zero and variance one. The c.d.f. is then denoted by $F(x) = \Phi(x)$ with

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt, \quad \text{for } x \in \mathbb{R}.$$

Unfortunately there is no explicit formula for the integral appearing in the c.d.f.!

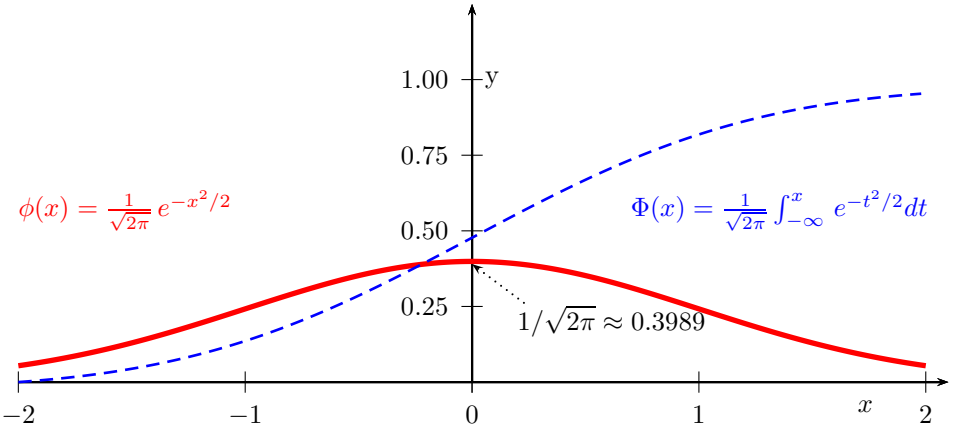


Figure 8.3: The red solid line depicts the standard Gaussian probability density function and the blue dashed line the corresponding cumulative distribution function.

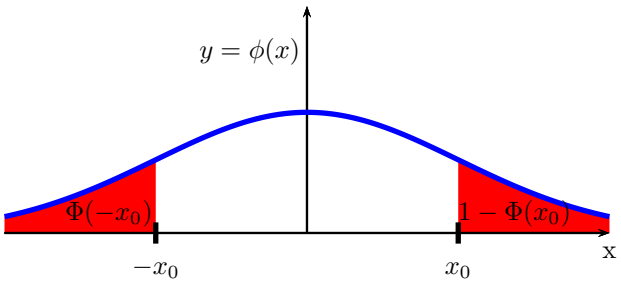


Figure 8.4: Note that the standard normal density is symmetric around 0, i.e. $\phi(x) = \phi(-x)$ for all x. This also implies that $\Phi(-x) = 1 - \Phi(x)$.

Definition 8.3.8 (Normal distribution)**.** *Let μ denote a real number and let $\sigma > 0$. A random variable X has the* normal/ Gaussian *distribution with mean μ and variance σ^2 if it has density function*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \textit{for } x \in \mathbb{R}.$$

Note that we typically write $X \sim \mathrm{N}(\mu, \sigma^2)$.

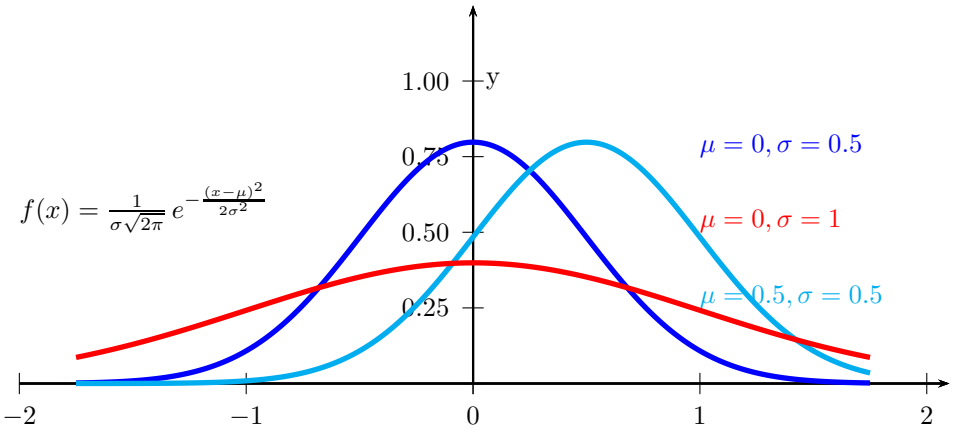


Figure 8.5: The red line depicts the standard Gaussian probability density function and the two blue lines show non-standard Gaussian probability density functions.

8.3.8 Cauchy distribution

Definition 8.3.9 (Cauchy distribution)**.** *A continuous random variable X is said to have the* Cauchy distribution*, if its density function is given by*

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \qquad \text{for } x \in \mathbb{R}.$$

Its cumulative distribution function is given by

$$F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}, \qquad \text{for } x \in \mathbb{R}.$$

We note that if we have two independent standard normal random variables $X, Y \sim \mathrm{N}(0,1)$, then their ratio $Z = X/Y$ follows the Cauchy distribution.

8.3.9 Student t-distribution

Definition 8.3.10 ((Student's) t-distribution)**.** *A continuous random variable X is said to have the* (Student's) t-distribution *with $\nu > 0$ degrees of freedom, if its density function is given by*

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad \text{for } x \in \mathbb{R}.$$

Its cumulative distribution function is not available in closed form.

End of lecture 12. .....

Chapter 9

Transformations of random variables

The material of this chapter is based on Blitzstein & Hwang (2019), p.123-129, Grimmett & Welsh (1986), p.28-29, 65-67.

Let us consider a random variable X on $(\Omega, \mathcal{F}, \mathrm{P})$ and a (deterministic) function $g : \mathbb{R} \to \mathbb{R}$. Clearly $Y = g(X)$ is a mapping from Ω to $\mathbb{R}$ with $Y(\omega) = g(X(\omega))$. In this chapter, we would like to study under which conditions Y is itself a random variable and we would like to study its distribution

9.1 The discrete case

Let us first consider the case when X is a discrete random variable. Then, for $Y = g(X)$, we have that $\mathrm{Im}(Y)$ is countable since $\mathrm{Im}(X)$ is countable. In this case, we can easily see that since

$$\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}, \text{ for all } x,$$

we immediately get that also

$$\{\omega \in \Omega : Y(\omega) = y\} = \{\omega \in \Omega : g(X(\omega)) = y\} \in \mathcal{F}, \text{ for all } y,$$

hence $Y = g(X)$ is indeed a random variable. We can compute its p.m.f. as follows:

$$p_Y(y) = \mathrm{P}(Y = y) = \mathrm{P}(g(X) = y) = \sum_{x \in \mathrm{Im}X : g(x) = y} \mathrm{P}(X = x), \tag{9.1.1}$$

so we are just summing up the probabilities for all values x for which $g(x) = y$. If g is invertible (i.e. bijective), then

$$p_Y(y) = p_X(g^{-1}(y)) \qquad \text{for all } y \in \mathrm{Im}Y.$$

9.2 The continuous case

For the continuous (or more general case) recall that $Y = g(X)$ is only a random variable if Y satisfies condition (8.1.1), i.e.

$$\{\omega \in \Omega : Y(\omega) \leq y\} \in \mathcal{F} \qquad \text{for all } y \in \mathbb{R}.$$

This condition is only satisfied if g satisfies some additional properties (e.g. if it is continuous or monotone[1]).

[1] More generally, we will need that g is Borel-measurable, but this concept is beyond the scope of this course.

Example 9.2.1. *Consider a linear transformation of the random variable X. I.e. let $a > 0, b \in \mathbb{R}$ and define $g(x) = ax + b$. Then $Y = g(X) = aX + b$ is indeed a random variable and, for any $y \in \mathbb{R}$ its c.d.f. is given by*

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(aX + b \leq y) = \mathrm{P}\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

We can now differentiate (with respect to y– using the chain rule) and obtain

$$f_Y(y) = F_Y'(y) = F_X'\left(\frac{y-b}{a}\right)\frac{1}{a} = \frac{1}{a}f_X\left(\frac{y-b}{a}\right).$$

In the previous example, we have seen that in the case that the function g can be inverted, we can find an explicit formula for the corresponding density of the transformed random variable. We can now state and prove this result in a more general form.

Theorem 9.2.2. *Suppose that X is a continuous random variable with density f_X and $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing and differentiable with inverse function denoted by g^{-1}, then $Y = g(X)$ has density*

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)], \qquad \text{for all } y \in \mathbb{R}.$$

Proof. As in the example, we first derive the c.d.f. of Y. For any $y \in \mathbb{R}$, we have

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(g(X) \leq y) = \mathrm{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)),$$

where we used the fact that g is increasing. Differentiating w.r.t. y and an application of the chain rule leads to

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

□

Remark 9.2.3. *Note that in the case when g is strictly* decreasing *and differentiable, we get*

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(g(X) \leq y) = \mathrm{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)),$$

and hence

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)].$$

Since in this case $\frac{d}{dy}[g^{-1}(y)] < 0$, we can also write

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}[g^{-1}(y)]\right|.$$

Remark 9.2.4. *We typically call the term $\left|\frac{d}{dy}[g^{-1}(y)]\right|$ the* Jacobian of the transformation.

Example 9.2.5. *Let $X \sim \mathrm{N}(0,1)$, $g(x) = x^2$ and set $Y = g(X) = X^2$. First we compute the c.d.f. of Y. Clearly, for $y < 0$, we have $F_Y(y) = \mathrm{P}(Y \leq y) = 0$ and hence $f_Y(y) = 0$. Now let $y \geq 0$, then*

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}(X^2 \leq y) = \mathrm{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}). \end{aligned}$$

Differentiating leads to

$$\begin{aligned} f_Y(y) &= \frac{1}{2}y^{-1/2}f_X(\sqrt{y}) - \left(-\frac{1}{2}y^{-1/2}\right)f_X(-\sqrt{y}) \\ &= \frac{1}{2\sqrt{y}}\left[\phi(\sqrt{y}) + \phi(-\sqrt{y})\right] \\ &= \frac{1}{2\sqrt{y}}\left[2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y\right)\right] = \frac{1}{\sqrt{2\pi y}}\exp\left(-\frac{1}{2}y\right), \end{aligned}$$

which is the density of a χ^2_1 random variable.

9.3 Summary

If X is discrete and we want to find the p.m.f. of $Y = g(X)$, then we compute

$$\begin{aligned} p_Y(y) &= \mathrm{P}(Y = y) = \mathrm{P}(g(X) = y) = \mathrm{P}(X \in \{x \in \mathrm{Im}X : g(x) = y\}) \\ &= \sum_{\{x \in \mathrm{Im}X : g(x) = y\}} p_X(x), \end{aligned}$$

see equation (9.1.1). If g is invertible (i.e. bijective), then

$$p_Y(y) = p_X(g^{-1}(y)) \qquad \text{for all } y \in \mathrm{Im}Y.$$

If X is continuous and we want to find the c.d.f. of $Y = g(X)$, then we compute

$$\begin{aligned} F_Y(y) &= \mathrm{P}(Y \leq y) = \mathrm{P}(g(X) \leq y) = \mathrm{P}(X \in \{x \in \mathrm{Im}X : g(x) \leq y\}) \\ &= \int_{\{x \in \mathrm{Im}X : g(x) \leq y\}} f_X(x)dx. \end{aligned}$$

If Y is continuous, we would then differentiate its c.d.f. to obtain its p.d.f..

In the case when g is differentiable and strictly increasing/decreasing, we get

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}[g^{-1}(y)] \right|.$$

You might remember the above formula better when writing $x = g^{-1}(y)$ and noting that

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|,$$

in the case when g is strictly increasing you can remove the absolute value signs and you get the pretty symmetric formula:

$$f_Y(y)dy = f_X(x)dx.$$

End of lecture 13. . . . . . .

Chapter 10

Expectation of random variables

The material of this chapter is based on Blitzstein & Hwang (2019), p.149-174, Grimmett & Welsh (1986), p.29-32, 67-70, 90-92.

This chapter introduces the *expectation* of a random variable. We will distinguish the two cases of a discrete and continuous random variable and study the so-called *law of the unconscious statistician* (LOTUS). We will also learn that the expectation is a *linear* operator and we will introduce the concept of a *variance* and other (higher) *moments*.

10.1 Definition of the expectation

Next we define the expectation of a discrete random variable.

Definition 10.1.1 (Expectation of discrete random variable)**.** *Let X denote a discrete random variable, then the* expectation *of X is defined as*

$$\mathrm{E}(X) = \sum_{x \in \mathrm{Im} X} x \mathrm{P}(X = x)$$

whenever the sum on the right hand side converges absolutely, i.e. when we have $\sum_{x \in \mathrm{Im} X} |x| \mathrm{P}(X = x) < \infty$.[1]

The expectation of X is also called *expected value* or *mean*. Note that we typically simplify the notation and write

$$\mathrm{E}(X) = \sum_{x} x \mathrm{P}(X = x) = \sum_{x} x p_X(x).$$

Definition 10.1.2 (Expectation of a continuous random variable)**.** *For a continuous random variable X with density f_X, we define the* expectation *of X as*

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx,$$

provided that $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$.

As in the discrete case, we often refer to the expectation as *mean* or *expected value*.

Remark 10.1.3. *Recall that we said that $p_X(x)$ for a discrete random variable is comparable to $f_X(x)dx$ for a continuous random variable. Also, in the discrete case, we deal with sums, whereas in the continuous case we have integrals. Using these analogies it makes sense to use the definition*

$$\mathrm{E}(X) = \begin{cases} \sum_x x p_X(x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

[1] This assumption matters in the case when $\mathrm{Im} X$ is infinite. If the sum converges absolutely, then the sum takes the same value irrespectively of the order of summation.

10.2 Law of the unconscious statistician (LOTUS)

Consider the situation that we have a transformation of a random variable $Y = g(X)$ and we would like to find its expectation. The law of the unconscious statistician will tell us that we do not need to find the p.m.f./p.d.f. of the transformed variable but rather use the following formula:

Theorem 10.2.1 (LOTUS: Discrete case)**.** *Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathrm{E}(g(X)) = \sum_{x \in \mathrm{Im}X} g(x)\mathrm{P}(X = x),$$

whenever the sum on the right hand side converges absolutely.

Proof. We note that if $Y = g(X)$, then according to equation (9.1.1) the p.m.f. of Y is given by

$$\mathrm{P}(Y = y) = \mathrm{P}(g(X) = y) = \sum_{x \in \mathrm{Im}X : g(x) = y} \mathrm{P}(X = x).$$

Hence

$$\begin{aligned}\mathrm{E}(Y) &= \sum_y y\mathrm{P}(Y = y) = \sum_y y\mathrm{P}(g(X) = y) = \sum_y y \sum_{x:g(x)=y} \mathrm{P}(X = x) \\ &= \sum_y \sum_{x:g(x)=y} y\mathrm{P}(X = x) = \sum_y \sum_{x:g(x)=y} g(x)\mathrm{P}(X = x) \\ &= \sum_x g(x)\mathrm{P}(X = x).\end{aligned}$$

□

Example 10.2.2. *Consider a Bernoulli random variable $X \sim \mathrm{Ber}(p)$. We compute its mean as follows:*

$$\mathrm{E}(X) = \sum_x x\mathrm{P}(X = x) = 0 \cdot \mathrm{P}(X = 0) + 1 \cdot \mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

Next, we want to find $E(X^2)$. Using Theorem 10.2.1, we find

$$\mathrm{E}(X^2) = \sum_x x^2\mathrm{P}(X = x) = 0^2 \cdot \mathrm{P}(X = 0) + 1^2 \cdot \mathrm{P}(X = 1) = \mathrm{P}(X = 1) = p.$$

We will now state (without proof) the LOTUS for the continuous case:

Theorem 10.2.3 (LOTUS: Continuous case)**.** *Let X be a continuous random variable with density f_X, consider a function $g : \mathbb{R} \to \mathbb{R}$, then*

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx,$$

provided that $\int_{-\infty}^{\infty} |g(x)|f_X(x)dx < \infty$.

Example 10.2.4. *Take $g(x) = x^k$ for $k \in \mathbb{N}$. Then $\mathrm{E}(X^k)$ is called the kth moment of X (provided it exists).*

The LOTUS theorems imply the linearity of the expectation in the following sense:

Theorem 10.2.5. *Consider a discrete/continuous random variable X with finite expectation.*

1. *If X is non-negative, then $\mathrm{E}(X) \geq 0$.*
2. *If $a, b \in \mathbb{R}$, then $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.*

Proof. The proof is left as an exercise, see Exercise 6- 2. □

10.3 Variance

While the expectation tells you something about the centre of the distribution, in many applications we also want to know about the dispersion of X about its mean value. Hence we introduce the so-called *variance*

Definition 10.3.1 (Variance)**.** *Let X be a discrete/continuous random variable. Then its variance is defined as*

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

provided that it exists. Often we write $\sigma^2 = \mathrm{Var}(X)$.

From Theorem 10.2.5 we deduce that the variance of a random variable is always non-negative.

If we are considering a random variable which is just given by a deterministic constant, e.g. for some $c \in \mathbb{R}$ we have $\mathrm{P}(X = c) = 1$, then $\mathrm{E}(X) = \sum_x x\mathrm{P}(X = x) = c \cdot \mathrm{P}(X = c) = c$ and $\mathrm{Var}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{E}[(c-c)^2] = 0$. That means that only true randomness generates a non-zero variance.

In practice, it is often easier to work with a slightly different expression for the variance which we shall derive next.

Theorem 10.3.2. *For a discrete/continuous random variable with finite variance we have that*

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

Proof. In order to simplify the notation we write $\mu = \mathrm{E}(X)$. In the discrete case we have, using Theorem 10.2.1,

$$\begin{aligned}
\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] &= \sum_x (x^2 - 2\mu x + \mu^2) p_X(x) \\
&= \sum_x x^2 p_X(x) + \sum_x (-2\mu x) p_X(x) + \sum_x \mu^2 p_X(x) \\
&= \sum_x x^2 p_X(x) - 2\mu \sum_x x p_X(x) + \mu^2 \sum_x p_X(x) \\
&= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.
\end{aligned}$$

In the continuous case, we have after applying Theorem 10.2.3,

$$\begin{aligned}
\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)dx + \int_{-\infty}^{\infty} (-2\mu x) f_X(x)dx + \int_{-\infty}^{\infty} \mu^2 f_X(x)dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x)dx - 2\mu \int_{-\infty}^{\infty} x f_X(x)dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)dx \\
&= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.
\end{aligned}$$

□

A very useful property of the variance is that it is not affected by deterministic additions and a multiplicative constant can be taken out of the variance provided we square it:

Theorem 10.3.3. *Let X be a discrete/continuous random variable with finite variance and consider deterministic constants $a, b \in \mathbb{R}$. Then*

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X).$$

Proof. This is left as an exercise, see Exercise 6- 3. □

End of lecture 14. . . . . . .

Chapter 11

Multivariate random variables

The material of this chapter is based on Blitzstein & Hwang (2019), p.129-133, 303-306, 312-313, Grimmett & Welsh (1986), p.36-43, 75-88.

11.1 Multivariate distributions and independence

Let us now consider two (arbitrary, i.e. not restricted to discrete or continuous) random variables X and Y on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We would like to understand how they relate to each other and whether or not they are independent. We will write them as a random vector (X, Y) taking values in $\mathbb{R}^2$.

Definition 11.1.1 (Joint distribution function)**.** *The* joint distribution function *of the random vector (X, Y) is defined as the mapping $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by*

$$F_{X,Y}(x, y) = \mathrm{P}(\{\omega \in \Omega : X(\omega) \leq x, Y(\omega) \leq y\}), \qquad \textit{for any } x, y \in \mathbb{R}.$$

Using our shortened notation, we typically write

$$F_{X,Y}(x, y) = \mathrm{P}(X \leq x, Y \leq y), \qquad \text{for any } x, y \in \mathbb{R}.$$

We can now list some of the key properties of joint distribution functions:

- $F_{X,Y}$ is non-increasing in each variable, meaning that
$$F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2) \qquad \text{if } x_1 \leq x_2 \text{ and } y_1 \leq y_2.$$
- We have the following two limits:
$$\lim_{x \to -\infty, y \to -\infty} F_{X,Y}(x, y) = 0, \qquad \lim_{x \to \infty, y \to \infty} F_{X,Y}(x, y) = 1.$$
- They determine the *marginal distributions* uniquely, i.e.
$$F_X(x) = \lim_{y \to \infty} F_{X,Y}(x, y), \qquad F_Y(y) = \lim_{x \to \infty} F_{X,Y}(x, y).$$

We have now all the tools to define what we mean by independence of (general) random variables: We call random variables X and Y independent if the events $\{\omega \in \Omega : X(\omega) \leq x\}$ and $\{\omega \in \Omega : Y(\omega) \leq y\}$ are independent for all $x, y \in \mathbb{R}$. I.e. we define:

Definition 11.1.2 (Independence of random variables)**.** *The random variables X and Y are independent if and only if*

$$\mathrm{P}(X \leq x, Y \leq y) = \mathrm{P}(X \leq x)\mathrm{P}(Y \leq y), \qquad \textit{for all } x, y \in \mathbb{R},$$

which is equivalent to saying that the joint distribution function factorises as the product of the two marginal distribution functions:

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \qquad \textit{for all } x, y \in \mathbb{R}.$$

11.1.1 The n-dimensional case

The extension to the n-dimensional case (for $n \in \mathbb{N}$) is now straightforward: We consider random variables $X_1, \dots, X_n$ on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We write $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{x} = (x_1, \dots, x_n)$. Then the joint distribution function of $\mathbf{X}$ is given by $F_{\mathbf{X}} : \mathbb{R}^n \to [0,1]$:

$$F_{\mathbf{X}}(\mathbf{x}) = \mathrm{P}(X_1 \leq x_1, \dots, X_n \leq x_n), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n.$$

We call the random variables $X_1, \dots, X_n$ *independent* if

$$\mathrm{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathrm{P}(X_1 \leq x_1) \cdots \mathrm{P}(X_n \leq x_n), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

or equivalently if

$$F_{\mathbf{X}}(\mathbf{x}) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

Definition 11.1.3 (Pairwise independence for $n \in \mathbb{N}, n > 2$ random variables)**.** *We call the random variables $X_1, \dots, X_n$* pairwise independent *if*

$$F_{X_i, X_j}(x_i, x_j) = F_{X_i}(x_i) F_{X_j}(x_j), \quad \text{for all } x_i, x_j \in \mathbb{R} \text{ whenever } i \neq j.$$

Remark 11.1.4. *Independence of random variables implies pairwise independence, the reverse statement, however, is not true in general.*

Finally, we define what we mean by independence of a family of (infinitely many) random variables.

Definition 11.1.5 (Independence of a family of random variables)**.** *Let $\mathcal{I} \subset \mathbb{R}$ denote an index set. A family of random variables $\{X_i : i \in \mathcal{I}\}$ is said to be* independent *if for all finite subsets $\mathcal{J} \subseteq \mathcal{I}$ and all $x_j \in \mathbb{R}, j \in \mathcal{J}$, the following product rule holds:*

$$\mathrm{P}(\cap_{j \in \mathcal{J}} \{X_j \leq x_j\}) = \prod_{j \in \mathcal{J}} \mathrm{P}(X_j \leq x_j).$$

11.2 Multivariate discrete distributions and independence

Definition 11.2.1 (Joint probability mass function)**.** *Let X, Y denote discrete random variables on $(\Omega, \mathcal{F}, \mathrm{P})$. Their* joint probability mass function *denoted by $p_{X,Y}$ is defined as the function $p_{X,Y} : \mathbb{R}^2 \to [0,1]$ given by*

$$p_{X,Y}(x,y) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x, Y(\omega) = y\}),$$

which is typically shortened to

$$p_{X,Y}(x,y) = \mathrm{P}(X = x, Y = y).$$

We have that $p_{X,Y}(x,y) \geq 0$ for all $x, y \in \mathbb{R}$ and $\sum_x \sum_y p_{X,Y}(x,y) = 1$.

The marginal probability mass functions *of X and Y are then given by*

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad \text{and } p_Y(y) = \sum_x p_{X,Y}(x,y).$$

It turns out that for any "nice" set $A \subseteq \mathbb{R}^2$, we obtain that

$$\mathrm{P}((X,Y) \in A) = \sum \sum_{(x,y) \in A} \mathrm{P}(X = x, Y = y).$$

11.2.1 Independence

Definition 11.1.2 covers the case of general random variables. In the discrete (or continuous) case, we can formulate equivalent independence conditions:

Definition 11.2.2 (Independence of discrete random variables)**.** *Suppose that X and Y are discrete random variables on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. X and Y are said to be* independent *if the pair of events $\{\omega \in \Omega : X(\omega) = x\}$ and $\{\omega \in \Omega : Y(\omega) = y\}$ are independent for all $x, y \in \mathbb{R}$, i.e. if*

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y), \quad \text{for all } x, y \in \mathbb{R}. \tag{11.2.1}$$

Condition (11.2.1) is equivalent to saying that

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad \text{for all } x, y \in \mathbb{R}.$$

Random variables which are not independent are called *dependent*.

11.3 Multivariate continuous distributions and independence

We can also extend the concept of a continuous random variable to random vectors. Again, we will be focussing on the bivariate case, but the n-dimensional case works in exactly the same way.

Definition 11.3.1 (Continuous random vector)**.** *We call the random vector (X, Y) on $(\Omega, \mathcal{F}, \mathrm{P})$* (jointly) continuous *if*

$$F_{X,Y}(x, y) = \int_{u=-\infty}^{x} \int_{v=-\infty}^{y} f_{X,Y}(u, v)dvdu,$$

for a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ satisfying

(i) $f_{X,Y}(u, v) \geq 0$ *for all* $u, v \in \mathbb{R}$,

(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v)dvdu = 1$.

We call $f_{X,Y}$ the (joint) density *function of (X, Y).*

Similar to the univariate case, we typically obtain the joint density by differentiating the joint distribution function. I.e. we take

$$f_{X,Y}(x, y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y), & \text{if this derivative exists at } (x, y), \\ 0, & \text{otherwise.} \end{cases}$$

It turns out that for any "nice" set $A \subseteq \mathbb{R}^2$, we obtain that

$$\mathrm{P}((X, Y) \in A) = \int \int_{(x,y)\in A} f_{X,Y}(x, y)dxdy.$$

We do not prove this result here formally.

We note that the *marginal densities* can be obtained from the joint density as follows:

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) = \frac{d}{dx} \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f_{X,Y}(u, v)dvdu \\ &= \int_{v=-\infty}^{\infty} f_{X,Y}(x, v)dv, \end{aligned}$$

and also

$$f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u, y)du.$$

11.3.1 Independence

From our definition of independence of random variables, see Definition 11.1.2, we can immediately deduce by differentiating/integrating that jointly continuous random variables X and Y are independent if and only if their joint density factorises:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \text{for all } x,y \in \mathbb{R}.$$

11.3.2 Example

Example 11.3.2. *Suppose the joint density of (X,Y) is given by*

$$f_{X,Y}(x,y) = \begin{cases} \frac{7}{\sqrt{2\pi}} e^{-x^2/2-7y}, & \text{if } -\infty < x < \infty, y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We want to check whether or not X and Y are independent and compute $\mathrm{P}(X > 2, Y < 1)$. We note that for $x \in \mathbb{R}, y > 0$, we can write

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \cdot 7e^{-7y},$$

which is in fact the product of a standard normal random variable and an $\mathrm{Exp}(7)$ random variable. Hence, we have $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x,y \in \mathbb{R}$ which implies independence. Hence

$$\begin{aligned} \mathrm{P}(X > 2, Y < 1) &= \mathrm{P}(X > 2)\mathrm{P}(Y < 1) = (1 - \Phi(2))F_Y(1) \\ &= (1 - \Phi(2))(1 - e^{-7}). \end{aligned}$$

11.4 Transformations of random vectors: The bivariate case [Reading material]

In Chapter we discussed how we can compute p.m.f. and p.d.f. of transformed random variables. Here we will give an outlook on how the methodology works in a multivariate setting. Please read through this section in your own time and, in particular, work through the working example, Example 11.4.1, below.

Consider the case of jointly continuous random variables (X,Y) with density $f_{X,Y}$. Let $u, v : \mathbb{R}^2 \to \mathbb{R}$ denote deterministic functions and define a new pair of random variables by

$$U = u(X,Y), \qquad V = v(X,Y).$$

We would like to find the joint density of (U,V).

We define the mapping $T : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$T(x,y) = (u(x,y), v(x,y)),$$

and assume that T is a bijection from the domain $D = \{(x,y) : f_{X,Y}(x,y) > 0\} \subseteq \mathbb{R}^2$ to some range $S \subseteq \mathbb{R}^2$. Then we can write $T^{-1} : S \to D$ for the inverse mapping of T, i.e. $(x,y) = T^{-1}(u,v)$. For the first component we write $x = x(u,v)$ and for the second $y = y(u,v)$. The *Jacobian* of T^{-1} is defined as the determinant

$$J(u,v) = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Then the joint density of (U,V) is given by

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}(x(u,v), y(u,v))|J(u,v)|, & \text{if } (u,v) \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Example 11.4.1. *Let us demonstrate how the methodology works in practice, see Grimmett & Welsh (1986, p. 87).*

Suppose that $X, Y \sim \text{Exp}(1)$ *are independent. Define* $U := X + Y, V := X/(X + Y)$. *We want to find the joint density of* (U, V) *and the marginal densities of* U *and* V.

First, we note that the joint density of (X, Y) *is –due to independence– given by* $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ *for all* x, y, *hence*

$$f_{X,Y}(x, y) = e^{-(x+y)}, \text{ if } x, y > 0,$$

and zero otherwise.

In our case, the mapping T *is given by*

$$T(x, y) = (u, v) = (x + y, x/(x + y)),$$

where T *maps the set*

$$D = \{(x, y) : x, y > 0\},$$

onto

$$S = \{(u, v) : 0 < u < \infty, 0 < v < 1\}.$$

Next, we find the inverse function of T:

$$T^{-1}(u, v) = (x, y) = (uv, (1 - v)u).$$

The Jacobian of T^{-1} *is given by*

$$J(u, v) = \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} v & u \\ (1 - v) & -u \end{pmatrix} = -uv - (1 - v)u = -u.$$

Then, for $(u, v) \in S$, *we have*

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v))|J(u, v)| = \exp(-(uv + (1 - v)u))| - u| = u \exp(-u),$$

and zero otherwise. I.e.

$$f_{U,V}(u, v) = \begin{cases} ue^{-u}, & \text{for } u > 0, 0 < v < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The marginal density of U *is given by*

$$f_U(u) = \int_0^1 f_{U,V}(u, v)dv = \int_0^1 u \exp(-u)dv = u \exp(-u), \text{ for } u > 0,$$

and zero otherwise. Hence, $U \sim \text{Gamma}(2, 1)$. *The marginal density of* V *is given by*

$$f_V(v) = \int_0^\infty f_{U,V}(u, v)du = \int_0^\infty u \exp(-u)du = \Gamma(2) = 1, \text{ for } 0 < v < 1,$$

and zero otherwise. Hence $V \sim \text{U}(0, 1)$.

We notice that $f_{U,V}(u, v) = f_U(u) f_V(v)$ *for all* u, v, *which implies that* U *and* V *are independent.*

End of lecture 15.

11.5 Two dimensional law of the unconscious statistician (2D LOTUS)

Using exactly the same arguments as in the univariate case, we can also formulate a law of the unconscious statistician applied to a function of a (bivariate) random vector. We will only state the result here.

Theorem 11.5.1 (2D LOTUS: discrete case)**.** *Let X, Y denote discrete random variables on $(\Omega, \mathcal{F}, \mathrm{P})$ and let $g : \mathbb{R}^2 \to \mathbb{R}$. Then $Z = g(X, Y)$ is also a discrete random variable on $(\Omega, \mathcal{F}, \mathrm{P})$ and its expectation is given by*

$$\mathrm{E}(g(X,Y)) = \sum_{x \in \mathrm{Im}X} \sum_{y \in \mathrm{Im}Y} g(x,y)\mathrm{P}(X = x, Y = y).$$

Theorem 11.5.2 (2D LOTUS: continuous case)**.** *Let X, Y be jointly continuous random variables with density $f_{X,Y}$ and let $h : \mathbb{R}^2 \to \mathbb{R}$. Then*

$$\mathrm{E}[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y) dxdy.$$

Theorems 11.5.1,11.5.2 can be used to prove the linearity of the expectation:

Theorem 11.5.3 (Linearity of expectation)**.** *Let X, Y denote jointly discrete/continuous random variables on $(\Omega, \mathcal{F}, \mathrm{P})$, and $a, b \in \mathbb{R}$, then*

$$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y),$$

provided that $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ exist.

Proof. In the discrete case, we apply Theorem 11.5.1 with $g(x, y) = ax + by$. Then

$$\begin{aligned}
\mathrm{E}(aX + bY) &= \sum_x \sum_y (ax + by)\mathrm{P}(X = x, Y = y) \\
&= a \sum_x x \sum_y \mathrm{P}(X = x, Y = y) + b \sum_y y \sum_x \mathrm{P}(X = x, Y = y) \\
&= a \sum_x x\mathrm{P}(X = x) + b \sum_y y\mathrm{P}(Y = y) \\
&= a\mathrm{E}(X) + b\mathrm{E}(Y).
\end{aligned}$$

In the continuous case, we apply Theorem 11.5.2 with $g(x, y) = ax + by$. Then

$$\begin{aligned}
\mathrm{E}(aX + bY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{X,Y}(x,y) dxdy \\
&= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dxdy + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dxdy \\
&= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} y f_Y(y) dy \\
&= a\mathrm{E}(X) + b\mathrm{E}(Y).
\end{aligned}$$

□

Using induction, one can easily deduce that for $n \in \mathbb{N}$ and random variables $X_1, \ldots, X_n$ with finite expectations and constants $a_1, \ldots, a_n \in \mathbb{R}$ we have

$$\mathrm{E}(a_1X_1 + \cdots + a_nX_n) = a_1\mathrm{E}(X_1) + \cdots + a_n\mathrm{E}(X_n). \tag{11.5.1}$$

Remark 11.5.4. *It is important to remember that the linearity of the expectation* (11.5.1) *holds in general without assuming any independence between the random variables.*

11.6 Covariance and correlation between random variables

Definition 11.6.1. *Consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$. The* covariance *of X and Y is defined as*

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

if the expectation on the right hand side takes a finite value.

Also, $\mathrm{COR}(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$

When we set $X = Y$, then the covariance simplifies to the variance:

$$\mathrm{Cov}(X, X) = \mathrm{E}[(X - \mu_X)^2] = \mathrm{Var}(X).$$

For concrete computations it is often useful to work with the following alternative expression for the covariance.

Theorem 11.6.2 (Covariance)**.** *For jointly discrete/continuous random variables X, Y with finite expectations, we have*

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

Proof. See Exercise 6- 9. □

Remark 11.6.3. *It is important to note that independent random variables always have zero covariance, but the converse does not hold in general!*

A very important property of independent random variables is the fact that the expectation of their product can be written as the product of their expectation (a property which does not hold in general!):

Theorem 11.6.4. *Let X, Y denote independent and jointly discrete/continuous random variables with finite expectation, then*

$$\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y). \tag{11.6.1}$$

Proof. In the case when X, Y are jointly discrete, we use Theorem 11.5.1 with $g(x, y) = xy$. Then

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_x \sum_y xy\mathrm{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy\mathrm{P}(X = x)\mathrm{P}(Y = y) \quad \text{(by independence)} \\
&= \sum_x x\mathrm{P}(X = x) \sum_y y\mathrm{P}(Y = y) \quad \text{(using the existence of } \mathrm{E}(X), \mathrm{E}(Y)) \\
&= \mathrm{E}(X)\mathrm{E}(Y).
\end{aligned}$$

Using Theorem 11.5.2 and similar computations as above gives us the result for the jointly continuous case. □

Remark 11.6.5. *It is important to note that if $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, then this does not in general imply that X and Y are independent, see Exercise 6- 6.*

The results stated in Theorem 11.6.4 can be extended to the $n \in \mathbb{N}$ dimensional case by induction: If $X_1, \ldots, X_n$ are independent, then

$$\mathrm{E}(X_1 \cdots X_n) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_n).$$

In statistics, we often deal with sums of random variables. How can we compute their variance? The following theorem gives an answer.

Theorem 11.6.6 (Variance of a sum of random variables)**.** *Let X, Y denote two jointly discrete/continuous random variables with finite variances. Then*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

Proof. See Exercise 6- 10. □

End of lecture 16. .....

Chapter 12

Generating functions

The material of this chapter is based on Blitzstein & Hwang (2019), p.279-293, Grimmett & Welsh (1986), p.45-52.

In probability theory we often use so-called generating functions to derive/prove statements regarding the distribution of random variables/vectors or to compute moments. In this course, we study so-called *probability generating functions* and *moment generating functions* and we will give an outlook on what *characteristic functions* are.

12.1 Probability generating functions

First of all, we introduce so-called *probability generating functions* and explain why they are extremely useful!

Throughout this section, we will only consider **discrete random variables** taking values in the **non-negative integers**, i.e. $\mathrm{Im}X \subseteq \mathbb{N} \cup \{0\} = \{0, 1, 2, \dots\}$.

Definition 12.1.1 (Probability generating function (p.g.f.))**.** *Let X denote a discrete random variable with $\mathrm{Im}X \subseteq \mathbb{N} \cup \{0\}$. We denote by*

$$\mathcal{S}_X = \left\{ s \in \mathbb{R} : \sum_{x=0}^{\infty} |s|^x \mathrm{P}(X = x) < \infty \right\}.$$

Then the probability generating function *(pgf) of X is defined as the function $G_X : \mathcal{S}_X \to \mathbb{R}$ given by*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x \mathrm{P}(X = x).$$

We observe that the pgf is well-defined for $|s| \leq 1$ since

$$\sum_{x=0}^{\infty} |s|^x \mathrm{P}(X = x) \leq \sum_{x=0}^{\infty} \mathrm{P}(X = x) = 1 < \infty.$$

Also, $G_X(0) = \mathrm{P}(X = 0)$ and $G_X(1) = 1$.

The reason why probability generating functions are extremely useful is that they uniquely determine the probability mass function (i.e. the distribution) of a discrete random variable:

Theorem 12.1.2. *Let X, Y denote discrete random variables with $\mathrm{Im}X, \mathrm{Im}Y \subseteq \mathbb{Z} \cup \{0\}$. Their p.g.f.s are denoted by G_X and G_Y, respectively. Then*

$$G_X(s) = G_Y(s), \text{ for all } s \in \mathcal{S}_X \cap \mathcal{S}_Y, \tag{12.1.1}$$

if and only if

$$\mathrm{P}(X = x) = \mathrm{P}(Y = x), \text{ for all } x = 0, 1, 2, \dots. \tag{12.1.2}$$

Proof. Assume that (12.1.1) holds. First we note that $G_X(1) = G_Y(1)$ implies $\mathrm{P}(X = 0) = \mathrm{P}(Y = 0)$. When we differentiate[1] the pgfs we get

$$G'_X(s) = \sum_{x=1}^{\infty} x s^{x-1} \mathrm{P}(X = x).$$

When we plug in $s = 0$ in the first derivative, we get $G'_X(0) = \mathrm{P}(X = 1)$. Hence, $G'_X(0) = \mathrm{P}(X = 1) = \mathrm{P}(Y = 1) = G'_Y(0)$. This procedure can be repeated and we obtain

$$\frac{d^n}{ds^n} G_X(s)\Big|_{s=0} = n! \mathrm{P}(X = n),$$

the same can be done for G_Y, and the identity of the two pgfs implies the result.

The other direction of the proof is trivial. □

12.1.1 Common probability generating functions

We will now list the p.g.f.s of some common discrete distributions.

Example 12.1.3 (Bernoulli distribution)**.** *Let $X \sim \mathrm{Ber}(p)$. Then*

$$G_X(s) = \mathrm{E}(s^X) = s^0 \mathrm{P}(X = 0) + s^1 \mathrm{P}(X = 1) = 1 - p + sp$$

for all $s \in \mathbb{R}$.

Example 12.1.4 (Binomial distribution)**.** *Let $X \sim \mathrm{Bin}(n, p)$. Then*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{n} \binom{n}{x} s^x p^x (1-p)^{n-x} = (1 - p + sp)^n,$$

for all $s \in \mathbb{R}$, by an application of the binomial theorem.

Example 12.1.5 (Poisson distribution)**.** *Let $X \sim \mathrm{Poi}(\lambda)$. Then*

$$G_X(s) = \mathrm{E}(s^X) = \sum_{x=0}^{\infty} s^x \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(s\lambda)^x}{x!} = e^{-\lambda} e^{s\lambda} = \exp(\lambda(s-1)),$$

for all $s \in \mathbb{R}$. Here we used the series expansion of the exponential function.

12.1.2 Probability generating function of a sum of independent discrete random variables

Theorem 12.1.6. *Let X, Y be independent discrete random variables with $\mathrm{Im}X, \mathrm{Im}Y \subseteq \mathbb{N} \cup \{0\}$. Then*

$$G_{X+Y}(s) = G_X(s) G_Y(s), \text{ for all } s \in \mathcal{S}_X \cap \mathcal{S}_Y.$$

Proof. Let $s \in \mathcal{S}_X \cap \mathcal{S}_Y$. Since X and Y are independent, Exercise 6- 7 implies that s^X and s^Y are also independent. Hence using Theorem 11.6.4 we conclude that

$$G_{X+Y}(s) = \mathrm{E}(s^{X+Y}) = \mathrm{E}(s^X s^Y) = \mathrm{E}(s^X)\mathrm{E}(s^Y) = G_X(s) G_Y(s).$$

□

An immediate consequence of the above results is, that for independent non-negative integer-valued random variables $X_1, \ldots, X_n$ ($n \in \mathbb{N}$), we have

$$G_{\sum_{i=1}^n X_i}(s) = \prod_{i=1}^{n} G_{X_i}(s),$$

for all $s \in \cap_{i=1}^n \mathcal{S}_{X_i}$.

[1] You will learn in the real analysis course under which conditions we are allowed to interchange the infinite sum and the derivative. For the purpose of this course, we will just assume that the above computation is valid.

12.1.3 Moments

We have already introduced, the mean and variance of a (discrete) random variable X. More generally, for $k \in \mathbb{N}$, we call $\mathrm{E}(X^k)$ the kth moment of X provided it exists. It turns out that we can use the probability generating function for deriving moments of random variables. More precisely, we differentiate the pgf k times and plug in $s = 1$:

Theorem 12.1.7. *Let X be a discrete random variable with $\mathrm{Im}X \in \mathbb{N} \cup \{0\}$. Let $k \in \mathbb{N}$. Then the kth derivative of the pgf is given by*

$$\frac{d^k}{ds^k} G_X(s)\Big|_{s=1} = G_X^{(k)}(1) = \mathrm{E}[X(X-1)\cdots(X-k+1)].$$

Proof. As before, we assume that we are allowed to interchange derivatives and summation under suitable conditions. Then

$$\frac{d}{ds} G_X(s) = \frac{d}{ds}\mathrm{E}(s^X) = \mathrm{E}(X s^{X-1}).$$

Hence

$$\frac{d}{ds} G_X(s)\Big|_{s=1} = \mathrm{E}(X).$$

Similarly,

$$\frac{d^k}{ds^k} G_X(s) = \frac{d}{ds}\mathrm{E}(s^X) = \mathrm{E}[X(X-1)\cdots(X-k+1)s^{X-k}].$$

Hence

$$\frac{d^k}{ds^k} G_X(s)\Big|_{s=1} = \mathrm{E}[X(X-1)\cdots(X-k+1)].$$

□

Example 12.1.8 (Computing the variance using pgfs)**.** *The above theorem can be used for computing the variance of a discrete non-negative integer-valued random variable X. We note that*

$$G_X''(1) = \mathrm{E}[X(X-1)] = \mathrm{E}(X^2) - \mathrm{E}(X).$$

Hence,

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = G_X''(1) + G_X'(1) - (G_X'(1))^2.$$

End of lecture 17.

12.2 Moment generating functions

Definition 12.2.1 (Moment generating functions)**.** *Let X be a random variable. Then its* moment generating function *(m.g.f.) is defined as*

$$M_X(t) = \mathrm{E}(e^{tX}),$$

provided the expectation exists in some neighbourhood of zero, i.e. the expectation exists for all $|t| < \epsilon$ for some $\epsilon > 0$.

We compute the m.g.f. as follows:

$$M_X(t) = \mathrm{E}(e^{tX}) = \begin{cases} \sum_x e^{tx} p_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, & \text{if } X \text{ is continuous,} \end{cases}$$

whenever the sum/integral is absolutely convergent.

Example 12.2.2. *Let $X \sim \mathrm{N}(0,1)$. Then we use the trick of "completing the square" in the second line:*

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2 + tx} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2tx + t^2)} e^{\frac{t^2}{2}} dx \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-t)^2} dx = e^{\frac{t^2}{2}}, \end{aligned}$$

since the latter integral is equal to 1 since it is the integral of a $\mathrm{N}(t,1)$ density function. We note that the m.g.f. exists for all $t \in \mathbb{R}$ in this case.

Let us now consider an example where the m.g.f. does not exist for all $t \in \mathbb{R}$.

Example 12.2.3. *Let $X \sim \mathrm{Exp}(\lambda)$, then*

$$\begin{aligned} M_X(t) = \mathrm{E}(e^{tX}) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx = \int_0^{\infty} e^{(t-\lambda)x} \lambda dx \\ &= \begin{cases} \frac{\lambda}{\lambda - t}, & \text{if } t < \lambda, \\ \infty, & \text{if } t \geq \lambda. \end{cases} \end{aligned}$$

12.2.1 Properties

Theorem 12.2.4. *If X has a m.g.f., then for $k \in \mathbb{N}$, the kth moment of X is given by*

$$\mathrm{E}(X^k) = M_X^{(k)}(0) = \left. \frac{d^k}{dt^k} M_X(t) \right|_{t=0}.$$

Proof. We give a sketch proof of the theorem. Assuming we can interchange expectation and differentiation, we write

$$\frac{d^k}{dt^k} M_X(t) = \frac{d^k}{dt^k} \mathrm{E}(e^{tX}) = \mathrm{E}\left(\frac{d^k}{dt^k} e^{tX} \right) = \mathrm{E}\left(X^k e^{tX} \right),$$

and then plug in $t = 0$. □

Theorem 12.2.5. *For $a, b \in \mathbb{R}$, we have $M_{aX+b} = e^{bt} M_X(at)$.*

Proof.

$$M_{aX+b}(t) = \mathrm{E}\{\exp[t(aX+b)]\} = e^{tb} \mathrm{E}[\exp(taX)] = e^{tb} M_X(at).$$

□

Example 12.2.6. *Let $X \sim \mathrm{N}(0,1)$. Let $\mu \in \mathbb{R}, \sigma > 0$, then*

$$M_{\mu+\sigma X} = e^{t\mu} M_X(\sigma t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \sigma^2 t^2/2},$$

which is the m.g.f. of an $\mathrm{N}(\mu, \sigma^2)$ distributed random variable.

Theorem 12.2.7. *Let $X_1, \ldots, X_n$ denote a sequence of independent random variables with m.g.f.s $M_{X_1}, \ldots, M_{X_n}$. Then*

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t).$$

Proof. Since the expectation of a product of independent random variables is the product of their corresponding expectation, see Theorem 11.6.4, we have

$$M_{\sum_{i=1}^n X_i}(t) = \mathrm{E}\left[\exp\left(t\sum_{i=1}^n X_i\right)\right] = \mathrm{E}\left[\prod_{i=1}^n \exp(tX_i)\right] = \prod_{i=1}^n \mathrm{E}(e^{tX_i}) = \prod_{i=1}^n M_{X_i}(t).$$

□

We will now state without proof the famous characterisation theorem:

Theorem 12.2.8 (Characterisation)**.** *If the m.g.f.s of the random variables X and Y exist and $M_X(t) = M_Y(t)$ in a neighbourhood of zero, then*

$$F_X(u) = F_Y(u) \qquad \textit{for all } u.$$

The above theorem states that m.g.f. characterise the distribution of a random variable uniquely.

12.3 Outlook: Characteristic function and Laplace transform

Note that moment generating functions do not exist for all distributions.

Hence we often work with the *characteristic function* of a random variable X instead which is defined as

$$\phi_X(t) = \mathrm{E}(e^{itX}), \qquad \text{for all } t \in \mathbb{R},$$

$$= E[\cos(tX) + i\sin(tX)]$$

where $i = \sqrt{-1}$. It turns out that characteristic functions exist indeed for all distributions and hence they are a useful tool for general proofs in probability theory. However, we will defer the detailed discussion of complex-valued objects to a later probability/statistics (and analysis!) course.

For a non-negative random variable X we sometimes work with the *Laplace transform* instead which is defined as

$$\mathcal{L}_X(t) = \mathrm{E}(e^{-tX}), \qquad \text{for all } t \geq 0.$$

We note that $\mathcal{L}_X(t) = M_X(-t)$ for $t \geq 0$.

Example 12.3.1. *Let $X \sim \mathrm{Exp}(\lambda)$, then, for $t \geq 0$,*

$$\mathcal{L}_X(t) = \mathrm{E}(e^{-tX}) = \int_0^\infty e^{-tx}\lambda e^{-\lambda x}dx = \int_0^\infty e^{-(t+\lambda)x}\lambda dx$$
$$= \frac{\lambda}{\lambda + t}.$$

End of lecture 18. .....

Chapter 13

Conditional distribution and conditional expectation

The material of this chapter is based on Blitzstein & Hwang (2019), p.306-311, 313-321, Grimmett & Welsh (1986), p.32-33, 88-89, 92-95.

Let us now study *conditional distributions* both for discrete and continuous random variables. They allow us to define the *conditional expectation*, which is a really useful concept as we shall see when stating the *law of total expectation*.

13.1 Discrete case: Conditional expectation and the law of total expectation

We have already introduced the notation of conditional probabilities. Now we are going to define the conditional distribution of a discrete random variable.

Definition 13.1.1 (Conditional distribution and conditional expectation)**.** *Let X denote a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Consider an event $B \in \mathcal{F}$ such that $\mathrm{P}(B) > 0$. The* conditional distribution of X given B *is defined as*

$$\mathrm{P}(X = x|B) = \frac{\mathrm{P}(\{X = x\} \cap B)}{\mathrm{P}(B)}, \text{ for } x \in \mathbb{R}.$$

Further, the conditional expectation of X given B *is defined as*

$$\mathrm{E}(X|B) = \sum_{x \in \mathrm{Im}X} x\mathrm{P}(X = x|B),$$

provided the sum is absolutely convergent.

Similarly to the ideas presented in the law of total probability, it can often be useful to consider a partition of the probability space to compute an (unconditional) expectation via conditional expectations as we describe in the following theorem.

Theorem 13.1.2. *[Law of total expectation] Consider a partition $\{B_i : i \in \mathcal{I}\}$ of Ω with $\mathrm{P}(B_i) > 0$ for all $i \in \mathcal{I}$. Let X denote a discrete random variable. Then*

$$\mathrm{E}(X) = \sum_{i \in \mathcal{I}} \mathrm{E}(X|B_i)\mathrm{P}(B_i),$$

whenever the sum converges absolutely.

Proof. First we use the definition of the expectation, followed by the law of total probability (Theorem 5.4.4):

$$\begin{aligned}\mathrm{E}(X) &= \sum_x x\mathrm{P}(X=x) = \sum_x x \sum_{i\in\mathcal{I}} \mathrm{P}(X=x|B_i)\mathrm{P}(B_i)\\ &= \sum_{i\in\mathcal{I}} \mathrm{P}(B_i) \sum_x x\mathrm{P}(X=x|B_i) = \sum_{i\in\mathcal{I}} \mathrm{P}(B_i)\mathrm{E}(X|B_i).\end{aligned}$$

We use the fact that the series is absolutely convergent to justify that we are allowed to change the order of summation. □

13.1.1 Conditioning on a random variable

Suppose (X, Y) are jointly discrete random variables. In the above definition, consider the event $B = \{X = x\}$ for some $x \in \mathbb{R}$ such that $p_X(x) = \mathrm{P}(X = x) > 0$. Then the *conditional distribution/probability mass function of* Y *given* $X = x$ is given by

$$p_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)}, \quad \text{for } y \in \mathbb{R}.$$

Also, the *conditional expectation of* Y *given* $X = x$ is given by

$$\mathrm{E}(Y|X = x) = \sum_y y p_{Y|X}(y|x),$$

provided the sum is absolutely convergent.

Also, the LOTUS for conditional expectations says that

$$\mathrm{E}(g(Y)|X = x) = \sum_y g(y) p_{Y|X}(y|x).$$

Note that we can also formulate an independence condition in terms of conditional p.m.f.s: Discrete X and Y are independent if and only if

$$\mathrm{P}(Y = y|X = x) = \mathrm{P}(Y = y)$$

for all x, y such that $\mathrm{P}(X = x) > 0$. Also, we get a Bayes' type result of the form

$$p_{Y|X}(y|x)p_X(x) = p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y),$$

for all x, y for which $p_X(x), p_Y(y) > 0$.

13.1.2 Example

Let us study an example:

Example 13.1.3. *Suppose you sit in Heathrow waiting for your flight to go on your well deserved holiday. You denote by* N *the total (random) number of planes arriving while you wait and you assume that, for some* $\lambda > 0$, $N \sim \mathrm{Poi}(\lambda)$. *Each plane, independently, turns out to be a British Airways plane with probability* $p \in (0, 1)$, *hence with probability* $1 - p$ *it will be a plane from another airline. We write* $N = X + Y$ *where* X *represents the number of British Airways planes and* Y *the number of planes from other airlines. You are wondering what might be the joint probability mass function of* X *and* Y.

You recall that the Bernoulli distribution describes binary outcomes with success probability p. *So, every time a plane you observe turns out to be a British Airways plane, you view this as a success and a failure otherwise.*

You recall that the number of success given the total number of trials follows a Binomial distribution, so more precisely, in your case you have for $n \in \mathbb{N}$

$$X|N = n \sim \text{Bin}(n, p), \text{ and } Y|N = n \sim \text{Bin}(n, 1-p).$$

Given this information, you try to compute $\text{P}(X = x, Y = y)$*. For this, it would be really useful to know* N*, so let us apply the law of total probability given information on* N*. For* $x, y \in \mathbb{N} \cup \{0\}$*:*

$$\text{P}(X = x, Y = y) = \sum_{n=0}^{\infty} \text{P}(X = x, Y = y|N = n)\text{P}(N = n).$$

Clearly $\text{P}(X = x, Y = y|N = n) > 0 \Leftrightarrow x + y = n$*. So, in the sum, we can get rid off all the terms which result in conditional probabilities being equal to 0.*

$$\begin{aligned}\text{P}(X = x, Y = y) &= \sum_{n=0}^{\infty} \text{P}(X = x, Y = y|N = n)\text{P}(N = n) \\ &= \sum_{n:x+y=n} \text{P}(X = x, Y = y|N = n)\text{P}(N = n) \\ &= \text{P}(X = x, Y = y|N = x + y)\text{P}(N = x + y).\end{aligned}$$

Conditional on the event that $\{N = x + y\}$*, the events* $\{X = x\}$ *and* $\{Y = y\}$ *contain exactly the same information, hence we get*

$$\text{P}(X = x, Y = y|N = x + y)\text{P}(N = x + y) = \text{P}(X = x|N = x + y)\text{P}(N = x + y).$$

It remains to plug in the Binomial and Poisson p.m.f.s:

$$\begin{aligned}\text{P}(X = x, Y = y) &= \text{P}(X = x|N = x + y)\text{P}(N = x + y) \\ &= \binom{x+y}{x} p^x (1-p)^y \frac{\lambda^{x+y}}{(x+y)!} e^{-\lambda} \\ &= \frac{(x+y)!}{x!y!} p^x (1-p)^y \frac{\lambda^{x+y}}{(x+y)!} e^{-\lambda} \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p} \cdot \frac{(1-p)^y \lambda^y}{y!} e^{-\lambda(1-p)},\end{aligned}$$

which is in fact the product of the p.m.f. of a $\text{Poi}(p\lambda)$ *and a* $\text{Poi}((1-p)\lambda)$ *random variable. Hence, we conclude that* X *and* Y *are independent and* $X \sim \text{Poi}(p\lambda)$ *and* $Y \sim \text{Poi}((1-p)\lambda)$*.*

13.2 Continuous case: Conditional density, conditional distribution and conditional expectation

Let us now consider two jointly continuous random variables (X, Y). We cannot proceed as above to define the conditional distribution $\text{P}(Y \leq y|X = x)$ since we now have that $\text{P}(X = x)$ for all $x \in \mathbb{R}$. Hence we need to condition on an event with non-zero probability. Let $\epsilon > 0$, then we have

$$\begin{aligned}\text{P}(Y \leq y|x \leq X \leq x + \epsilon) &= \frac{\text{P}(Y \leq y, x \leq X \leq x + \epsilon)}{\text{P}(x \leq X \leq x + \epsilon)} \\ &= \frac{\int_{u=x}^{x+\epsilon} \int_{v=-\infty}^{y} f_{X,Y}(u, v) du dv}{\int_{x}^{x+\epsilon} f_X(u) du}.\end{aligned}$$

Now we let $\epsilon \to 0$ and get

$$\lim_{\epsilon\to 0} \mathrm{P}(Y \le y | x \le X \le x+\epsilon) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)} = \int_{-\infty}^{y} \frac{f_{X,Y}(x,v)dv}{f_X(x)} = G(y).$$

So G is a distribution function with density function

$$g(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \text{ for } y \in \mathbb{R}.$$

The derivations above only work in the case when $f_X(x) > 0$. Let us now state our formal definition:

Definition 13.2.1 (Conditional distribution and conditional density)**.** *For two jointly continuous random variables X, Y, we define the* conditional density of Y given $X = x$ *as*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{13.2.1}$$

for all $y \in \mathbb{R}$ and for all $x \in \mathbb{R}$ for which $f_X(x) > 0$. The corresponding conditional distribution function of Y given $X = x$ *is then given by*

$$F_{Y|X=x}(y|x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)},$$

for all $y \in \mathbb{R}$ and for all $x \in \mathbb{R}$ for which $f_X(x) > 0$.

Note that we can now also formulate an independence condition in terms of conditional p.d.f.s: Jointly continuous random variables X and Y are independent if and only if

$$f_{Y|X}(y|x) = f_Y(y)$$

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

Remark 13.2.2. *Note that* (13.2.1) *also implies a Bayes' type formula:*

$$f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y),$$

provided that $f_X(x), f_Y(y) > 0$.

End of lecture 19. .....

Similarly to the discrete case, we can now define the conditional expectation and formulate the law of total expectation.

Definition 13.2.3 (Conditional expectation)**.** *For two jointly continuous random variables X, Y, we define the* conditional expectation of Y given $X = x$ *as*

$$\mathrm{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)}dy,$$

provided that $f_X(x) > 0$.

Recall that, in the discrete case, Theorem 13.1.2 implies that for jointly discrete random variables X, Y, we have

$$\mathrm{E}(Y) = \sum_{x:\mathrm{P}(X=x)>0} \mathrm{E}(Y|X=x)\mathrm{P}(X=x),$$

whenever the sum converges absolutely. The continuous analogue reads as follows:

Theorem 13.2.4 (Law of total expectation)**.** *For jointly continuous random variable X, Y we have*

$$\mathrm{E}(Y) = \int_{\{x: f_X(x)>0\}} \mathrm{E}(Y|X=x) f_X(x) dx.$$

Proof. We use the definition of the expectation, the fact that the marginal density of Y can be obtained by integrating out the joint density and equation (13.2.1):

$$\begin{aligned}
\mathrm{E}(Y) &= \int y f_Y(y) dy = \int\int y f_{X,Y}(x,y) dx dy \\
&= \int\int y f_{Y|X}(y|x) f_X(x) dx dy \\
&= \int \left(\int y f_{Y|X}(y|x) dy \right) f_X(x) dx \\
&= \int \mathrm{E}(Y|X=x) f_X(x) dx,
\end{aligned}$$

where we assume that the integrals range over the appropriate values for x and y. □

13.2.1 Example

Example 13.2.5. *Let $\rho \in (-1,1)$. The standard bivariate normal distribution has joint density given by*

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(x^2 - 2\rho xy + y^2\right) \right]$$

for $x, y \in \mathbb{R}$. We want to demonstrate some of the concepts introduced earlier.

1. *What is the marginal density of X? We compute*

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(x^2 - 2\rho xy + y^2\right) \right] dy \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left[(y-\rho x)^2 + x^2\left(1-\rho^2\right)\right] \right] dy \\
&= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(y-\rho x)^2}{2(1-\rho^2)} \right] dy.
\end{aligned}$$

   We observe that the integrand in the above integral is the density of an $\mathrm{N}(\rho x, 1-\rho^2)$ random variable and hence the integral equals 1. Hence

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

   I.e. $X \sim \mathrm{N}(0,1)$ (and also $Y \sim \mathrm{N}(0,1)$).

2. *What is the conditional density of Y given $X = x$? We have*

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(x^2 - 2\rho xy + y^2\right) \right] \sqrt{2\pi} e^{x^2/2}
\end{aligned}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right) + x^2/2\right]$$
$$= \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,(1-\rho^2)}\left(x^2 - 2\rho xy + y^2 - x^2(1-\rho^2)\right)\right]$$
$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(y-\rho x)^2}{2\,(1-\rho^2)}\right].$$

This is in fact the density of an $\mathrm{N}(\rho x, 1-\rho^2)$ *random variable.*

3. *What is the conditional expectation of* Y *given* $X = x$*?*

   Using the definition of the conditional expectation, we have

$$\mathrm{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dx = \rho x,$$

   given our above finding that $Y|X = x \sim \mathrm{N}(\rho x, 1-\rho^2)$*.*

4. *Formulate a condition which ensures that* X *and* Y *are independent.*

   We know that X *and* Y *are independent, if and only if* $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ *for all* $x, y \in \mathbb{R}$*. For any* $x, y \in \mathbb{R}$ *we have*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$
$$\iff$$
$$\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right] = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$
$$\iff$$
$$\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\,(1-\rho^2)}\left(x^2 - 2\rho xy + y^2\right)\right] = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}$$
$$\iff \rho = 0.$$

   So X *and* Y *are independent if and only if* $\rho = 0$*.*

5. *Find the covariance between* X *and* Y*.*

   We note that since $\mathrm{E}(X) = 0 = \mathrm{E}(Y)$*, we have that*

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(XY) = \int_{-\infty}^{\infty} \mathrm{E}(XY|X=x) f_X(x)dx,$$

   where we used the law of the total expectation, see Theorem 13.2.4. Note that $\mathrm{E}(XY|X=x) = \mathrm{E}(xY|X=x) = x\mathrm{E}(Y|X=x) = \rho x^2$*, hence*

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty} \rho x^2 f_X(x)dx = \rho \int_{-\infty}^{\infty} x^2 f_X(x)dx = \rho\mathrm{E}(X^2)$$
$$= \rho\mathrm{Var}(X) = \rho.$$

   So, we have

$$\rho = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y),$$

   which, with our findings above, implies the following important result: **Assume that** X, Y **follow a bivariate (standard) normal distribution.** *Then* X *and* Y *are independent if and only if* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$*.* **Warning:** *As soon as you drop the assumption that you are dealing with jointly normal random variables, then we only know that if we assume that they are independent, then the product formula for the expectations holds. However, if we have only verified that the product formula for the expectations holds, then that does not imply in general independence of the random variables.*

End of lecture 20. . . . . . .

Bibliography

Anderson, D. F., Seppäläinen, T. & Valkó, B. (2018), *Introduction to probability*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge.

Blitzstein, J. K. & Hwang, J. (2019), *Introduction to probability*, Texts in Statistical Science Series, CRC Press, Boca Raton, FL. Second edition.

Feller, W. (1957), *An introduction to probability theory and its applications. Vol. I*, John Wiley and Sons, Inc., New York; Chapman and Hall, Ltd., London. 2nd ed.

Grimmett, G. & Welsh, D. (1986), *Probability: an introduction*, Oxford Science Publications, The Clarendon Press, Oxford University Press, New York.

Hájek, A. (2012), Interpretations of probability, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Winter 2012 edn, Metaphysics Research Lab, Stanford University.

Proschan, M. A. & Shaw, P. A. (2016), *Essentials of probability theory for statisticians*, Chapman & Hall/CRC Texts in Statistical Science Series, CRC Press, Boca Raton, FL.

Ross, S. (2014), *A first course in probability*, ninth edn, Macmillan Co., New York; Collier Macmillan Ltd., London.

Spiegelhalter, D., Pearson, M. & Short, I. (2011), 'Visualizing uncertainty about the future', *Science* **333**(6048), 1393–1400.