C:_Teaching\Topics\Prob.ifloresc.pdf

Math 611 Probability

Instructor: Ionut Florescu
Office: Kidde 227 *Email:* ifloresc@stevens.edu
Phone: (201) 216-5452 *Office hours:* M 4:00pm -6:00pm and by appt.
website: http://www.math.stevens.edu/~ifloresc/Teaching/2009-2010/index611.html

Some Topics to be presented:

Elements of Probability Measure, Conditional Probability and Independence, Random Variables and Distributions, Conditional Distribution and Conditional Expectation, The Poisson Process, Generating Functions and their applications, Characteristic Function, Convergence of random variates, The Central Limit Theorem, Markov Chains[1], Random Walks[2].

Textbook(s):

This semester we will use as the main textbook:

- *Introduction to Probability Models*, 9th edition, by Sheldon M. Ross, Academic Press, 2006, ISBN-10: 0125980620 ISBN-13: 978-0125980623.

I choose this book mainly for the examples and exercises it contains.

However, the material which we cover goes beyond this book. On the course website (link above) I will post several chapters that detail the specific material covered in this class. Eventually, they will make a book but for now I only have these draft chapters. I am going to ask that if you find mistakes or missprints to mark them on the notes and give them to me at the end of the semester.

The following books are given as reference. They are on the list of reserved books in the library:

- *Probability: Theory and Examples*, by Richard Durrett, Thomson Learning 2004
- *Probability and Measure*, by Patrick Billingsley, Wiley series in probability and mathematical statistics 1995

[1] Time permitting
[2] idem

Math 611 Probability

Instructor: Ionut Florescu

Office: Kidde 227

Phone: (201) 216-5452

Email: ifloresc@stevens.edu

Office hours: M 4:00pm -6:00pm and by appt

website: http://www.math.stevens.edu/~ioresc/Teaching/2009-2010/index611.html

http://www.math.stevens.edu/~ifloresc/Teaching/2009-2010/MA611F09/

/~ifloresc/Teaching/2009-2010/MA611F09

	Parent Directory	-	
[ ]	Lecture1.pdf	30-Aug-2011 10:26	157K
[ ]	Lecture2.pdf	30-Aug-2011 10:26	138K
[ ]	Lecture3.pdf	30-Aug-2011 10:26	110K
[ ]	Lecture4.pdf	30-Aug-2011 10:26	64K
[ ]	MA611Sillabus.pdf	30-Aug-2011 10:26	56K
[ ]	hwk1.pdf	30-Aug-2011 10:26	18K
[ ]	hwk2.pdf	30-Aug-2011 10:26	18K
[ ]	hwk3.pdf	30-Aug-2011 10:26	35K

- *A course in probability theory*, by Kai Lai Chung, Academic Press 2000
- *Probability with Martingales*, by David Williams, Cambridge University Press 1991
- *Probability and Random Processes* by Geoffrey Grimmett and David Stirzaker, Oxford University Press 2001.

Homework, Exams and Grading:

We will have one midterm and a final exam. Their dates will be agreed on during the semester. We will have assignments during the semester. They will be graded and counting for the final grade. However, the most weight for the final grade will be coming from the final examination.

Chapter 1
Elements of Probability Measure

The axiomatic approach of Kolmogorov is followed by most Probability Theory books. This is the approach of choice for most graduate level probability courses. However, the immediate applicability of the theory learned as such is questionable and many years of study are required to understand and unleash its full power.

On the other hand the Applied probability books completely disregard this approach and they go more or less directly into presenting applications, thus leaving gaps into the reader's knowledge. At a cursory glance this approach appears to be very useful (the presented problems are all very real and most are difficult), however I question the utility of this approach when confronted with problems that are slightly different from the ones presented in such books.

Unfortunately, there is no middle ground between these two, hence the necessity of the present lecture notes. I will start with the axiomatic approach and present as much as I feel is going to be necessary for a complete understanding of the Theory of Probabilities. I will skip proofs which I consider will not bring something new to the development of the student's understanding.

1.1 Probability Spaces

Let Ω be an abstract set. This is sometimes denoted with S and is called the sample space. It is a set containing all the possible outcomes or results of a random experiment or phenomenon. I called it abstract because it could contain anything. For example if the experiment consists in tossing a coin once the space Ω could be represented as $\{Head, Tail\}$. However, it could just as well be represented as $\{Cap, Pajura\}$, these being the romanian equivalents of $Head$ and $Tail$. The space Ω could just as well contain an infinite number of elements. For example measuring the diameter of a doughnut could result in all possible numbers inside a whole range. Furthermore, measuring in inches or in centimeters would produce different albeit equivalent spaces.

We will use $\omega \in \Omega$ to denote a generic outcome or a sample point.

Any collection of outcomes is called an event. That is, any subset of Ω is an event. We shall use capital letters from the beginning of the alphabet A, B, C to denote these events.

So far so good. The proper definition of Ω is one of the most important issues when treating a problem probabilistically. However, this is not enough. We have to make sure that we can calculate the probability of all the items of interest.

Think of the following possible situation: Poles of various sizes are painted in all the possible nuances of colors. In other words the poles have two characteristics of interest size and color. Suppose that in this model we have to calculate the probability of things like the next pole would be shorter than 15 inches and painted a nuance of red or blue. In order to answer such questions we have to define properly the sample space Ω and furthermore give a definition of probability that will be consistent. Specifically, we need to give a definition of the elements of Ω which **can be** measured.

To this end we have to group these events into some way that would allow us to say: yes we can calculate the probability of all the events in this group. In other words, we need to talk about the notion of collection of events.

We will introduce the notion of σ-algebra (or σ-field) to deal with the problem of the proper domain of definition for the probability. Before we do that, we introduce a special collection of events:

$$\mathscr{P}(\Omega) = \text{The collection of all possible subsets of } \Omega \tag{1.1}$$

We could define probability on this very large set. However, this would mean that we would have to define probability for every single element of $\mathscr{P}(\Omega)$. This will prove impossible except on the case when Ω is finite. However, even in this case we have to do it consistently. For example if say the set $\{1,2,3\}$ is in Ω and has probability 0.2, how do we define the probability of $\{1,2\}$? How about probability of $\{1,2,5\}$? A much better approach would be to define probability only on the generators of the collection $\mathscr{P}(\Omega)$ or on the generators of a collection of sets as close as we can possibly make to $\mathscr{P}(\Omega)$.

How do we do this? Fortunately, algebra comes to the rescue. The elements of a collection of events are the events. So first we define operations with them: *union, intersection, complement* and slightly less important *difference and symmetric difference*.

$$\begin{cases} A \cup B & = \text{set of elements that are } \textbf{either} \text{ in} A \textbf{ or} \text{ in } B \\ A \cap B & = AB = \text{set of elements that are } \textbf{both} \text{ in } A \textbf{ and} \text{ in } B \\ A^c & = \bar{A} = \text{set of elements that are in } \Omega \text{ but } \textbf{not} \text{ in } A \end{cases} \tag{1.2}$$

$$\begin{cases} A \setminus B = & \text{set of elements that are in } A \text{ but } \textbf{not} \text{ in } B \\ A \triangle B = & (A \setminus B) \cup (B \setminus A) \end{cases}$$

We can of course express every operation in terms of union and intersection. There are important relations between these operations, I will stop here to mention the De Morgan laws:

$$\begin{cases} (A \cup B)^c & = A^c \cap B^c \\ (A \cap B)^c & = A^c \cup B^c \end{cases} \tag{1.3}$$

There is much more to be found out about set operations but for our purpose this is enough. Look at Billingsley (1995) or Chung (2000) for a wealth of more details.

Definition 1.1 (Algebra on Ω). A collection $\mathscr{F}$ of events in Ω is called an algebra (or field) on Ω iff:

1. $\Omega \in \mathscr{F}$
2. Closed under complementarity: If $A \subseteq \mathscr{F}$ then $A^c \subseteq \mathscr{F}$
3. Closed under finite union: If $A, B \subseteq \mathscr{F}$ then $A \cup B \subseteq \mathscr{F}$

Remark 1.1. The first two properties imply that $\varnothing \in \mathscr{F}$. The third is equivalent with $A \cap B \subseteq \mathscr{F}$ by the second property and the de Morgan laws (1.3).

Definition 1.2 (σ-Algebra on Ω). If $\mathscr{F}$ is an algebra on Ω and in addition it is closed under countable unions then it is a σ-algebra (or σ-field) on Ω

Note: Closed under countable unions means that the third property in Definition 1.1 is replaced with: If $n \in \mathbb{N}$ is a natural number and $A_n \subseteq \mathscr{F}$ for all n then

$$\bigcup_{n \in \mathbb{N}} A_n \subseteq \mathscr{F}$$

The σ-algebra provides an appropriate domain of definition for the probability function. However, it is such an abstract thing that it will be hard to work with it. This is the reason for the next definition, it will be much easier to work on the generators of a *sigma*-algebra. *This will be a recurring theme in probability, in order to show a property for a big class we show the property for a small generating set of the class and then use standard arguments to extend to the whole class.*

Definition 1.3 (σ algebra generated by a class $\mathscr{C}$ of sets in Ω).
Let $\mathscr{C}$ be a collection (class) of subsets of Ω. Then $\sigma(\mathscr{C})$ is the smallest σ-algebra on Ω that contains $\mathscr{C}$.

Mathematically:

1. $\mathscr{C} \subseteq \sigma(\mathscr{C})$
2. $\sigma(\mathscr{C})$ is a σ-field
3. If $\mathscr{C} \subseteq \mathscr{G}$ and $\mathscr{G}$ is a σ-field then $\sigma(\mathscr{C}) \subseteq \mathscr{G}$

The idea of this definition is to verify a statement on the set $\mathscr{C}$. Then, due to the properties that would be presented later the same statement will be valid for all the sets in $\sigma(\mathscr{C})$.

Proposition 1.1. *Properties of σ-algebras:*

- *$\mathscr{P}(\Omega)$ is a σ-algebra, the largest possible σ-algebra on Ω*
- *If $\mathscr{C}$ is already a σ-algebra then $\sigma(\mathscr{C}) = \mathscr{C}$*
- *If $\mathscr{C} = \{\varnothing\}$ or $\mathscr{C} = \{\Omega\}$ then $\sigma(\mathscr{C}) = \{\varnothing, \Omega\}$, the smallest possible σ-algebra on Ω*
- *If $\mathscr{C} \subseteq \mathscr{C}'$ then $\sigma(\mathscr{C}) \subseteq \sigma(\mathscr{C}')$*
- *If $\mathscr{C} \subseteq \mathscr{C}' \subseteq \sigma(\mathscr{C})$ then $\sigma(\mathscr{C}') = \sigma(\mathscr{C})$*

In general listing the elements of a sigma algebra explicitly is hard. It is only in simple cases that this is done.

Remark 1.2 (Finite space Ω). When the sample space is finite, we can and typically will take the sigma algebra to be $\mathscr{P}(\Omega)$. Indeed, any event of a finite space can be trivially expressed in terms of individual outcomes. In fact, if the finite space Ω contains M possible outcomes, then the number of possible events is finite and is equal with 2^M.

Example 1.1. Suppose a set $A \subset \Omega$. Let us calculate $\sigma(A)$. Clearly, by definition Ω is in $\sigma(A)$. Using the complementarity property we clearly see that A^c and $\emptyset$ are also in $\sigma(A)$. We only need to take unions of these sets and see that there are no more new sets. Thus:

$$\sigma(A) = \{\Omega, \emptyset, A, A^c\}.$$

□

Proposition 1.2 (Intersection and union of σ-algebras). *Suppose that $\mathscr{F}_1$ and $\mathscr{F}_2$ are two σ-algebras on Ω. Then:*

1. *$\mathscr{F}_1 \cap \mathscr{F}_2$ is a sigma algebra.*
2. *$\mathscr{F}_1 \cup \mathscr{F}_2$ is **not** a sigma algebra. The smallest σ algebra that contains both of them is: $\sigma(\mathscr{F}_1 \cup \mathscr{F}_2)$ and is denoted $\mathscr{F}_1 \vee \mathscr{F}_2$*

Proof. For part 2 there is nothing to show. Perhaps a counterexample. Take for instance two sets $A, B \subset \Omega$ such that $A \cap B \neq \emptyset$. Then take $\mathscr{F}_1 = \sigma(A)$ and $\mathscr{F}_2 = \sigma(B)$. Use the previous example and Exercise 1.2, part *c*.

For part 1 we just need to verify the definition of the sigma algebra. For example, take A in $\mathscr{F}_1 \cap \mathscr{F}_2$. So A belongs to both collections of sets. Since $\mathscr{F}_1$ is a sigma algebra by definition $A^c \in \mathscr{F}_1$. Similarly $A^c \in \mathscr{F}_2$. Therefore, $A^c \in \mathscr{F}_1 \cap \mathscr{F}_2$. The rest of the definition is verified in a similar manner. □

An example: Borel σ-algebra

Let Ω be a topological space (think geometry is defined in this space and this assures us that the open subsets exist in this space).

Definition 1.4. We define:

$$\begin{aligned}\mathscr{B}(\Omega) &= \text{The Borel } \sigma\text{-algebra} \\ &= \sigma\text{-algebra generated by the class of open subsets of } \Omega\end{aligned} \tag{1.4}$$

In the special case when $\Omega = \mathbb{R}$ we denote $\mathscr{B} = \mathscr{B}(\mathbb{R})$, the Borel sets of $\mathbb{R}$. This $\mathscr{B}$ is the most important σ-algebra. The reason for this fact is that most experiments can be brought to equivalence with $\mathbb{R}$ (as we shall see when we will talk about random variables). Thus, if we define a probability measure on $\mathscr{B}$, we have a way to calculate probabilities for most experiments. □

Most subsets of $\mathbb{R}$ are in $\mathscr{B}$. However, it is possible (though very difficult) to explicitly construct a subset of $\mathbb{R}$ which is not in $\mathscr{B}$. See (Billingsley, 1995, page 45) for such a construction in the case $\Omega = (0,1]$.

There is nothing special about the open sets, except for the fact that they can be defined in any topological space. In $\mathbb{R}$ we have alternate definitions which you will have to show are equivalent with the one given above in problem 1.7.

Probability measure

We are finally in the position to give the domain for the probability measure.

Definition 1.5 (Measurable Space.). A pair $(\Omega, \mathscr{F})$, where Ω is a set and $\mathscr{F}$ is a σ-algebra on Ω is called a *measurable space*.

Definition 1.6 (Probability measure. Probability space). Given a measurable space $(\Omega, \mathscr{F})$, a probability measure is any function $\mathbf{P} : \mathscr{F} \to [0,1]$ with the following properties:

i) $\mathbf{P}(\Omega) = 1$
ii) (countable additivity) For any sequence $\{A_n\}_{n\in\mathbb{N}}$ of disjoint events in $\mathscr{F}$ (i.e. $A_i \cap A_j = \varnothing$, for all $i \neq j$):

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

The triple $(\Omega, \mathscr{F}, \mathbf{P})$ is called a Probability Space.

Note that the probability measure is a set function (i.e., a function defined on sets).

The next two definitions are given for completeness only. However, we will use them later in this class. They are both presenting more general notions than a probability measure and they will be used later in hypotheses of some theorems to show that the results apply to even more general measures than probability measures.

Definition 1.7 (Finite Measure). Given a measurable space $(\Omega, \mathscr{F})$, a finite measure is a set function $\mu : \mathscr{F} \to [0,1]$ with the same countable additivity property as

defined above and the measure of the space finite instead of one. More specifically the first property above is replaced with:

$$\mu(\Omega) < \infty$$

Definition 1.8 (σ-finite Measure). A measure μ defined on a measurable space $(\Omega, \mathscr{F})$ is called σ-finite if it is countably additive and there exist a partition[1] of the space Ω, $\{\Omega_i\}_{i\in I}$, and $\mu(\Omega_i) < \infty$ for all $i \in I$. Note that the index set I is allowed to be countable.

Example 1.2 (Discrete Probability Space).

Let Ω be a countable space. Let $\mathscr{F} = \mathscr{P}(\Omega)$. Let $p : \Omega \to [0, N)$ be a function on Ω such that $\sum_{\omega\in\Omega} p(\omega) = N < \infty$, where N is a finite constant. Define:

$$\mathbf{P}(A) = \frac{1}{N} \sum_{\omega\in A} p(\omega)$$

We can show that $(\Omega, \mathscr{F}, \mathbf{P})$ is a Probability Space. Indeed, from the definition:

$$\mathbf{P}(\Omega) = \frac{1}{N} \sum_{\omega\in\Omega} p(\omega) = \frac{1}{N} N = 1.$$

To show the countable additivity property let A a set in Ω such that $A = \bigcup_{i=1}^{\infty} A_i$, with A_i disjoint sets in Ω. Since the space is countable we may write $A_i = \{\omega_1^i, \omega_2^i, \ldots\}$, where any of the sets may be finite, but $\omega_j^i \neq \omega_l^k$ for all i, j, k, l where either $i \neq k$ or $j \neq l$. Then using the definition we have:

$$\begin{aligned}\mathbf{P}(A) &= \frac{1}{N} \sum_{\omega\in\bigcup_{i=1}^{\infty} A_i} p(\omega) = \frac{1}{N} \sum_{i\geq 1, j\geq 1} p(\omega_j^i) \\ &= \frac{1}{N} \sum_{i\geq 1} \left(p(\omega_1^i) + p(\omega_2^i) + \ldots \right) = \sum_{i\geq 1} \mathbf{P}(A_i)\end{aligned}$$

□

This is a very simple example but it shows the basic probability reasoning.

Remark 1.3. The previous exercise gives a way to construct discrete probability measures (distributions). For example take $\Omega = \mathbb{N}$ the natural numbers and take $N = 1$ in the definition of probability of an event. Then:

- $p(\omega) = \begin{cases} 1-p & \text{, if } \omega = 0 \\ p & \text{, if } \omega = 1 \\ 0 & \text{, otherwise} \end{cases}$, gives the Bernoulli(p) distribution.
- $p(\omega) = \begin{cases} \binom{n}{\omega} p^{\omega}(1-p)^{n-\omega} & \text{, if } \omega \leq n \\ 0 & \text{, otherwise} \end{cases}$, gives the Binomial(n,p) distribution.

[1] a partition of the set A is a collection of sets A_i, disjoint ($A_i \cap A_j = \emptyset$, if $i \neq j$) such that $\cup_i A_i = A$

- $p(\omega) = \begin{cases} \binom{\omega-1}{r-1} p^r (1-p)^{\omega-r} & \text{, if } \omega \geq r \\ 0 & \text{, otherwise} \end{cases}$, gives the Negative Binomial(r,p) distribution.
- $p(\omega) = \frac{\lambda^\omega}{\omega!} e^{-\lambda}$, gives the Poisson (λ) distribution.

Example 1.3 (Uniform Distribution on (0,1)). As another example let $\Omega = (0,1)$ and $\mathscr{F} = \mathscr{B}((0,1))$ the Borel sigma algebra. Define a probability measure U as follows: for any open interval $(a,b) \subseteq (0,1)$ let $U((a,b)) = b-a$ the length of the interval. For any other open interval O define $U(O) = U(O \cap (0,1))$.

Note that we did not specify $U(A)$ for all Borel sets A, rather only for the generators of the Borel σ-field. This illustrates the probabilistic concept presented above. In our specific situation, under very mild conditions on the generators of the σ-algebra any probability measure defined only on the generators can be uniquely extended to a probability measure on the whole σ-algebra (Carathèodory extension theorem). In particular when the generators are open sets these conditions are true and we can restrict the definition to the open sets alone. This example is going to be extended in Section 1.5.

Proposition 1.3 (Elementary properties of Probability Measure). *Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a Probability Space. Then:*

1. *$\forall A, B \in \mathscr{F}$ with $A \subseteq B$ then $\mathbf{P}(A) \leq \mathbf{P}(B)$*
2. *$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B), \forall A, B \in \mathscr{F}$*
3. *(General Inclusion-Exclusion formula, also named Poincaré formula):*

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{i<j\leq n} \mathbf{P}(A_i \cap A_j) \tag{1.5}$$
$$+ \sum_{i<j<k\leq n} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^n \mathbf{P}(A_1 \cap A_2 \cdots \cap A_n)$$

   Note that successive partial sums are alternating between over-and-under estimating.
4. *(Finite subadditivity, sometimes called Boole's inequality):*

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathcal{P}(A_i), \quad \forall A_1, A_2, \ldots, A_n \in \mathscr{F}$$

1.1.1 Null element of $\mathscr{F}$. Almost sure (a.s.) statements. Indicator of a set.

An event $N \in \mathscr{F}$ is called a null event if $P(N) = 0$.

Definition 1.9. A statement $\mathcal{S}$ about points $\omega \in \Omega$ is said to be true *almost surely* (a.s.), almost everywhere (a.e.) or with probability 1 (w.p.1) if the set M defined as:

$$M := \{\omega \in \Omega | \mathcal{S}(\omega) \text{ is true}\},$$

is in $\mathscr{F}$ and $\mathbf{P}(M) = 1$, (or, equivalently M^c is a null set).

We will use the notions a.s., a.e., and w.p.1. to denote the same thing – the definition above. For example we will say $X \geq 0$ a.s. and mean: $\mathbf{P}\{\omega | X(\omega) \geq 0\} = 1$ or equivalently $\mathbf{P}\{\omega | X(\omega) < 0\} = 0$. The notion of almost sure is a fundamental one in probability. Unlike in deterministic cases where something has to always be true no matter what, in probability we care about "the majority of the truth". In other words probability recognizes that some phenomena may have extreme outcomes, but if they are extremely improbable then we do not care about them. Fundamentally, it is mathematics applied to reality.

Definition 1.10. We define the indicator function of an event A as the (simple) function $\mathbf{1}_A : \Omega \to \{0,1\}$,

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

Sometimes this function is denoted with I_A.

Note that the indicator function is a regular function (not a set function). Indicator functions are very useful in probability theory. Here are some useful relationships:

$$\mathbf{1}_{\mathbf{A}\cap\mathbf{B}}(\cdot) = 1_A(\cdot)1_B(\cdot)$$

If $\{B_i\}$ form a partition of Ω (i.e. the sets A_i are disjoint and $\Omega = \bigcup_{i=1}^n A_i$):

$$\mathbf{1}_{\mathbf{A}}(\cdot) = \sum_i 1_{A\cap B_i}(\cdot)$$

1.2 Conditional Probability

Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a Probability Space. Then for $A, B \in \mathscr{F}$ we define the conditional probability of A given B as usual by:

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

We can immediately rewrite the formula above to obtain the *multiplicative rule*:

$$\begin{aligned}
&\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\mathbf{P}(B), \\
&\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A|B \cap C)\mathbf{P}(B|C)\mathbf{P}(C), \quad \text{etc.}
\end{aligned}$$

Total probability formula: Given $A_1, A_2, \ldots, A_n$ a partition of Ω (i.e. the sets A_i are disjoint and $\Omega = \bigcup_{i=1}^{n} A_i$), then:

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B|A_i)\mathbf{P}(A_i), \quad \forall B \in \mathscr{F} \tag{1.6}$$

Bayes Formula: If $A_1, A_2, \ldots, A_n$ form a partition of Ω:

$$\mathbf{P}(A_j|B) = \frac{\mathbf{P}(B|A_j)\mathbf{P}(A_j)}{\sum_{i=1}^{n} \mathbf{P}(B|A_i)\mathbf{P}(A_i)}, \quad \forall B \in \mathscr{F}. \tag{1.7}$$

Example 1.4. A biker leaves the point O in the figure below. At each crossroad the biker chooses a road at random. What is the probability that he arrives at point A ?

Let B_k, $k = 1,2,3,4$ be the event that the biker passes through point B_k. These four events are mutually exclusive and they form a partition of the space. Moreover, they are equiprobable ($\mathbf{P}(B_k) = 1/4, \forall k \in \{1,2,3,4\}$). Let A denote the event "the biker reaches the destination point A". Conditioned on each of the possible points B_1-B_4 of passing we have:

$$\begin{aligned}
&\mathbf{P}(A|B_1) = 1/4 \\
&\mathbf{P}(A|B_2) = 1/2 \\
&\mathbf{P}(A|B_3) = 1
\end{aligned}$$

At B_4 is slightly more complex. We have to use the multiplicative rule:

$$\begin{aligned}
\mathbf{P}(A|B_4) &= 1/4 + \mathbf{P}(A \cap B_5|B_4) + \mathbf{P}(A \cap B_6 \cap B_5|B_4) \\
&= 1/4 + \mathbf{P}(A|B_5 \cap B_4)\mathbf{P}(B_5|B_4) + \mathbf{P}(A|B_6 \cap B_5 \cap B_4)\mathbf{P}(B_6|B_5 \cap B_4)\mathbf{P}(B_5|B_4) \\
&= 1/4 + 1/3(1/4) + 1(1/3)(1/4) = 3/12 + 2/12 = 5/12
\end{aligned}$$

Finally, by the law of total probability:

$$\begin{aligned}
\mathbf{P}(A) &= \mathbf{P}(A|B_1)\mathbf{P}(B_1) + \mathbf{P}(A|B_2)\mathbf{P}(B_2) + \mathbf{P}(A|B_3)\mathbf{P}(B_3) + \mathbf{P}(A|B_4)\mathbf{P}(B_4) \\
&= 1/4(1/4) + 1/2(1/4) + 1/4(1) + 5/12(1/4) = 13/24
\end{aligned}$$

□

Example 1.5 (De Méré's Paradox). As a result of extensive observation of dice games the French gambler Chevaliér De Méré noticed that the total number of spots

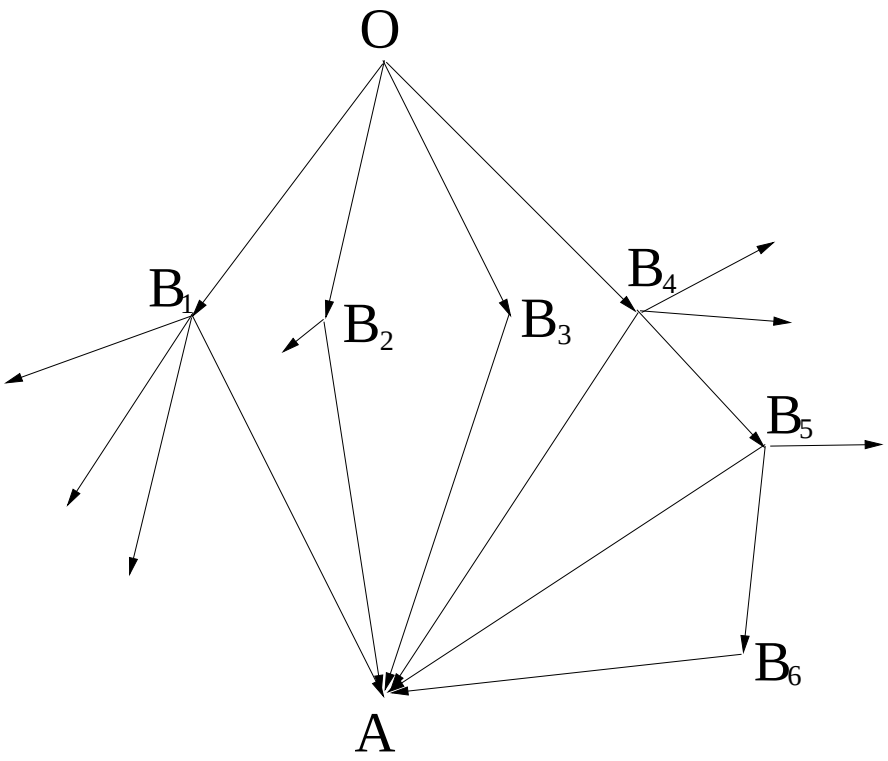


Fig. 1.1 The possible trajectories of the biker. O is the origin point and A is the arrival point. B_k's are intermediate points. Note that not all the ways lead to Rome, i.e. the probability of reaching Rome is less than 1.

showing on 3 dice thrown simultaneously turn out to be 11 more often than 12. However, from his point of view this is not possible since 11 occurs in six ways :

$(6:4:1);(6:3:2);(5:5:1);(5:4:2);(5:3:3);(4:4:3)$,

while 12 also in six ways:

$(6:5:1);(6:4:2);(6:3:3);(5:5:2);(5:4:3);(4:4:4)$

What is the fallacy in the argument?

Solution 1.1 (Solution due to Pascal). The argument would be correct if these "ways" would have the same probability. However this is not true. For example: (6:4:1) occurs in 3! ways, (5:5:1) occurs in 3 ways and (4:4:4) occurs in 1 way.

As a result we can easily calculate: $\mathbf{P}(11) = 27/216$; $\mathbf{P}(12) = 25/216$, and indeed his observation is correct and he should bet on 11 rather than on 12 if they have the same game payoff. □

Example 1.6 (Another De Mére's Paradox:). What is more probable?

1. Throw 4 dice and obtain at least one 6

2. Throw 2 dice 24 time and obtain at least once a double 6

Solution 1.2. For option 1: $1 - \mathbf{P}(\text{No 6}) = 1 - (5/6)^4 = 0.517747$.

For option 2: $1 - \mathbf{P}(\text{None of the 24 trials has a double 6}) = 1 - (35/36)^{24} = 0.491404$

Example 1.7 (Monty Hall problem). This is a problem named after the host of the American television show "Let's make a deal". Simply put at the end of a game you are left to chose between 3 closed doors. Two of them have nothing behind and one contains a prize. You chose one door but the door is not opened automatically. Instead, the presenter opens another door that contains nothing. He then gives you the choice of changing the door or sticking with the initial choice.

Most people would say that it does not matter what you do at this time, but that is not true. In fact everything depends on the host behavior. For example, if the host knows in advance where the prize is and always reveals at random some other door that does not contain anything then it is always better to switch.

Solution 1.3. This problem generated a lot of controversy since its publication (in 1970's) since the solution seems so counterintuitive. Articles talking about this problem in more detail Morgan et al. (1991), Mueser and Granberg (1991). We are presenting it here since it exemplifies the conditional probability reasoning. The key in any such problem is the sample space which has to be complete enough to be able to answer the questions asked.

Let D_i be the event that the price is behind door i. Let SW be the event that switching wins the price[2].

It does not matter which door we chose initially the reasoning is identical with all the three doors. So, we assume that initially we pick door 1.

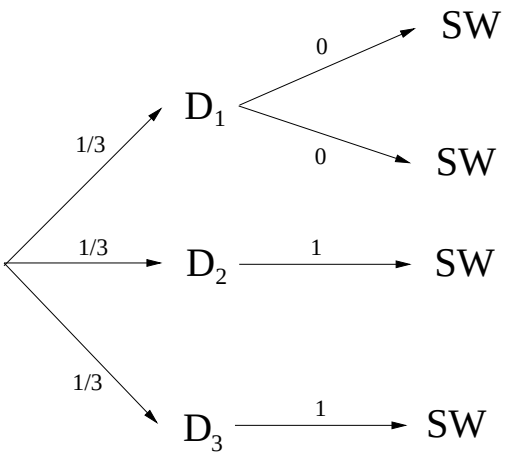


Fig. 1.2 The tree diagram of conditional probabilities. Note that the presenter has two choices in case D_1 neither of which results in winning if switching the door.

[2] As a side note this event is the same as the event "not switching loses"

Events D_i $i = 1,2,3$ are mutually exclusive and we can write:

$$\mathbf{P}(SW) = \mathbf{P}(SW|D_1)\mathbf{P}(D_1) + \mathbf{P}(SW|D_2)\mathbf{P}(D_2) + \mathbf{P}(SW|D_3)\mathbf{P}(D_3).$$

When the prize is behind door 1 since we chose door 1 the presenter has two choices for the door to show us. However, neither would contain the prize and in either case switching does not result in winning the prize, therefore $\mathbf{P}(SW|D_1) = 0$. If the car is behind door 2 since our choice is door 1 the presenter has no alternative but to show us the other door 3 which contains nothing. Thus switching in this case results in winning the price. The same reasoning works if the prize is behind door 3. Therefore:

$$\mathbf{P}(SW) = 1\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3} = \frac{2}{3}$$

Thus switching has a higher probability of winning than not switching.

A generalization to n doors shows that it still is advantageous to switch but the advantage decreases as $n \to \infty$. Specifically, in this case $\mathbf{P}(D_i) = 1/n$; $\mathbf{P}(SW|D_1) = 0$ still, but $\mathbf{P}(SW|D_i) = 1/(n-2)$ if $i \neq 1$. Which gives:

$$\mathbf{P}(SW) = \sum_{i=2}^{n} \frac{1}{n}\frac{1}{n-2} = \frac{n-1}{n-2}\frac{1}{n} > \frac{1}{n}$$

Furthermore, different presenter strategies produce different answers. For example, if the presenter offers the option to switch only when the player chooses the right door then switching is always bad. If the presenter offers switching only when the player has chosen incorrectly then switching always wins. These and other cases can be analyzed in Rosenthal (2008).

Example 1.8 (Bertrand's box paradox). This problem was first formulated by Joseph Louis François Bertrand in his Calcul de Probabilités (Bertrand, 1889). In some sense this problem is related to the previous problem but it does not depend on any presenter strategy and the solution is much more clear. Solving this problem is an exercise in Bayes formula.

Suppose that we know that three boxes contain respectively: one box contains two gold coins, a second box with two silver coins, and a third box with one of each. We chose a box at random and from that box we chose a coin also at random. Then we look at the coin chosen. Given that the coin chosen was gold what is the probability that the other coin in the box chosen is also gold. At a first glance it may seem that this probability is $1/2$ but after calculation this probability turns out to be 2/3.

Solution 1.4. We plot the sample space in Figure 1.3. Using this tree we can calculate the probability:

$$\mathbf{P}(\text{Second coin is } G|\text{First coin is } G) = \frac{\mathbf{P}(\text{Second coin is } G \text{ and First coin is } G)}{\mathbf{P}(\text{First coin is } G)}.$$

Now, using the probabilities from the tree we continue:

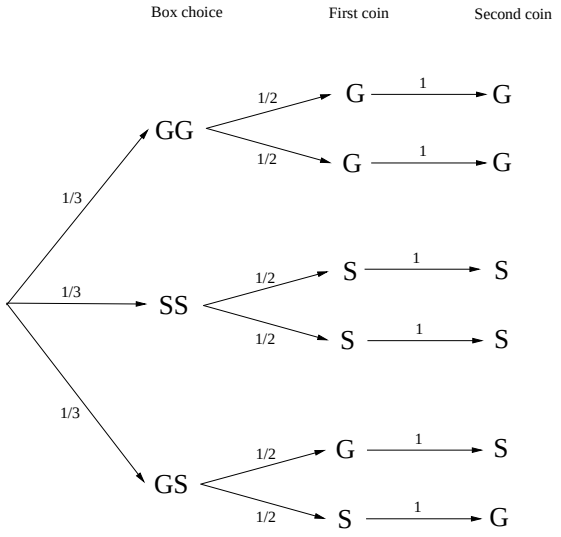


Fig. 1.3 The tree diagram of conditional probabilities.

$$= \frac{\frac{1}{3}\frac{1}{2}1 + \frac{1}{3}\frac{1}{2}1}{\frac{1}{3}\frac{1}{2}1 + \frac{1}{3}\frac{1}{2}1 + \frac{1}{3}\frac{1}{2}1} = \frac{2}{3}.$$

Now that we have seen the solution we can recognize a logical solution to the problem as well. Given that the coin seen is gold we can throw away the middle box. Then if this would be box 1 then we have two possibilities that the other coin is gold (depending on which we have chosen in the first place). If this is the box 2 then there is one possibility (the remaining coin is silver). Thus the probability should be 2/3 since we have two out of three chances. Of course this "logical" argument does not work if we do not choose the boxes with the same probability. □

Example 1.9. A blood test is 95% effective in detecting a certain disease when it is in fact present. However, the test yields also a false positive result for 1% of the people tested. If 0.5% of the population actually has the disease, what is the probability that the person is diseased given that the test is positive?

Solution 1.5. This problem illustrates once again the application of the Bayes rule. I do not like to use the rule literally instead work from first principles one will also obtain the Bayes rule without memorizing anything. We start by describing the sample space. Refer to the Figure 1.4 for this purpose.

So given that the test is positive means that we have to calculate a conditional probability. We may write:

$$\mathbf{P}(D|+) = \frac{\mathbf{P}(D \cap +)}{\mathbf{P}(+)} = \frac{\mathbf{P}(+|D)\mathbf{P}(D)}{\mathbf{P}(+)} = \frac{0.95(0.005)}{0.95(0.005) + 0.01(0.995)} = 0.323$$

How about if only 0.05% (i.e. 0.0005) of the population has the disease?

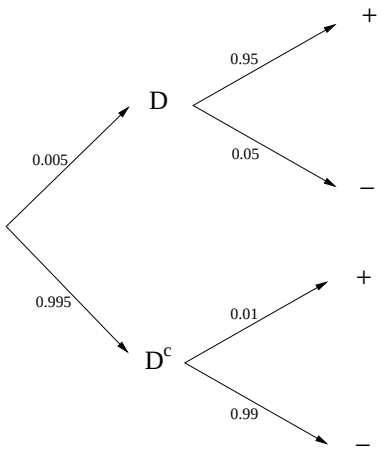


Fig. 1.4 Blood test probability diagram

$$\mathbf{P}(D|+) = \frac{0.95(0.0005)}{0.95(0.0005)+0.01(0.9995)} = 0.0454$$

This problem is an exercise in thinking. It is the same test device. In the first case the disease is relatively common and thus the test device is more or less reliable (though 32% right is very low). In the second case however the disease is very rare and thus the precision of the device goes way down. □

Example 1.10 (Gambler's Ruin Problem). We conclude this section with an example which we shall see many times throughout this book. I do not know who to credit with the invention of the problem since it is so mentioned so often in every probability treaties[3].

The formulation is simple. A game of heads or tails with a fair coin. Player wins 1 dollar if he successfully calls the side of the coin which lands upwards and loses \$1 otherwise. Suppose the initial capital is X dollars and he intends to play until he wins m dollars but no longer. What is the probability that the gambler will be ruined?

Solution 1.6. We will display what is called as a first step analysis.

Let $p(x)$ denote the probability that the player is going to be eventually ruined if he starts with x dollars.

If he wins the next game then he will have \$ $x+1$ and he is ruined from this position with prob $p(x+1)$.

If he loses the next game then he will have \$ $x-1$ so he is ruined from this position with prob $p(x-1)$.

Let R be the event he is eventually ruined. Let W be the event he wins the next trial. Let L be the event he loses this trial. Using the total prob. formula we get:

$$\mathbf{P}(R) = \mathbf{P}(R|W)\mathbf{P}(W) + \mathbf{P}(R|L)\mathbf{P}(L) \Rightarrow p(x) = p(x+1)(1/2) + p(x-1)(1/2)$$

[3] The formalization may be due to Huygens (1629-1695) in the XVII-th century

Is this true for all x? No. This is true for $x \geq 1$ and $x \leq w-1$. In the rest of cases we obviously have $p(0) = 1$ and $p(m) = 0$ which give the boundary conditions for the equation above.

This is a linear difference equation with constant coefficients. Please look at the general methodology in the following subsection on how to solve such equations.

Applying the method in our case gives the characteristic equation:

$$y = \frac{1}{2}y^2 + \frac{1}{2} \Rightarrow y^2 - 2y + 1 = 0 \Rightarrow (y-1)^2 = 0 \Rightarrow y_1 = y_2 = 1$$

In our case the two solutions are equal thus we seek a solution of the form $p(x) = (C + Dx)1^n = C + Dx$. Using the initial conditions we get: $p(0) = 1 \Rightarrow C = 1$ and $p(m) = 0 \Rightarrow C + Dm = 0 \Rightarrow D = -C/m = -1/m$, thus the general probability of ruin starting with wealth x is:

$$p(x) = 1 - x/m.$$

□

Solving difference equations with constant coefficients

This methodology is given for second order difference equations but higher order equations are solved in a very similar way. Suppose we are given an equation of the form:

$$a_n = Aa_{n-1} + Ba_{n-2},$$

with some boundary conditions.

The idea is to look for solutions of the form $a_n = cy^n$, with c some constant and y needs to be determined. Note that if we have two solutions of this form (say $c_1y_1^n$ and $c_2y_2^n$), then any linear combination of them is also a solution. We substitute this proposed form and obtain:

$$y^n = Ay^{n-1} + By^{n-2}.$$

Dividing by y^{n-2} we obtain the characteristic equation:

$$y^2 = Ay + B.$$

Next, we solve this equation and obtain real solutions y_1 and y_2 (if they exist). It may be possible that the characteristic equation does not have solutions in $\mathbb{R}$ in which case the difference equation does not have solutions either. Now we have two cases:

1. If y_1 and y_2 are distinct then the solution is $a_n = Cy_1^n + Dy_2^n$ where C, D are constants that are going to be determined from the initial conditions.

2. If $y_1 = y_2$ the solution is $a_n = Cy_1^n + Dny_1^n$. Again, C and D are determined from the initial conditions.

In the case when the difference equation contains p terms the procedure is identical even replicating the multiplicity issues. For more information one can consult any book on Ordinary Differential Equations such as Boyce and DiPrima (2004).

1.3 Independence

Definition 1.11. Two events A and B are called independent if and only if

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$$

.

The events $A_1, A_2, A_3, \dots$ are called *mutually independent* (or sometimes simply independent) if for every subset J of $\{1, 2, 3, \dots\}$ we have:

$$\mathbf{P}\left(\bigcup_{j \in J} A_j\right) = \prod_{j \in J} \mathbf{P}(A_j)$$

The events $A_1, A_2, A_3, \dots$ are called *pairwise independent* (sometimes jointly independent) if:

$$\mathbf{P}(A_i \cup A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j), \quad \forall i, j.$$

Note that jointly independent does not imply independence.

Two *sigma fields* $\mathscr{G}, \mathscr{H} \in \mathscr{F}$ *are* $\mathbf{P}$*–independent* if:

$$\mathbf{P}(G \cap H) = \mathbf{P}(G)\mathbf{P}(H), \quad \forall G \in \mathscr{G}, \forall H \in \mathscr{H}.$$

See Billingsley (1995) for the definition of independence of $k \geq 2$ sigma-algebras.

1.4 Monotone Convergence properties of probability

Let us take a step back for a minute and comment on what we have seen thus far. The σ-algebra differs from the regular algebra in that it allows us to deal with countable (not finite) number of sets. In fact this is a recurrent theme in probability, learning to deal with infinity. On finite spaces things are more or less simple. One has to define the probability of each individual outcome and everything proceeds from there. However, even in these simple cases imagine that one repeats an experiment over and over. Then again we are forced to cope with infinity. This section introduces a way to deal with this infinity problem.

Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a Probability Space.

Lemma 1.1. *The following are true:*

1. *If $A_n, A \in \mathscr{F}$ and $A_n \uparrow A$ (i.e., $A_1 \subseteq A_2 \subseteq \ldots A_n \subseteq \ldots$ and $A = \bigcup_{n\geq 1} A_n$), then: $\mathbf{P}(A_n) \uparrow \mathbf{P}(A)$ as a sequence of numbers.*
2. *If $A_n, A \in \mathscr{F}$ and $A_n \downarrow A$ (i.e., $A_1 \supseteq A_2 \supseteq \ldots A_n \supseteq \ldots$ and $A = \bigcap_{n\geq 1} A_n$), then: $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$ as a sequence of numbers.*
3. *(Countable subadditivity) If $A_1, A_2, \ldots,$ and $\bigcup_{i=1}^{\infty} A_n \in \mathscr{F}$, with A_i's not necessarily disjoint then:*

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathcal{P}(A_n)$$

Proof. 1. Let $B_1 = A_1, B_2 = A_2 \setminus A_1, \ldots, B_n = A_n \setminus A_{n-1}$. Because the sequence is increasing we have that the B_i's are disjoint thus:

$$\mathbf{P}(A_n) = \mathbf{P}(B_1 \cup B_2 \cup \cdots \cup B_n) = \sum_{i=1}^{n} \mathbf{P}(B_i).$$

Thus using countable additivity:

$$\mathbf{P}\left(\bigcup_{n\geq 1} A_n\right) = \mathbf{P}\left(\bigcup_{n\geq 1} B_n\right) = \sum_{i=1}^{\infty} \mathbf{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{P}(B_i) = \lim_{n\to\infty} \mathbf{P}(A_n)$$

2. Note that $A_n \downarrow A \Leftrightarrow {A_n}^c \uparrow A^c$ and from part 1 this means $1 - \mathbf{P}(A_n) \uparrow 1 - \mathbf{P}(A)$.

3. Let $B_1 = A_1, B_2 = A_1 \cup A_2, \ldots, B_n = A_1 \cup \cdots \cup A_n, \ldots$. From the finite subadditivity property in Proposition 1.3 we have that $\mathbf{P}(B_n) = \mathbf{P}(A_1 \cup \cdots \cup A_n) \leq \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)$.

$\{B_n\}_{n\geq 1}$ is an increasing sequence of events, thus from part 1 we get that $\mathbf{P}(\bigcup_{n=1}^{\infty} B_n) = \lim_{n\to\infty} \mathbf{P}(B_n)$. Combining the two relations above we obtain:

$$\mathbf{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbf{P}(\bigcup_{n=1}^{\infty} B_n) \leq \lim_{n\to\infty} (\mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n)) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

□

Lemma 1.2. *The union of a countable number of* **P***-null sets is a* **P***-null set*

This Lemma is a direct consequence of the countable subadditivity.

Recall from analysis: For a sequence of numbers $\{x_n\}_n$ lim sup and lim inf are defined:

$$\limsup x_n = \inf_m \{\sup_{n\geq m} x_n\} = \lim_{m\to\infty} (\sup_{n\geq m} x_n)$$
$$\liminf x_n = \sup_m \{\inf_{n\geq m} x_n\} = \lim_{m\to\infty} (\inf_{n\geq m} x_n),$$

and they represent the highest (respectively lowest) limiting point of a subsequence included in $\{x_n\}_n$.

Note that if z is a number such that $z > \limsup x_n$ then $x_n < z$ eventually[4].
Likewise, if $z < \limsup x_n$ then $x_n > z$ infinitely often[5].
These notions are translated to probability in the following way.

Definition 1.12. Let $A_1, A_2, \ldots$ be an infinite sequence of events, in some probability space $(\Omega, \mathscr{F}, \mathbf{P})$. We define the events:

$$\limsup_{n\to\infty} A_n = \bigcap_{n\geq 1} \bigcup_{m=n}^{\infty} A_m = \{\omega : \omega \in A_n \text{ for infinitely many n}\} = \{A_n \text{ i.o.}\}$$

$$\liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} \bigcap_{m=n}^{\infty} A_m = \{\omega : \omega \in A_n \text{ for all n large enough}\} = \{A_n \text{ eventually}\}$$

Let us clarify the notions of "infinitely often" and "eventually" a bit more. We say that an outcome ω happens infinitely often for the sequence $A_1, A_2, \ldots, A_n, \ldots$ if ω is in the set $\bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m$. This means that for any n (no matter how big) there exist an $m \geq n$ and $\omega \in A_m$.

We say that an outcome ω happens eventually for the sequence $A_1, A_2, \ldots, A_n, \ldots$ if ω is in the set $\bigcup_{n=1}^{\infty} \bigcap_{m\geq n} A_m$. This means that there exist an n such that for any $m \geq n$, $\omega \in A_m$, so from this particular n and up ω is in all the sets.

Why so complicate definitions? The basic intuition is the following: say you roll a die infinitely many times, then it is obvious what it means for the outcome 1 to appear infinitely often. Also, we can say the average of the rolls will eventually be arbitrarily close to 3.5 (this will be shown later). It is not so clear cut in general. The framework above provides a generalization to these notions.

The Borel Cantelli lemmas

With this definitions we are now capable to give two important lemmas.

Lemma 1.3 (First Borel-Cantelli). *If $A_1, A_2, \ldots$ is any infinite sequence of events with the property $\sum_{n\geq 1} \mathbf{P}(A_n) < \infty$ then*

$$\mathbf{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m\right) = \mathbf{P}\left(A_n \textit{ events are true infinitely often}\right) = 0$$

This lemma essentially says that if the probabilities of events go to zero and the sum is convergent then necessarily A_n will stop occurring. However, the reverse of the statement is not true. To make it hold we need a very strong condition (independence).

[4] i.e., there is some n_0 very large so that $x_n < z$, for all $n \geq n_0$

[5] i.e., for any n there exists an $m \geq n$ such that $x_m > z$

Lemma 1.4 (Second Borel-Cantelli). *If $A_1, A_2, \ldots$ is an infinite sequence of **independent** events then:*

$$\sum_{n\geq 1} \mathbf{P}(A_n) = \infty \quad \Leftrightarrow \quad \mathbf{P}(A_n \text{ i.o.}) = 1.$$

Proof. First Borel-Cantelli.

$$\mathbf{P}(A_n \text{ i.o.}) = \mathbf{P}\left(\bigcap_{n\geq 1}\bigcup_{m=n}^{\infty} A_m\right) \leq \mathbf{P}\left(\bigcup_{n=m}^{\infty} A_m\right) \leq \sum_{m=n}^{\infty} \mathbf{P}(A_m), \forall n$$

where we used the definition and countable subadditivity. By the hypothesis the sum on the right is the tail end of a convergent series, therefore converges to zero as $n \to \infty$. Thus we are done. □

Proof. Second Borel-Cantelli:

"$\Rightarrow$" Clearly, showing that $\mathbf{P}(A_n \text{ i.o.}) = \mathbf{P}(\limsup A_n) = 1$ is the same as showing that $\mathbf{P}((\limsup A_n)^c) = 0$.

By the definition of lim sup and the DeMorgan's laws,

$$(\limsup A_n)^c = \left(\bigcap_{n\geq 1}\bigcup_{m=n}^{\infty} A_m\right)^c = \bigcup_{n\geq 1}\bigcap_{m=n}^{\infty} A_m^c.$$

Therefore, it is enough to show that $\mathbf{P}(\bigcap_{m=n}^{\infty} A_m^c) = 0$ for all n (recall that a countable union of null sets is a null set). However,

$$\begin{aligned}
\mathbf{P}\left(\bigcap_{m=n}^{\infty} A_m^c\right) &= \lim_{r\to\infty} \mathbf{P}\left(\bigcap_{m=n}^{r} A_m^c\right) = \underbrace{\lim_{r\to\infty} \prod_{m=n}^{\infty} \mathbf{P}(A_m^c)}_{\text{by independence}} \\
&= \lim_{r\to\infty} \prod_{m=n}^{r} (1 - \mathbf{P}(A_m)) \leq \underbrace{\lim_{r\to\infty} \prod_{m=n}^{r} e^{-\mathbf{P}(A_m)}}_{1-x\leq e^{-x} \text{ if } x\geq 0} \\
&= \lim_{r\to\infty} e^{-\sum_{m=n}^{r} \mathbf{P}(A_m)} = e^{-\sum_{m=n}^{\infty} \mathbf{P}(A_m)} = 0
\end{aligned}$$

The last equality follows since $\sum \mathbf{P}(A_n) = \infty$.
Note that we have used the following inequality: $1 - x \leq e^{-x}$ which is true if $x \in [0, \infty)$. One can prove this inequality with elementary analysis.

"$\Leftarrow$" This implication is the same as the first lemma. Indeed, assume by absurd that $\sum \mathbf{P}(A_n) < \infty$. By the First Borel-Cantelli Lemma this implies that $\mathbf{P}(A_n \text{ i.o.}) = 0$, a contradiction with the hypothesis. □

The Fatou lemmas

Again assume that $A_1, A_2, \ldots$ is a sequence of events.

Lemma 1.5 (Fatou lemma for sets). *Given any measure (not necessarily finite) μ we have:*

$$\mu(A_n \textit{ eventually}) = \mu(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mu(A_n)$$

Proof. Recall that $\liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} \bigcap_{m=n}^{\infty} A_m$, and denote this set with A. Let $B_n = \bigcap_{m=n}^{\infty} A_m$, which is an increasing sequence (less intersections as n increases) and $B_n \uparrow A =$. By the monotone convergence property of measure (Lemma 1.1) $\mu(B_n) \to \mu(A)$. However,

$$\mu(B_n) = \mu(\bigcap_{m=n}^{\infty} A_m) \leq \mu(A_m), \forall m \geq n,$$

thus $\mu(B_n) \leq \inf_{m\geq n} \mu(A_m)$. Therefore:

$$\mu(A) \leq lim_{n\to\infty} \inf_{m\geq n} \mu(A_m) = \liminf_{n\to\infty} \mu(A_n)$$

□

Lemma 1.6 (The reverse of the Fatou lemma). *If $\mathbf{P}$ is a finite measure (e.g., probability measure) then:*

$$\mathbf{P}(A_n \textit{ i.o.}) = \mathbf{P}(\limsup_{n\to\infty} A_n) \geq \limsup_{n\to\infty} \mathbf{P}(A_n)$$

.

Proof. This proof is entirely similar. Recall that $\limsup_{n\to\infty} A_n = \bigcap_{n\geq 1} \bigcup_{m=n}^{\infty} A_m$, and denote this set with A. Let $B_n = \bigcup_{m=n}^{\infty} A_m$. Then clearly B_n is a decreasing sequence and $B_n \downarrow A =$. By the monotone convergence property of measure (Lemma 1.1) and since the measure is finite $\mathbf{P}(B_1) < \infty$ so $\mathbf{P}(B_n) \to \mathbf{P}(A)$. However,

$$\mathbf{P}(B_n) = \mathbf{P}(\bigcup_{m=n}^{\infty} A_m) \geq \mathbf{P}(A_m), \forall m \geq n,$$

thus $\mathbf{P}(B_n) \geq \sup_{m\geq n} \mathbf{P}(A_m)$, again since the measure is finite . Therefore:

$$\mathbf{P}(A) \geq lim_{n\to\infty} \sup_{m\geq n} \mathbf{P}(A_m) = \limsup_{n\to\infty} \mathbf{P}(A_n)$$

□

Kolmogorov zero-one law

I like to present this theorem since it introduces the concept of *a sequence of* σ*-algebras*, a notion essential for stochastic processes.

For a sequence $A_1, A_2, \dots$ of events in the probability space $(\Omega, \mathscr{F}, \mathscr{P})$ consider the generated sigma algebras $\mathscr{T}_n = \sigma(A_n, A_{n+1}, \dots)$ and their intersection

$$\mathscr{T} = \bigcap_{n=1}^{\infty} \mathscr{T}_n = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \dots),$$

called the tail σ-field.

Theorem 1.1 (Kolmogorov's 0-1 Law). *If $A_1, A_2, \dots$ are independent then for every event A in the tail σ field ($A \in \mathscr{T}$) its probability $\mathbf{P}(A)$ is either* 0 *or* 1.

Proof. Skipped. The idea is to show that A is independent of itself thus $\mathbf{P}(A \cap A) = \mathbf{P}(A)\mathbf{P}(A) \Rightarrow \mathbf{P}(A) = \mathbf{P}(A)^2 \Rightarrow \mathbf{P}(A)$ is either 0 or 1. The steps of this proof are as follows:

1. First define $\mathscr{A}_n = \sigma(A_1, \dots, A_n)$ and show that is independent of $\mathscr{T}_{n+1}$ for all n.
2. Since $\mathscr{T} \subseteq \mathscr{T}_{n+1}$ and $\mathscr{A}_n$ is independent of $\mathscr{T}_{n+1}$, then $\mathscr{A}_n$ and $\mathscr{T}$ are independent for all n.
3. Define $\mathscr{A}_\infty = \sigma(A_1, A_2, \dots)$. Then from the previous step we deduce that $\mathscr{A}_\infty$ and $\mathscr{T}$ are independent.
4. Finally since $\mathscr{T} \subseteq \mathscr{A}_\infty$ by the previous step $\mathscr{T}$ is independent of itself and the result follows.

Note that $\limsup A_n$ and $\liminf A_n$ are tail events. However, it is only in the case when the original events are independent that we can apply Kolmogorov's theorem. Thus in that case $\mathbf{P}\{A_n \text{ i.o.}\}$ is either 0 or 1.

1.5 Lebesgue measure on the unit interval (0,1]

We conclude this chapter with the most important measure available. This is the unique measure that makes things behave in a normal way (e.g., the interval $(0.2, 0.5)$ has measure 0.3).

Let $\Omega = (0,1]$. Let $\mathscr{F}_0$=class of semiopen subintervals (a,b] of Ω. For an interval $I = (a,b] \in \mathscr{F}_0$ define $\lambda(I) = |I| = b - a$. Let $\varnothing \in \mathscr{F}_0$ the element of length 0. Let $\mathscr{B}_0$=the algebra of finite disjoint unions of intervals in (0,1]. Note that the problem 1.3 shows that this algebra is not a σ-algebra.

If $A = \sum_{i=1}^{n} I_n \in \mathscr{B}_0$ with I_n disjoint $\mathscr{F}_0$ sets; then

$$\lambda(A) = \sum_{i=1}^{n} \lambda(I_i) = \sum_{i=1}^{n} |I_i|$$

The goal is to show that λ is countably additive on the algebra $\mathscr{B}_0$. This will allow us to construct a measure (actually a prob. measure since we are working on (0,1]) using the next result (Caratheodory's theorem). The constructed measure is well defined and will be called the Lebesgue Measure.

Theorem 1.2 (Theorem for the length of intervals:). *Let $I = (a,b] \subseteq (0,1]$ and I_k of the form $(a_k,b_k]$ bounded but not necessarily in $(0,1]$.*

(i) If $\bigcup_k I_k \subseteq I$ and I_k are disjoint then $\sum_k |I_k| \leq |I|$
(ii) If $I \subseteq \bigcup_k I_k$ (with the I_k not necessarily disjoint) then $|I| \leq \sum_k |I_k|$.
(iii) If $I = \bigcup_k I_k$ and I_k disjoint then $|I| = \sum_k |I_k|$.

Proof. Exercise (*Hint:* use induction)

Note: Part (iii) shows that the function λ is well defined.

Theorem 1.3. *λ is a (countably additive) probability measure on the field $\mathscr{B}_0$. λ is called the Lebesgue measure restricted to the algebra $\mathscr{B}_0$*

Proof. Let $A = \bigcup_{k=1}^{\infty} A_k$, where A_k are disjoint $\mathscr{B}_0$ sets. By definition of $\mathscr{B}_0$,

$$A_k = \bigcup_{j=1}^{m_k} J_{k_j}, \quad A = \bigcup_{i=1}^{n} I_i,$$

where the J_{k_j} are disjoint. Then,

$$\lambda(A) = \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} (\sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |I_i \cap J_{k_j}|) = \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} (\sum_{i=1}^{n} |I_i \cap J_{k_j}|)$$

and since $A \cap J_{k_j} = J_{k_j} \Rightarrow |A \cap J_{k_j}| = \sum_{i=1}^{n} |I_i \cap J_{k_j}| = |J_{k_j}|$, the above is continued:

$$= \sum_{k=1}^{\infty} \underbrace{\sum_{j=1}^{m_k} |J_{k_j}|}_{=|A_k|} = \sum_{k=1}^{\infty} \lambda(A_k)$$

□

The next theorem will extend the Lebesgue measure to the whole $(0,1]$, thus we define the probability space $((0,1], \mathscr{B}((0,1]), \lambda)$. The same construction with minor modifications works in $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ case.

Theorem 1.4 (Caratheodory's Extension Theorem). *A probability measure on an algebra has a unique extension to the generated σ-algebra.*

Note: The Caratheodory Theorem practically constructs all the interesting probability models. However, once we construct our models we have no further need of the theorem. It also reminds us of the central idea in the theory of probabilities: If one wants to prove something for a big set one needs to look first at the generators of that set.

Proof. (skipped), in the exercises.

Definition 1.13 (Monotone Class). A class $\mathscr{M}$ of subsets in Ω *is monotone* if it is closed under the formation of monotone unions and intersections, i.e.:

(i) $A_1, A_2, \cdots \in \mathscr{M}$ and $A_n \subset A_{n+1}, \bigcup_n A_n = A \Rightarrow A \in \mathscr{M}$
(ii) $A_1, A_2, \cdots \in \mathscr{M}$ and $A_n \supset A_{n+1} \Rightarrow \bigcap_n A_n \in \mathscr{M}$

The next theorem is only needed for the proof of the Caratheodory theorem. However, the proof is interesting and that is why is presented here.

Theorem 1.5. *If $\mathscr{F}_0$ is an algebra and $\mathscr{M}$ is a monotone class, then $\mathscr{F}_0 \subseteq \mathscr{M} \Rightarrow \sigma(\mathscr{F}_0) \subseteq \mathscr{M}$.*

Proof. Let $m(\mathscr{F}_0) =$ minimal monotone class over $\mathscr{F}_0 =$ the intersection of all monotone classes containing $\mathscr{F}_0$

We will prove that $\sigma(\mathscr{F}_0) \subseteq m(\mathscr{F}_0)$.

To show this it is enough to prove that $m(\mathscr{F}_0)$ is an algebra. Then exercise 1.11 will show that $m(\mathscr{F}_0)$ is a σ algebra. Since $\sigma(\mathscr{F}_0)$ is the smallest the conclusion follows.

To this end, let $\mathscr{G} = \{A : A^c \in m(\mathscr{F}_0)\}$.

(i) Since $m(\mathscr{F}_0)$ is a monotone class so is $\mathscr{G}$.
(ii) Since $\mathscr{F}_0$ is an algebra its elements are in $\mathscr{G} \Rightarrow \mathscr{F}_0 \subset \mathscr{G}$

(i) and (ii) $\Rightarrow m(\mathscr{F}_0) \subseteq \mathscr{G}$. Thus $m(\mathscr{F}_0)$ is closed under complementarity.

Now define $\mathscr{G}_1 = \{A : A \cup B \in m(\mathscr{F}_0), \forall B \in \mathscr{F}_0\}$.

We show that $\mathscr{G}_1$ is a monotone class:

Let $A_n \nearrow$ an increasing sequence of sets, $A_n \in \mathscr{G}_1$. By definition of $\mathscr{G}_1$, for all n $A_n \cup B \in m(\mathscr{F}_0), \forall B \in \mathscr{F}_0$.

But $A_n \cup B \supseteq A_{n-1} \cup B$ and thus the definition of $m(\mathscr{F}_0)$ implies:

$$\bigcup_n (A_n \cup B) \in m(\mathscr{F}_0), \forall B \in \mathscr{F}_0 \Rightarrow \left(\bigcup_n A_n\right) \cup B \in m(\mathscr{F}_0), \forall B,$$

and thus $\bigcup_n A_n \in \mathscr{G}_1$.

This shows that $\mathscr{G}_1$ is a monotone class. But since $\mathscr{F}_0$ is an algebra its elements (the contained sets) are in $\mathscr{G}_1$[6], thus $\mathscr{F}_0 \subset \mathscr{G}_1$. Since $m(\mathscr{F}_0)$ is the smallest monotone class containing $\mathscr{F}_0$ we immediately have $m(\mathscr{F}_0) \subseteq \mathscr{G}_1$.

Let $\mathscr{G}_2 = \{B : A \cup B \in m(\mathscr{F}_0), \forall A \in m(\mathscr{F}_0)\}$

$\mathscr{G}_2$ **is a monotone class.** (identical proof- see problem 1.10)

Let $B \in \mathscr{F}_0$. Since $m(\mathscr{F}_0) \subseteq \mathscr{G}_1$ for any set $A \in m(\mathscr{F}_0) \Rightarrow A \cup B \in m(\mathscr{F}_0)$. Thus, by the definition of $\mathscr{G}_2 \Rightarrow B \in \mathscr{G}_2 \Rightarrow \mathscr{F}_0 \subseteq \mathscr{G}_2$.

The previous implication and the fact that $\mathscr{G}_2$ is a monotone class implies that $m(\mathscr{F}_0) \subseteq \mathscr{G}_2$.

Therefore, $\forall A, B \in m(\mathscr{F}_0) \Rightarrow A \cup B \in m(\mathscr{F}_0) \Rightarrow m(\mathscr{F}_0)$ is an algebra. □

[6] one can just verify the definition of $\mathscr{G}_1$ for this.

Problems

1.1. Roll a die. Then $\Omega = \{1,2,3,4,5,6\}$. An example of a event is $A = \{$ Roll an even number$\} = \{2,4,6\}$. Find the cardinality (number of elements) of $\mathscr{P}(\Omega)$ in this case.

1.2. Suppose two events A and B are in some space Ω. List the elements of the generated σ algebra $\sigma(A,B)$ in the following cases:
a) $A \cap B = \emptyset$
b) $A \subset B$
c) $A \cap B \neq \emptyset$; $A \setminus B \neq \emptyset$ and $B \setminus A \neq \emptyset$

1.3. An algebra which is not a σ-algebra
Let $\mathscr{B}_0$ be the collection of sets of the form: $(a_1, a_1'] \cup (a_2, a_2'] \cup \cdots \cup (a_m, a_m']$, for any $m \in \mathbb{N}^* = \{1,2\ldots\}$ and all $a_1 < a_1' < a_2 < a_2' < \cdots < a_m < a_m'$ in $\Omega = (0,1]$
Verify that $\mathscr{B}_0$ is an algebra. Show that $\mathscr{B}_0$ is not a σ-algebra.

1.4. Let $\mathscr{F} = \{A \subseteq \Omega | A$ finite **or** A^c is finite$\}$.
a) Show that $\mathscr{F}$ is an algebra
b) Show that if Ω is finite then $\mathscr{F}$ is a σ-algebra
c) Show that if Ω is infinite then $\mathscr{F}$ is **not** a σ-algebra

1.5. A σ-Algebra does not necessarily contain all the events in Ω
Let $\mathscr{F} = \{A \subseteq \Omega | A$ countable **or** A^c is countable$\}$. Show that $\mathscr{F}$ is a σ-algebra. Note that if Ω is uncountable implies that it contains a set A such that both A and A^c are uncountable thus $A \notin \mathscr{F}$.

1.6. Show that the Borel sets of $\mathbb{R}$ $\mathscr{B} = \sigma(\{(-\infty, x] | x \in \mathbb{R}\})$.

Hint: show that the generating set is the same i.e., show that any set of the form $(-\infty, x]$ can be written as countable union (or intersection) of open intervals and viceversa that any open interval in $\mathbb{R}$ can be written as countable union (or intersection) of sets of the form $(-\infty, x]$.

1.7. Show that the following classes all generate the Borel σ-algebra, or put differently show the equality of the following collections of sets:

$$\begin{aligned}\sigma((a,b) : a < b \in \mathbb{R}) = \sigma([a,b] : a < b \in \mathbb{R}) = \sigma((-\infty, b) : b \in \mathbb{R}) \\ = \sigma((-\infty, b) : b \in \mathbb{Q}),\end{aligned}$$

where $\mathbb{Q}$ is the set of rational numbers.

1.8. Properties of probability measures
Prove properties 1-4 in the Proposition 1.3 on page 13.

Hint: You only have to use the definition of probability. The only thing non-trivial in the definition is the countable additivity property.

1.9. No mater how many zeros do not add to more than zero
Prove the Lemma 1.2 on page 23.
Hint: You may use countable subadditivity.

1.10. If $\mathscr{F}_0$ is an algebra, $m(\mathscr{F}_0)$ is the minimal monotone class over $\mathscr{F}_0$ and $\mathscr{G}_2$ is defined as:

$$\mathscr{G}_2 = \{B : A \cup B \in m(\mathscr{F}_0), \forall A \in m(\mathscr{F}_0)\}$$

Then show that $\mathscr{G}_2$ is a monotone class.
Hint: Look at the proof of theorem 1.5 on page 29, and repeat the arguments therein.

1.11. A monotone algebra is a σ-algebra
Let $\mathscr{F}$ be an algebra that is also a monotone class. Show that $\mathscr{F}$ is a σ-algebra.

1.12. Prove the *total probability formula* equation (1.6) and the *Bayes Formula* equation 1.7.

1.13. If two events are such $A \cap B = \emptyset$ are A and B independent? Justify.

1.14. Show that $\mathbf{P}(A|B) = \mathbf{P}(A)$ is the same as independence of the events A and B.

1.15. Prove that if two events A and B are independent then so are their complements.

1.16. Generalize the previous problem to n sets using induction.

1.17. One urn contains w_1 white balls and b_1 black balls. Another urn contains w_2 white balls and b_2 black balls. A ball is drawn at random from each urn, then one of the two such chose are selected at random.
a) What is the probability that the final ball selected is white?
b) Given that the final ball selected was white what is the probability that in fact it came from the first urn (with w_1 and b_1 balls).

1.18. At the end of a well known course the final grade is decided with the help of an oral examination. There are a total of m possible subjects listed on some pieces of paper. Of them n are generally considered "easy".

Each student enrolled in the class, one after another, draws a subject at random then presents it. Of the first two students who has the better chance of drawing a "favorable" subject?

1.19. Suppose an event A has probability 0.3. How many independent trials must be performed to assert with probability 0.9 that the relative frequency of A differs from 0.3 by no more than 0.1.

1.20. Show using the Cantelli lemma that when you roll a die the outcome $\{1\}$ will appear infinitely often. Also show that eventually the average of all rolls up to roll n will be within ε of 3.5 where $\varepsilon > 0$ is any arbitrary real number.

1.21. Andre Agassi and Pete Sampras decide to play a number of games together. They play non-stop and at the end it turns out that Sampras won n games while Agassi m where $n > m$. Assume that in fact any possible sequence of games was possible to reach this result. Let $P_{n,m}$ denote the probability that from the first game until the last Sampras is always in the lead. Find:

1. $P_{2,1}$; $P_{3,1}$; $P_{n,1}$
2. $P_{3,2}$; $P_{4,2}$; $P_{n,2}$
3. $P_{4,3}$; $P_{5,3}$; $P_{5,4}$
4. Make a conjecture about a formula for $P_{n,m}$.

1.22. My friend Andrei has designed a system to win at the roulette. He likes to bet on red, but he waits until there have been 6 previous black spins and only then he bets on red. He reasons that the chance of winning is quite large since the probability of 7 consecutive back spins is quite small. What do you think of his system. Calculate the probability the he wins using this strategy.

Actually, Andrei plays his strategy 4 times and he actually wins three times out of the 4 he played. Calculate what was the probability of the event that just occurred.

1.23. Ali Baba is caught by the sultan while stealing his daughter. The sultan is being gentle with him and he offers Ali Baba a chance to regain his liberty.
There are 2 urns and m white balls and n black balls. Ali Baba has to put the balls in the 2 urns however he likes with the only condition that no urn is empty. After that the sultan will chose an urn at random then pick a ball from that urn. If the chosen ball is white Ali Baba is free to go, otherwise Ali Baba's head will be at the same level as his legs.
How should Ali Baba divide the balls to maximize his chance of survival?

References

Bertrand, J. L. F. (1889). *Calcul des probabilités*. Paris: Gauthier-Villars et fils.

Billingsley, P. (1995). *Probability and measure* (3 ed.). Wiley.

Blæsild, P. and J. Granfeldt (2002). *Statistics with Applications in Biology and Geology*. CRC Press.

Boyce, W. E. and R. C. DiPrima (2004). *Elementary Differential Equations and Boundary Value Problems* (8 ed.). Wiley.

Cauchy, A. L. (1821). *Analyse algébrique*. Imprimerie Royale.

Chung, K. L. (2000). *A Course in Probability Theory Revised* (2nd ed.). Academic Press.

Dembo, A. (2008). Lecture notes in probability. available on http://www-stat.stanford.edu/∼adembo/.

Good, I. J. (1986). Some statistical applications of poisson's work. *Statistical Science 1*(2), 157–170.

Gross, D. and C. M. Harris (1998). *Fundamentals of Queueing Theory*. Wiley.

Jona-Lasinio, G. (1985). *Some recent applications of stochastic processes in quantum mechanics*, Volume 1159 of *Lecture Notes in Mathematics*, pp. 130–241. Springer Berlin / Heidelberg.

Karlin, S. and H. M. Taylor (1975). *A first course in stochastic processes* (2 ed.). Academic Press.

Kingman, J. F. C. (1993). *Poisson processes*. Oxford University Press.

Lu, T.-C., Y.-S. Hou, and R.-J. Chen (1996). A parallel poisson generator using parallel prefix. *Computers & Mathematics with Applications 31*(3), 33 – 42.

Morgan, J. P., N. R. Chaganty, R. C. Dahiya, and M. J. Doviak (1991). Let's make a deal: The player's dilemma. *American Statistician 45*, 284–287.

Mueser, P. R. and D. Granberg (1991). The monty hall dilemma revisited: Understanding the interaction of problem definition and decision making. working paper 99-06, University of Missouri.

Øksendal, B. (2003). *Stochastic Differential Equations* (5 ed.). Springer Verlag.

Rosenthal, J. (2008, September). Monty hall, monty fall, monty crawl. *Math Horizons*, 5–7.

Ross, S. (1995). *Stochastic Processes* (2nd ed.). Wiley.

Chapter 2
Random Variables

All the definitions with sets presented in Chapter 1 are consistent, however if we wish to calculate and compute numerical values related to abstract spaces we need to standardize the spaces. The first step is to give the following definition.

Definition 2.1 (Measurable Function (m.f.)). Let $(\Omega_1, \mathscr{F}_1)$, $(\Omega_2, \mathscr{F}_2)$ be two measurable spaces. Let $f : \Omega_1 \longrightarrow \Omega_2$ be a function. f is called a measurable function if and only if for any set $B \in \mathscr{F}_2$ we have $f^{-1}(B) \in \mathscr{F}_1$. The inverse function is a set function defined in terms of the pre-image. Explicitly, for a given set $B \in \mathscr{F}_2$,

$$f^{-1}(B) = \{\omega_1 \in \Omega_1 : f(\omega_1) \in B\}$$

Note: This definition makes it possible to extend probability measures to other spaces. For instance, let f be a measurable function and assume that there exists a probability measure P_1 on the first space $(\Omega_1, \mathscr{F}_1)$. Then we can construct a probability measure on the second space $(\Omega_2, \mathscr{F}_2)$ by $(\Omega_2, \mathscr{F}_2, P_1 \circ f^{-1})$. Note that since f is measurable $f^{-1}(B)$ is in $\mathscr{F}_1$, thus $P_1 \circ f^{-1}(B) = P_1(f^{-1}(B))$ is well defined.

Reduction to $\mathbb{R}$. Random variables

Definition 2.2. Any measurable function with codomain $(\Omega_2, \mathscr{F}_2) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is called a random variable.

Consequence: Since the Borel sets in $\mathbb{R}$ are generated by $(-\infty, x]$ then we can have the definition of a random variable directly by:

$$f : \Omega_1 \longrightarrow \mathbb{R} \text{ such that } f^{-1}(-\infty, x] \in \mathscr{F} \text{ or } \{\omega : f(\omega) \leq x\} \in \mathscr{F}, \forall x \in \mathbb{R}.$$

We shall sometimes use $f(\omega) \leq x$ to denote $f^{-1}(-\infty, x)$. Traditionally, the random variables are denoted with capital letters from the end of the alphabet $X, Y, Z, \ldots$ and their values are denoted with corresponding small letters $x, y, z, \ldots$.

Definition 2.3 (Distribution of Random Variable). Assume that on the measurable space $(\Omega, \mathscr{F})$ we define a probability measure $\mathbf{P}$ so that it becomes a probability space $(\Omega, \mathscr{F}, \mathbf{P})$. If a random variable $X : \Omega \to \mathbb{R}$ is defined then we call its distribution, the set function μ defined on the Borel sets of $\mathbb{R}$: $\mathscr{B}(\mathbb{R})$, with values in $[0,1]$:

$$\mu(B) = \mathbf{P}(\{\omega : X(\omega) \in B\}) = \mathbf{P}\left(X^{-1}(B)\right) = \mathbf{P} \circ X^{-1}(B)$$

Remark 2.1. First note that the measure μ is defined on sets in $\mathbb{R}$ and takes values in the interval $[0,1]$. Therefore, the random variable X allows us to apparently eliminate the abstract space Ω. However, this is not the case since we still have to calculate probabilities using $\mathbf{P}$ in the definition of μ above.

However, there is one simplification we can make. If we recall the result of the exercises 1.6 and1.7, we know that all Borel sets are generated by the same type of sets. Using the same idea as before it is enough to describe how to calculate μ for the generators. We could of course specify any type of generating sets we wish (open sets, closed sets, etc) but it turns out the simplest way is to use sets of the form $(-\infty, x]$, since we only need to specify one end of the interval (the other is always $-\infty$). With this observation we only need to specify the measure $\mu = P \circ X^{-1}$ directly on the generators to completely characterize the probability measure.

Definition 2.4. [The distribution function of a random variable] The distribution function of a random variable X is $F : \mathbb{R} \to [0,1]$ with:

$$F(x) = \mu(-\infty, x] = \mathbf{P}(\{\omega : X(\omega) \in (-\infty, x]\}) = \mathbf{P}(\{\omega : X(\omega) \leq x\})$$

But wait a minute, this is exactly the definition of the cumulative distribution function (cdf) which you can find in any lower level probability classes. It is exactly the same thing except that in an effort to dumb down (in whomever opinion it was to teach the class that way) the meaning is lost and we cannot proceed with more complicated things. From the definition above we can deduce all the elementary properties of the cdf that you have learned (right-continuity, increasing, taking values between 0 and 1). In fact let me ask you to prove this in exercise .

Proposition 2.1. *The distribution function for any random variable X has the following properties:*

(i) F is increasing (i.e. if $x \leq y$ then $F(x) \leq F(y)$)[1]
(ii) F is right continuous (i.e. $\lim_{h \downarrow 0} F(x+h) = F(x)$)
(iii) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$

Example 2.1 (Indicator random variable). Recall the indicator function from Definition 1.10. Let $\mathbf{1}_A$ be the indicator function of a set $A \subseteq \Omega$. This is a function

[1] In other math books a function with this property is called non-decreasing. I do not like the negation and I prefer to call a function like this increasing with the distinction that a function with the following property $x < y$ implies $F(x) < F(y)$ is going to be called a **strictly increasing** function

defined on Ω with values in $\mathbb{R}$. Therefore, it may be a random variable. According to the definition it is a random variable if the function is measurable. It is simple to show that this happens if and only if $A \in \mathscr{F}$ the σ-algebra associated with the probability space. Assuming that $A \in \mathscr{F}$, what is the distribution function of this random variable?

According to the definition we have to calculate $\mathbf{P} \circ \mathbf{1}_A^{-1}((-\infty, x])$ for any x. However, the function 1_A only takes two values 0 and 1. We can calculate immediately:

$$\mathbf{1}_A^{-1}((-\infty, x]) = \begin{cases} \emptyset & , \text{ if } x < 0 \\ A^c & , \text{ if } x \in [0,1) \\ \Omega & , \text{ if } x > 1 \end{cases}.$$

Therefore,

$$F(x) = \begin{cases} 0 & , \text{ if } x < 0 \\ \mathbf{P}(A^c) & , \text{ if } x \in [0,1) \\ 1 & , \text{ if } x \geq 1 \end{cases}.$$

Proving the following lemma is elementary using the properties of the probability measure (Proposition 1.3) and is left as an exercise.

Lemma 2.1. *Let F be the distribution function of X. Then:*

(i) $\mathbf{P}(X \geq x) = 1 - F(x)$
(ii) $\mathbf{P}(x < X \leq y) = F(x) - F(y)$
(iii) $\mathbf{P}(X = x) = F(x) - F(x-)$, *where* $F(x-) = \lim_{y \nearrow x} F(y)$ *the left limit of F at x.*

Above, we define a random variable as a measurable function with codomain $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. A more specific case is obtained when the random variable has the domain also equal to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. In this case the random variable is called a Borel function.

Definition 2.5 (Borel measurable function). A function $g : \mathbb{R} \to \mathbb{R}$ is called Borel (measurable) function if g is a measurable function from $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ into $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

Example 2.2. Show that any continuous function $g : \mathbb{R} \to \mathbb{R}$ is Borel measurable.

Solution 2.1. This is very simple. Recall that the Borel sets are generated by open sets. So it is enough to see what happens to the pre-image of a open set B. But g is a continuous function therefore $g^{-1}(B)$ is an open set and thus $g^{-1}(B) \in \mathscr{B}(\mathbb{R})$. Therefore by definition g is Borel measurable.

2.1 Discrete and Continuous Random Variables

Definition 2.6 (pdf pmf and all that). Note that the distribution function F always exists. In general the distribution function F is not necessarily derivable. However, if it is, we call its derivative $f(x)$ the *probability density function* (pdf):

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

Traditionally, a variable X with this property is called *a continuous random variable*.

Furthermore if F is piecewise constant (i.e., constant almost everywhere), or in other words there exist a countable sequence $\{a_1, a_2, \dots\}$ such that the function F is constant for every point except these a_i's and we denote $p_i = F(a_i) - F(a_i-)$, then the collection of p_i's is the traditional *probability mass function* (pmf) that characterizes a *discrete random variable*[2].

Remark 2.2. Traditional undergraduate textbooks segregate between discrete and continuous random variables. Because of this segregation they are the only variables presented and it appears that all the random variables are either discrete or continuous. In reality these are the only types that can be presented without following the general approach we take here. The definitions we presented here cover any random variable. Furthermore, the treatment of random variables is the same, no more segregation.

Important. So what is the point of all this? What did we just accomplish here?

The answer is: we successfully moved from the abstract space $(\Omega, \mathscr{F}, P)$ to something perfectly equivalent but defined on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Because of this we only need to define probability measures on $\mathbb{R}$ and show that anything coming from the original abstract space is equivalent with one of these distributions on $\mathbb{R}$. We have just constructed our first model.

Example 2.3 (Indicator r.v. (continued)). This indicator variable is also called the Bernoulli random variable. Notice that the variable only takes values 0 and 1 and the probability that the variable takes the value 1 may be easily calculated using the previous definitions:

$$\mathbf{P} \circ \mathbf{1}_A^{-1}(\{1\}) = \mathbf{P}\{\omega : \mathbf{1}_A(\omega) = 1\} = \mathbf{P}(A).$$

Therefore the variable is distributed as a Bernoulli random variable with parameter $p = \mathbf{P}(A)$. Alternately, we may obtain this probability using the previously computed distribution function:

$$\mathbf{P}\{\omega : \mathbf{1}_A(\omega) = 1\} = F(1) - F(1-) = 1 - \mathbf{P}(A^c) = \mathbf{P}(A)$$

Example 2.4. Roll a six sided fair die. Say $X(\omega) = 1$ if the die shows 1 ($\omega = 1$), $X = 2$ if the die shows 2, etc. Find $F(x) = \mathbf{P}(X \leq x)$.

Solution 2.2 (Solution).

If $x < 1$ then $\mathbf{P}(X \leq x) = 0$

[2] Again we used the notation $F(x-)$ for the left limit of function F at x or in a more traditional notation $\lim_{z \to x, z<x} F(z)$.

If $x \in [1,2)$ then $\mathbf{P}(X \leq x) = \mathbf{P}(X = 1) = 1/6$
If $x \in [2,3)$ then $\mathbf{P}(X \leq x) = \mathbf{P}(X(\omega) \in \{1,2\}) = 2/6$

We continue this way to get:

$$\mathbf{F}(x) = \begin{cases} 0 \text{ if } x < 1 \\ i/6 \text{ if } x \in [i, i+1) \text{ with } i = 1, \cdots, 5 \\ 1 \text{ if } x \geq 6 \end{cases}$$

Exercise 2.1 (Mixture of continuous and discrete random variable). Say a game asks you to toss a coin. If the coin lands Tail you lose 1$, if Head then you draw a number from $[1,2]$ at random and gain that number. Furthermore, suppose that the coins lands a Head with probability p. Let X be the amount of money won or lost after 1 game. Find the distribution of X.

Solution 2.3 (Solution). Let $\omega = (\omega_1, \omega_2)$ where $\omega_1 \in \{\text{Head}, \text{Tail}\}$ and ω_2 in the defining experiment space for the Uniform distribution. New define a random variable $Y(\omega_2)$ on the uniform $[1,2]$ space. Then the random variable X is defined as:

$$X(\omega) = \begin{cases} -1 & , \text{ if } \omega_1 = \text{Tail} \\ Y(\omega_2) & \text{if } \omega_1 = \text{Head} \end{cases}$$

If $x \in [-1,1)$ we get :

$$\mathbf{P}(X \leq x) = \mathbf{P}(X = -1) = \mathbf{P}(\omega_1 = \text{Tail}) = 1 - p$$

If $x \in [1,2)$ we get:

$$\begin{aligned} \mathbf{P}(X \leq x) &= \underbrace{\mathbf{P}(X = -1 \text{ or } X \in [1,x))}_{\text{the two events are disjoint}} = 1 - p + \mathbf{P}(\omega_1 = \text{ heads }, Y \leq x) \\ &= 1 - p + p \underbrace{\mathbf{P}(Y \in [1,x))}_{\text{Uniform}[1,2]} \\ &= 1 - p + p \int_1^x 1 dy = 1 - p + p(x-1) \\ &= 1 - 2p + px. \end{aligned}$$

Note that if the two parts of the game are not independent of each-other we cannot calculate this distribution.

Finally, we obtain:

$$\mathbf{F}(x) = \begin{cases} 0 \text{ if } x < -1 \\ 1 - p \text{ if } x \in [-1,1) \\ 1 - 2p + px \text{ if } x \in [1,2) \\ 1 \text{ if } x \geq 2 \end{cases}$$

Checking that our calculation is correct It is always a good idea to check the result. We can verify the distribution function properties, and we can plot the function to confirm this.

Examples of commonly encountered Random Variables:

Discrete random variables

For discrete random variables we give the probability mass function and it will describe completely the distribution (recall that the distribution function is piecewise linear).

(i) *Bernoulli Distribution,* the random variable only takes two values:

$$\mathbf{X} = \begin{cases} 1 & \text{with } \mathbf{P}(X=1) = p \\ 0 & \text{with } \mathbf{P}(X=0) = 1-p \end{cases}$$

We denote a random variable X with this distribution with $X \sim Bernoulli(p)$.

(ii) *Binomial*(n,p) *distribution,* the random variable takes values in $\mathbb{N}$ with:

$$\mathbf{P}(X=k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} \text{ for any } k \in \{0,1,2,\ldots,n\} \\ 0 \text{ otherwise} \end{cases}$$

Note: X has the same distribution as $Y_1 + \cdots Y_n$ where $Y_i \sim Bernoulli(p)$

We denote a random variable X with this distribution with $X \sim Binom(n,p)$.

(iii) *Geometric* (p) *distribution*:

$$\mathbf{P}(X=k) = \begin{cases} (1-p)^{k-1} p \text{ for any } k \in \{1,2\cdots\} \\ 0 \text{ otherwise} \end{cases}$$

This is sometimes called Geometric "number of trials" distribution. We can also talk about Geometric "number of failures distribution" distribution, defined:

$$\mathbf{P}(Y=k-1) = \begin{cases} (1-p)^{k-1} p \text{ for any } k \in \{1,2\cdots\} \\ 0 \text{ otherwise} \end{cases}$$

Most of the time when we write $X \sim Geometric(p)$ we mean that X has a Geometric number of trials distribution. In the rare cases when we use the other one we will specify very clearly.

(iv) *Negative Binomial* (r,p) *distribution*

$$\mathbf{P}(X=k) = \begin{cases} \binom{k-1}{r-1} (1-p)^{r-k} p^r \text{ for any } k \in \{r, r+1, \ldots\} \\ 0 \text{ otherwise} \end{cases}$$

Similarly with the $Geometric(p)$ distribution we can talk about "number of failures" distribution, but I will not give that definition.

Let us stop for a moment and see where these distributions are coming from. Suppose we do a simple experiment, we repeat an experiment many times. This experiment only has two possible outcomes "success" with probability p and "failure" with probability $1-p$.

- The variable X that takes value 1 if the experiment is a success and 0 otherwise has a $Bernoulli(p)$ distribution.
- Repeat the experiment n times in such a way that no experiment influences the outcome of any other experiment[3] and we count how many of the n repetition actually resulted in success. Let Y be the variable denoting this number. Then $Y \sim Binom(n,p)$.
- If instead of repeating the experiment a fixed number of times we repeat the experiment as many times as are needed to see the first success, then the number of trials needed is going to be distributed as a $Geometric(p)$ random variable. If we count failures until the first success we obtain the $Geometric(p)$ "number of failures" distribution.
- If we repeat the experiment until we see r successes, the number of trials needed is a $NegativeBinomial(r,p)$

(v) *Hypergeometric distribution(N,m,n,p)*,

$$\mathbf{P}(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \; k \in \{0,1\cdots m\}$$

This may be thought of as drawing n balls from an urn containing m white balls and $N-m$ black balls, where X represents the number of white balls in the sample.

(vi) *Poisson Distribution*, the random variable takes values in $\mathbb{N}$,

$$\mathbf{P}(X=k) = \frac{\lambda^k}{k!}e^{-k}, \quad k=0,1,2,\ldots$$

Continuous Random Variables.

In this case every random variable has a pdf and we will specify this function directly.

(i) *Uniform Distribution[a,b]*, the random variable represents the position of a point taken at random (without any preference) within the interval $[a,b]$.

[3] this is the idea of independence which we will discuss a bit later

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ , if } x \in [a,b] \\ 0 \text{ , otherwise} \end{cases}$$

(ii) *Exponential Distribution*(θ)

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0$$

(iii) *Normal Distribution*(μ, σ)

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

There are many more distributions, for our purpose the few presented will suffice.

A special random variable: Diraç Delta distribution

For a fixed a real number, consider the following distribution function:

$$F_\delta(x) = \begin{cases} 0 & \text{if } x < a \\ 1 & \text{if } x \geq a \end{cases}$$

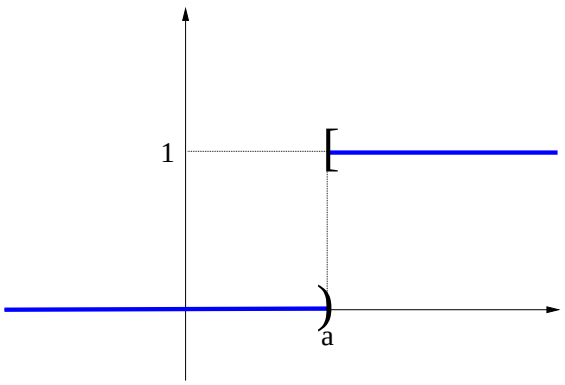


Fig. 2.1 A distribution function.

This function is plotted in Figure 2.1. Note that the function has all the properties of a distribution function (increasing, right continuous and limited by 0 and 1). However, the function is not derivable (the distribution does not have a pdf).

The random variable with this distribution is called a Diraç impulse function at a. It can only be described using measures. We will come back to this function when we develop the integration theory but for now let us say that if we define the associated set function:

$$\delta_{\{a\}}(A) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{otherwise} \end{cases}$$

this is in fact a probability measure with the property:

$$\int_{-\infty}^{\infty} f(x)d\delta_{\{a\}}(x) = f(a), \quad \text{for all continuous functions } f$$

This will be written later as $\mathbf{E}^{\delta_{\{a\}}}[f] = f(a)$. (In other sciences: $\delta_{\{a\}}(f) = f(a)$).

Also note that $\delta_{\{a\}}(A)$ is a set function (a is fixed) and has the same value as the indicator $\mathbf{1}_A(a)$ which is a regular function (A is fixed).

2.2 Existence of random variables with prescribed distribution. Skorohod representation of a random variable

In the previous section we have seen that any random variable has a distribution function F, what is called in other classes the c.d.f. Recall the essential properties of this function from Proposition 2.1 on page 36: right-continuity, increasing, taking values between 0 and 1. An obvious question is given a function F with these properties can we construct a random variable with the desired distribution?

In fact yes we can and this is the first step in a very important theorem we shall see later in this course: the Skorohod representation theorem. However, recall that a random variable has to have as domain some probability space. It actually is true that we can construct random variables with the prescribed distribution on any space but recall that the purpose of creating random variables was to have a uniform way of treating probability. It is actually enough to give the Skorohod's construction on the probability space $([0,1],\mathscr{B}([0,1]),\lambda)$, where λ is the Lebesque measure.

On this space define the following random variables:

$$X^+(\omega) = \inf\{z \in \mathbb{R} : F(z) > \omega\}$$
$$X^-(\omega) = \inf\{z \in \mathbb{R} : F(z) \geq \omega\}$$

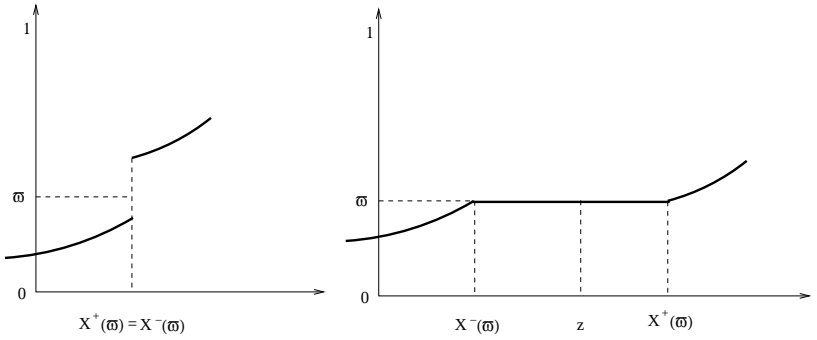
Note that in statistics X^- would be called the ω-quantile of the distribution F.

For most of the outcomes ω the two random variables are identical. Indeed, if at z with $\omega = F(z)$ the function F is non-constant then the two variables take the same values $X^+(\omega) = X^-(\omega) = z$. The two important cases when the variables take different values are depicted in Figure 2.2.

We need to show that the two variables have the desired distribution. To this end let $x \in \mathbb{R}$. Then we have:

$$\{\omega \in [0,1] : X^-(\omega) \leq x\} = [0, F(x)]$$

Indeed, if ω is in the left set then $X^-(\omega) \leq x$. By the definition of X^- then $\omega \leq F(x)$ and we have the inclusion $\subseteq$. If on the other hand $\omega \in [0,F(x)]$ then $\omega \leq F(x)$ and again by definition and right continuity of F, $X^-(\omega) \leq x$, thus we obtain $\supseteq$. Therefore, the distribution is:



(a) A point of discontinuity for F (b) An interval where the function F is constant

Fig. 2.2 Points where the two variables $X^{\pm}$ may have different outcomes

$$\lambda(\{\omega \in [0,1] : X^-(\omega) \leq x\}) = \lambda([0,F(x)]) = F(x) - 0 = F(x).$$

Finally, X^+ also has distribution function F and furthermore:

$$\lambda(X^+ \neq X^-) = 0.$$

By definition of X^+:

$$\{\omega \in [0,1] : X^-(\omega) \leq x\} \supseteq [0,F(x)),$$

and so $\lambda(X^+ \leq x) \geq F(x)$. Furthermore, since $X^- \leq X^+$ we have:

$$\{\omega \in \mathbb{R} : X^-(\omega) \neq X^+(\omega)\} = \bigcup_{x \in \mathbb{Q}} \{\omega \in \mathbb{R} : X^-(\omega) \leq x < X^+(\omega)\}$$

But for every such $x \in \mathbb{Q}$:

$$\lambda(\{\omega \in \mathbb{R} : X^-(\omega) \leq x < X^+(\omega)\}) = \lambda(\{X^- \leq x\} \setminus \{X^+ \leq x\}) \leq F(x) - F(x) = 0$$

Since $\mathbb{Q}$ is countable and any countable union of null sets is a null set the result follows.

2.3 Independence

In this section we extend the idea of independence originally defined for events to random variables. In order to do this we have to explain the joint distribution of several variables.

Example 2.5 (The idea of joint distribution). Suppose 2 points ξ_1, ξ_2 are tossed at random and independently onto a line segment of length L (ξ_1, ξ_2 are i.i.d.). What is the probability that the distance between the 2 points does not exceed 1?

Solution 2.4 (Solution). If $L \leq 1$ then the probability is trivially equal to 1.

Assume that $L > 1$ (the following also works if 1 is substituted by a $l \leq L$). What is the distribution of ξ_1 and ξ_2? They are both $Unif[0,L]$. We want to calculate $\mathbf{P}(|\xi_1 - \xi_2| \leq 1)$.

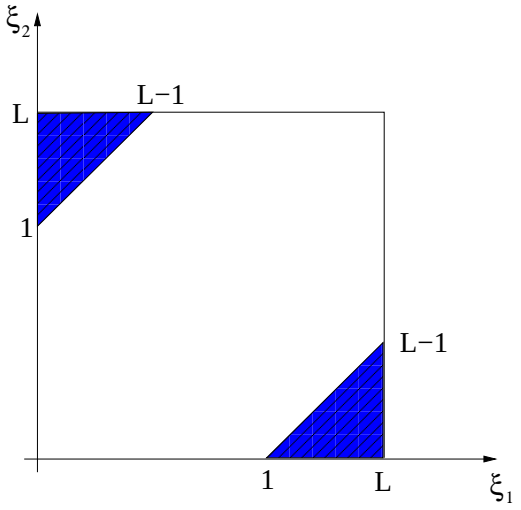


Fig. 2.3 The area we need to calculate. The blue parts need to be deleted.

We plot the surface we need to calculate in Figure 2.3. The area within the rectangle and not shaded is exactly the area we need. If we pick any point from within this area it will have the property that $|\xi_1 - \xi_2| \leq 1$. Since the points are chosen uniformly from within the rectangle the chance of a point being chosen is the ratio between the "good" area and the total area.

The unshaded area from within the rectangle is: $L^2 - \frac{(L-1)^2}{2} - \frac{(L-1)^2}{2} = 2L - 1$. Therefore, the desired probability is:

$$\mathbf{P}(|\xi_1 - \xi_2| \leq 1) = \frac{2L-1}{L^2}.$$

□

This geometrical proof works because the distribution is uniform and furthermore the points are chosen independently of each other. However if the distribution is anything else we need to go through the whole calculation. We shall see how to do this after we define joint probability. We need this to define the independence concept.

2.3.1 Joint distribution

We talked about σ-algebras in Chapter 1. Let us come back to them. If there is any hope of rigorous introduction into probability and stochastic processes, they are *unavoidable*. Later, when we will talk about stochastic processes we will find out the *crucial* role they play in quantifying the information available up to a certain time. For now let us play a bit with them.

Definition 2.7 (σ-algebra generated by a random variable). For a r.v. X we define the σ-algebra generated by X, denoted $\sigma(X)$ or sometime $\mathscr{F}_X$, the smallest σ-field $\mathscr{G}$ such that X is measurable on $(\Omega,\mathscr{G})$. It is the σ-algebra generated by the pre-images of Borel sets through X (recall that we have already presented this concept earlier in definition 1.3 on page 9). Because of this we can easily show[4]:

$$\sigma(X) = \sigma(\{\omega | X(\omega) \leq x\}, \text{ as } x \text{ varies in } \mathbb{R}).$$

Similarly, given $X_1, X_2, \ldots, X_n$ random variables, we define the sigma algebra generated by them as the smallest sigma algebra such that all are measurable with respect to it. It turns out we can show easily that it is the sigma algebra generated by the union of the individual sigma algebras or put more specifically $\sigma(X_i, i \leq n)$ is the smallest sigma algebra containing all $\sigma(X_i)$, for $i = 1,2,\ldots,n$, or $\sigma(X_1) \vee \sigma(X_2) \vee \cdots \vee \sigma(X_n)$, again recall proposition 1.2 on page 10.

In Chapter 1 we defined Borel sigma algebras corresponding to any space Ω. We consider the special case when $\Omega = \mathbb{R}^n$. This allows us to define a random vector on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbf{P})$ as $(X_1, X_2, \ldots, X_n)$ where each X_i is a random variable. The probability $\mathbf{P}$ is defined on $\mathscr{B}(\mathbb{R}^n)$.

We can talk about its distribution (the *"joint distribution"* of the variables $(X_1, X_2, \ldots, X_n)$) as the function:

$$\begin{aligned} F(x_1, x_2, \ldots, x_n) &= \mathbf{P} \circ (X_1, X_2, \ldots, X_n)^{-1} ((-\infty, x_1] \times \cdots \times (-\infty, x_n]) \\ &= \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \end{aligned}$$

which is well defined for any $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$

In the special case when F can be written as:

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_X(t_1, \cdots, t_n) dt_1 \cdots dt_n,$$

we say that the vector X has a *joint density* and f_X is the joint probability density function of the random vector X.

[4] Remember that the Borel sets are generated by intervals of the type $(-\infty, x]$

Definition 2.8 (Marginal Distribution). Given the joint distribution of a random vector $X = (X_1, X_2, \dots, X_n)$ we define the marginal distribution of X_1:

$$F_{X_1}(x_1) = \lim_{\substack{x_2 \to \infty \\ \dots \\ x_n \to \infty}} F_X(x_1 \cdots x_n)$$

and similarly for all the other variables.[5]

2.3.2 Independence of random variables

We can now introduce the notions of independence and joint independence using the definition in Section 1.3, the probability measure $= \mathbf{P} \circ (X_1, X_2, \dots, X_n)^{-1}$ and any Borel sets. Writing more specifically that definition is transformed here:

Definition 2.9. The variables $(X_1, X_2, \dots, X_n, \dots)$ are independent if for every subset $J = \{j_1, j_2, \dots, j_k\}$ of $\{1, 2, 3, \dots\}$ we have:

$$\mathbf{P}\left(X_{j_1} \le x_{j_1}, X_{j_2} \le x_{j_2}, \dots, X_{j_k} \le x_{j_k}\right) = \prod_{j \in J} \mathbf{P}(X_j \le x_j)$$

Remark 2.3. The formula in the Definition 2.8 allows to obtain the marginal distributions from the joint distribution. The converse is generally false meaning that if we know the marginal distributions we cannot regain the joint.

However, there is one case when this is possible: when X_i are independent. In this case $F_X(x) = \prod_{i=1}^n F_{X_i}(x_i)$. That is why the i.i.d case is the most important in probability (we can regain the joint from the marginals without any other special knowledge).

Independence (specialized cases)

(i) If X and Y are discrete r.v.'s with joint probability mass function $p_{X,Y}(\cdot, \cdot)$ then they are independent if and only if

$$p_{X,Y}(x,y) = p_X(x) p_Y(y), \quad \forall x, y$$

(ii) If X and Y are continuous r.v.'s with joint probability density function f then they are independent if and only if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \forall x, y$$

where we used obvious notations for marginal distributions. The above definition can be extended to n dimensional vectors in an obvious way.

[5] We can also define it simpler as $\int_{-\infty}^{x_1} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(t_1, \cdots, t_n) dt_1 \cdots dt_n$ if the joint pdf exists.

I.I.D. r.v.'s: (Independent Identically Distributed Random Variables). Many of the central ideas in probability involve sequences of random variables which are independent and identically distributed. That is a sequence of random variables $\{X_n\}$ such that X_n are independent and all have the same distribution function say $F(x)$.

Finally, we answer the question we asked in the earlier example: What to do if the variables ξ_1, ξ_2 are not uniformly distributed?

Suppose that ξ_1 had distribution F_{ξ_1} and ξ_2 had distribution F_{ξ_2}. Assuming that the two variables are independent we obtain the joint distribution:

$$F_{\xi_1\xi_2}(x_1,x_2) = F_{\xi_1}(x_1)F_{\xi_2}(x_2)$$

(If they are not independent we have to be given or infer the joint distribution).

The probability we are looking for is the area of the surface

$$\{(\xi_1,\xi_2)|\xi_1 \in [0,L], \xi_2 \in [0,L], \xi_1 - 1 \leq \xi_2 \leq \xi_1 + 1\}.$$

We shall find out how to calculate this probability using general distribution functions F_{ξ_1} and F_{ξ_2} in the next chapter. For now let us assume that the two variables have densities f_1 and f_2. Then, the desired probability is:

$$\int_0^L \int_0^L \mathbf{1}_{\{x_1-1\leq x_2\leq x_1+1\}}(x_1,x_2) f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_1 dx_2$$

which can be further calculated:

- When $L-1<1$ or $1<L<2$:

$$\int_0^{L-1}\int_0^{x_1+1} f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_2 dx_1 + (2-L)L + \int_1^L \int_{x_1-1}^L f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_2 dx_1$$

- When $L-1>1$ or $L>2$:

$$\int_0^1\int_0^{x_1+1} f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_2 dx_1 + \int_1^{L-1}\int_{x_1-1}^{x_1+1} f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_2 dx_1 + \int_{L-1}^L \int_{x_1-1}^L f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_2 dx_1$$

Above is given to remind about the calculation of a two dimensional integral.

2.4 Functions of random variables. Calculating distributions

Measurable functions allow us to construct new random variables. These new random variables possess their own distribution. This section is dedicated to calculating this new distribution. At this time it is not possible to work with abstract spaces (for that we will give a general theorem - the Transport formula in the next chapter) so all our calculations will be done in $\mathbb{R}^n$.

One dimensional functions

Let X be a random variable defined on some probability space $(\Omega, \mathscr{F}, \mathbf{P})$. Let $g : \mathbb{R} \longrightarrow \mathbb{R}$ be a Borel measurable function. Let $Y = g(X)$ which is a new random variable. Its distribution is deduced as:

$$\begin{aligned}\mathbf{P}(Y \leq y) &= \mathbf{P}(g(X) \leq y) = \mathbf{P}(g(X) \in (-\infty, y]) = \mathbf{P}\left(X \in g^{-1}((-\infty, y])\right) \\ &= \mathbf{P}\left(\{\omega : X(\omega) \in g^{-1}((-\infty, y])\}\right)\end{aligned}$$

where $g^{-1}((-\infty, y])$ is the preimage of $(-\infty, y]$ through the function g, i.e.,:

$$\{x \in \mathbb{R} : g(x) \leq y\}.$$

If the random variable X has p.d.f f then the probability has a simpler formula:

$$\mathbf{P}(Y \leq y) = \int_{g^{-1}(-\infty, y]} f(x)dx$$

Example 2.6. Let X be a random variable distributed as a Normal (Gaussian) with mean zero and variance 1, $X \sim N(0,1)$. Let $g(x) = x^2$, and take $Y = g(X) = X^2$. Then:

$$\mathbf{P}(Y \leq y) = \mathbf{P}(X^2 \leq y) = \begin{cases} 0 & \text{if } y < 0 \\ \mathbf{P}(-\sqrt{y} \leq X \leq \sqrt{y}) & \text{if } y \geq 0 \end{cases}$$

Note that the preimage of $(-\infty, y]$ through the function $g(x) = x^2$ is either $\emptyset$ if $y < 0$ or $[-\sqrt{y}, \sqrt{y}]$ if $y \geq 0$. This is how we obtain above. In the nontrivial case $y \geq 0$ we get:

$$\mathbf{P}(Y \leq y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - [1 - \Phi(\sqrt{y})] = 2\Phi(\sqrt{y}) - 1,$$

where Φ is the c.d.f of X, a $N(0,1)$ random variable. In this case $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. Since the function Φ is derivable Y has a p.d.f. which can be obtained:

$$f_Y(y) = \frac{d}{dy}[2\Phi(\sqrt{y})] = 2\Phi'(\sqrt{y})\frac{1}{2\sqrt{y}}$$
$$= \frac{1}{\sqrt{y}}\Phi'(\sqrt{y}) = \frac{1}{\sqrt{y}}\frac{1}{\sqrt{2\pi}}e^{-y/2}$$
$$= \frac{1}{\sqrt{2\pi y}}e^{-y/2}$$

□

We note that a random variable Y with the p.d.f. described above is said to have a chi-squared distribution with one degree of freedom (the notation is χ^2_1).

Two and more dimensional functions

If the variable X does not have a p.m.f or a p.d.f there is not much we can do. The same relationship holds as in the 1 dimensional case. Specifically, if X is a n-dim random vector and $g : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a measurable function which defines a new random vector $Y = g(X)$ then its distribution is determined using:

$$\mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}\left(\{\omega : X(\omega) \in g^{-1}((-\infty, y])\}\right)$$

and this is the same relationship as before.

In the case when the vector X has a density then things become more specific. We will exemplify using R^2 but the same calculation works in n dimensions with no modification (other than the dimension of course). Suppose that a two dimensional random vector (X_1, X_2) has joint density f. Let $g : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ be a measurable function:

$$g(x_1, x_2) = (g_1(x_1, x_2), g_2(x_1, x_2))$$

Suppose first that the function g is one-to-one[6]

Define a random vector $Y = (Y_1, Y_2) = g(X_1, X_2)$. First we find the support set of Y (i.e. the points where Y has nonzero probability). To this end let

$$\mathcal{A} = \{(x_1, x_2) : f(x_1, x_2) > 0\}$$
$$\mathcal{B} = \{(y_1, y_2) : y_1 = g_1(x_1, x_2) \text{ and } y_2 = g_2(x_1, x_2), \text{ for some } (x_1, x_2) \in \mathcal{A}\}$$

This $\mathcal{B}$ is the image of $\mathcal{A}$ through g, it is also the support set of Y. Since g is one-to-one, when restricted to $g : \mathcal{A} \to \mathcal{B}$ it is also surjective, therefore forms a bijection between $\mathcal{A}$ and $\mathcal{B}$. Thus, the inverse function $g^{-1}(y_1, y_2) = (g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2))$ is a unique, well defined function.

[6] this is why we use the same dimension n for both X and Y vectors

To calculate the density of Y we need the derivative of this g^{-1} and that role is played by the Jacobian of the transformation (the determinant of the matrix of partial derivatives):

$$J = J_{g^{-1}}(y_1, y_2) = \begin{vmatrix} \frac{\partial g_1^{-1}}{\partial y_1}(y_1, y_2) & \frac{\partial g_2^{-1}}{\partial y_1}(y_1, y_2) \\ \frac{\partial g_1^{-1}}{\partial y_2}(y_1, y_2) & \frac{\partial g_2^{-1}}{\partial y_2}(y_1, y_2) \end{vmatrix}$$

Then, the joint p.d.f. of the vector Y is given by:

$$f_Y(y_1, y_2) = f\left(g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)\right) \, |J| \, \mathbf{1}_{\mathcal{B}}(y_1, y_2)$$

where we used the indicator notation and $|J|$ is the absolute value of the Jacobian.

Suppose that the function g is not one-to-one

In this case we recover the previous one-to-one case by restricting the function. Specifically, define the sets $\mathcal{A}$ and $\mathcal{B}$ as before. Now, the restricted function $g : \mathcal{A} \to \mathcal{B}$ is surjective. We partition $\mathcal{A}$ into $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$. The set $\mathcal{A}_0$ may contain several points which are difficult to deal with, the only condition is that $\mathbf{P}((X_1, X_2) \in \mathcal{A}_0) = 0$ (it is a null set). Furthermore, for all $i \neq 0$, each restriction $g : \mathcal{A}_i \to \mathcal{B}$ is one-to one. Thus, for each such $i \geq 1$, an inverse can be found $g_i^{-1}(y_1, y_2) = (g_{i1}^{-1}(y_1, y_2), g_{i2}^{-1}(y_1, y_2))$. This i-th inverse gives for any $(y_1, y_2) \in \mathcal{B}$ a unique $(x_1, x_2) \in \mathcal{A}_i$ such that $(y_1, y_2) = g(x_1, x_2)$. Let J_i be the Jacobian associated with the i-th inverse transformation. Then the joint p.d.f. of Y is:

$$f_Y(y_1, y_2) = \sum_{i=1}^{k} f\left(g_{i1}^{-1}(y_1, y_2), g_{i2}^{-1}(y_1, y_2)\right) \, |J_i| \, \mathbf{1}_{\mathcal{B}}(y_1, y_2)$$

Example 2.7. Let (X_1, X_2) have some joint p.d.f. $f(\cdot, \cdot)$. Calculate the density of $X_1 X_2$.

Let us take $Y_1 = X_1 X_2$ and $Y_2 = X_1$ i.e. $g(x_1, x_2) = (x_1 x_2, \, x_1) = (y_1, y_2)$. The function thus constructed $g : \mathbb{R}^2 \to \mathbb{R}^2$ is bijective so $\mathcal{B} = \mathbb{R}^2$. To calculate its inverse:

$$\begin{aligned} x_1 &= y_2 \\ x_2 &= \frac{y_1}{x_1} = \frac{y_1}{y_2}, \end{aligned}$$

which gives:

$$g^{-1}(y_1, y_2) = \left(y_2, \frac{y_1}{y_2}\right)$$

We then get the Jacobian:

$$J_{g^{-1}}(y_1,y_2) = \begin{vmatrix} 0 & \frac{1}{y_2} \\ 1 & -\frac{y_1}{y_2^2} \end{vmatrix} = 0 - \frac{1}{y_2} = -\frac{1}{y_2}$$

Thus, the joint p.d.f of $Y = (Y_1, Y_2)$ is:

$$f_Y(y_1,y_2) = f\left(y_2, \frac{y_1}{y_2}\right)\left|\frac{1}{y_2}\right|,$$

where f is the given p.d.f. of X. To obtain the distribution of $X_1X_2 = Y_1$ we simply need the marginal p.d.f. obtained immediately by integrating out Y_2:

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f\left(y_2, \frac{1}{y_2}\right) \cdot \frac{1}{|y_2|} dy_2$$

□

Example 2.8 (A more specific example). Let X_1, X_2 be independent Exp(λ). Find the joint density of $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_2}$. Also show that the variables Y_1 and Y_2 are independent.

Let $g(x_1,x_2) = \left(x_1 + x_2, \frac{x_1}{x_2}\right) = (y_1,y_2)$. Let us calculate the domain of the transformation.

Remember that the p.d.f of the exponential distribution is:

$$f(x) = \lambda e^{-\lambda x} \mathbf{1}_{(0,\infty)}(x),$$

thus $\mathcal{A} = (0,\infty) \times (0,\infty)$. Since $x_1, x_2 > 0$ we get that $x_1 + x_2 > 0$ and $\frac{x_1}{x_2} > 0$, and so $\mathcal{B} = (0,\infty)^2$ as well. The function g restricted to this sets is bijective as we can easily show by solving the equations: $y_1 = x_1 + x_2$ and $y_2 = \frac{x_1}{x_2}$. We obtain:

$$\begin{aligned} x_1 = x_2y_2 &\Rightarrow y_1 = x_2y_2 + x_2 \\ &\Rightarrow x_2 = \frac{y_1}{1+y_2} \\ &\Rightarrow x_1 = \frac{y_1y_2}{1+y_2} \end{aligned}$$

Since the solution is unique the function g is one-to-one. Since the solution exists for all $(y_1,y_2) \in (0.\infty)^2$ the function is surjective. Its inverse is precisely:

$$g^{-1}(y_1,y_2) = \left(\frac{y_1y_2}{1+y_2}, \frac{y_1}{1+y_2}\right)$$

Furthermore, the Jacobian is:

$$J_{g^{-1}}(y_1,y_2) = \begin{vmatrix} \frac{y_2}{1+y_2} & \frac{1}{1+y_2} \\ \frac{y_1}{(1+y_2)^2} & -\frac{y_1}{(1+y_2)^2} \end{vmatrix} = -\frac{y_1y_2}{(1+y_2)^3} - \frac{y_1}{(1+y_2)^3} = -\frac{y_1}{(1+y_2)^2}$$

Thus the desired p.d.f is:

$$\begin{aligned} f_Y(y_1,y_2) &= f\left(\frac{y_1y_2}{1+y_2},\frac{y_1}{1+y_2}\right)\left|-\frac{y_1}{(1+y_2)^2}\right|\mathbf{1}_{(y_1,y_2)\in(0,\infty)^2} \\ &= \lambda e^{-\lambda\frac{y_1y_2}{1+y_2}}\lambda e^{-\frac{y_1}{1+y_2}}\frac{y_1}{(1+y_2)^2}\mathbf{1}_{\{y_1,y_2>0\}} \\ &= \lambda^2 e^{-\lambda y_1}\frac{y_1}{(1+y_2)^2}\mathbf{1}_{\{y_1,y_2>0\}} \end{aligned}$$

Finally, to end the example it is enough to recognize that the p.d.f. of Y can be decomposed into a product of two functions, one of them only in the variable y_1 and the other only a function of the variable y_2. Thus, if we apply the next lemma the example is solved. □

Lemma 2.2. *If the joint distribution f of a random vector (X,Y) factors as a product of functions of only x and y, i.e., there exist $g,h:\mathbb{R}\to\mathbb{R}$ such that $f(x,y)=g(x)h(y)$ then the variables X,Y are independent.*

Proof. Problem 2.12.

Example 2.9. Let X, Y be two random variables with joint p.d.f. $f(\cdot,\cdot)$. Calculate the density of $X+Y$.

Let $(U,V)=(X+Y,Y)$. We can easily calculate the domain and the inverse $g^{-1}(u,v)=(u-v,v)$. The Jacobian is:

$$J_{g^{-1}}(u,v)=\begin{vmatrix}1 & -1\\ 0 & 1\end{vmatrix}=1$$

As a result the desired p.d.f. is:

$$f_U(u)=\int_{-\infty}^{\infty}f(u-v,v)dv$$

We will observe this particular example later when we talk about convolutions.

Example 2.10. Let X_1 and X_2 be i.i.d. $N(0,1)$ random variables. Consider the function $g(x_1,x_2)=\left(\frac{x_1}{x_2},|x_2|\right)$. Calculate the joint distribution of $Y=g(X)$ and the distribution of the ratio of the two normals: X_1/X_2.

First, $\mathcal{A}=R^2$ and $\mathcal{B}=\mathbb{R}\times(0,\infty)$. Second, note that the transformation is not one-to-one. Also note that we have a problem when $x_2=0$[7]. Fortunately, we know

[7] 0 is in $\mathcal{A}$ since $f_{X_2}(0)>0$

how to deal with this situation. Take a partition of $\mathcal{A}$ as follows:

$$\mathcal{A}_0 = \{(x_1,0) : x_1 \in \mathbb{R}\},\ \mathcal{A}_1 = \{(x_1,x_2) : x_2 < 0\},\ \mathcal{A}_1 = \{(x_1,x_2) : x_2 > 0\}.$$

$\mathcal{A}_0$ has the desired property since $\mathbf{P}((X_1,X_2) \in \mathcal{A}_0) = \mathbf{P}(X_2 = 0) = 0$ (X_2 is a continuous random variable). Restricted to each A_i the function g is bijective and we can calculate its inverse in both cases:

$$g_1^{-1}(y_1,y_2) = (-y_1y_2, -y_2)$$
$$g_2^{-1}(y_1,y_2) = (y_1y_2, y_2)$$

In either case the Jacobian is identical $J_1 = J_2 = y_2$. Using the p.d.f. of a normal with mean zero and variance 1 ($f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$), and that X_1 and X_2 being independent the joint p.d.f. is the product of marginals we obtain:

$$\begin{aligned} f_Y(y_1,y_2) &= \left(\frac{1}{2\pi} e^{-(-y_1y_2)^2/2} e^{-(-y_2)^2/2} |y_2| + \frac{1}{2\pi} e^{-(y_1y_2)^2/2} e^{-(y_2)^2/2} |y_2| \right) \mathbf{1}_{\{y_2>0\}} \\ &= \frac{y_2}{\pi} e^{-\frac{(y_1^2+1)y_2^2}{2}} \mathbf{1}_{\{y_2>0\}}, \quad y_1 \in \mathbb{R}, \end{aligned}$$

and this is the desired joint distribution. To calculate the distribution of X_1/X_2 we calculate the marginal of Y_1 by integrating out y_2:

$$\begin{aligned} f_{Y_1}(y_1) &= \int_0^\infty \frac{y_2}{\pi} e^{-\frac{(y_1^2+1)y_2^2}{2}} dy_2 \quad (\text{ Change of variables } y_2^2 = t) \\ &= \int_0^\infty \frac{1}{2\pi} e^{-\frac{(y_1^2+1)}{2} t} dt = \frac{1}{2\pi} \frac{2}{y_1^2+1} \\ &= \frac{1}{\pi(y_1^2+1)}, \quad y_1 \in \mathbb{R} \end{aligned}$$

But this is the distribution of a Cauchy random variable. Thus we have just proven that the ratio of two independent $N(0,1)$ rv's has a Cauchy distribution. □

We conclude this chapter with a non-trivial application of the Borel-Cantelli lemmas. We have postponed this example until this point since we needed to learn about independent random variables first.

Example 2.11. Let $\{X_n\}$ a sequence of i.i.d. random variables, each exponentially distributed with rate 1, i.e.:

$$\mathbf{P}(X_n > x) = e^{-x}, \quad x > 0.$$

We wish to study how large are these variables when $n \to \infty$. To this end take $x = \alpha \log n$, for some $\alpha > 0$ and for any $n \geq 1$. Substitute into the probability above to obtain:

$$\mathbf{P}(X_n > \alpha \log n) = e^{-\alpha \log n} = n^{-\alpha} = \frac{1}{n^\alpha}.$$

But we know that the sum $\sum_n \frac{1}{n^\alpha}$ is divergent for the exponent $\alpha \leq 1$ and convergent for $\alpha > 1$. So we can apply the Borel-Cantelli lemmas since the events in question are independent. Thus,

If $\alpha \leq 1$ the sum is divergent and so $\sum_n \mathbf{P}(X_n > \alpha \log n) = \infty$, thus:

$$\mathbf{P}\left(\frac{X_n}{\log n} > \alpha \text{ i.o.}\right) = 1$$

If $\alpha > 1$ the sum is convergent, and $\sum_n \mathbf{P}(X_n > \alpha \log n) < \infty$, thus:

$$\mathbf{P}\left(\frac{X_n}{\log n} > \alpha \text{ i.o.}\right) = 0$$

We can express the same thing in terms of lim sup like so:

$$\mathbf{P}\left(\limsup_n \frac{X_n}{\log n} > \alpha\right) = \begin{cases} 0 & \text{, if } \alpha > 1 \\ 1 & \text{, if } \alpha \leq 1 \end{cases}$$

Since for all $\alpha \leq 1$ we have that $\mathbf{P}\left(\limsup_n \frac{X_n}{\log n} > \alpha\right) = 1$, then we necessarily have:

$$\mathbf{P}\left(\limsup_n \frac{X_n}{\log n} \geq 1\right) = 1$$

Take $\alpha = 1 + \frac{1}{k}$ and look at the other implication: $\mathbf{P}\left(\limsup_n \frac{X_n}{\log n} > 1 + \frac{1}{k}\right) = 0$, and this happens for all $k \in \mathbb{N}$, . But we can write:

$$\left\{\limsup_n \frac{X_n}{\log n} > 1\right\} = \bigcup_{k \in \mathbb{N}} \left\{\limsup_n \frac{X_n}{\log n} > 1 + \frac{1}{k}\right\},$$

and since any countable union of null sets is itself a null set, the probability of the event on the left must be zero. Therefore, $\limsup_n \frac{X_n}{\log n} \leq 1$ a.s. and combining with the finding above:

$$\limsup_n \frac{X_n}{\log n} = 1, \quad a.s.$$

This is very interesting since as we will see in the chapter dedicated to the Poisson process, these X_n are the inter-arrival times of this process. The example above tells us that if we look at the realizations of such a process then they form a sequence of numbers that has the upper limiting point equal to 1, or put differently there is no subsequence of inter-arrival times that in the limit is greater than the $\log n$.

Problems

2.1. Prove the Proposition 2.1. That is prove that the function F in Definition 2.4 is increasing, right continuous and taking values in the interval $[0,1]$, using only proposition 1.3 on page 13.

2.2. Show that any piecewise constant function is Borel measurable. (see description of piecewise constant functions in Definition 2.6

2.3. Give an example of two distinct random variables with the same distribution function.

2.4. Buffon's needle problem.
Suppose that a needle is tossed at random onto a plane ruled with parallel lines a distance L apart, where by a "needle" we mean a line segment of length $l \leq L$.
What is the probability of the needle intersecting one of the parallel lines?

Hint: Consider the angle that is made by the needle with the parallel lines as a random variable α uniformly distributed in the interval $[0,2\pi]$ and the position of the midpoint of the needle as another random variable ξ also uniform on the interval $[0,L]$. Then express the condition "needle intersects the parallel lines" in terms of the position of the midpoint of the needle and the angle α. Do a calculation similar with example 2.5.

2.5. A random variable X has distribution function

$$F(x) = a + b \arctan \frac{x}{2} \quad , -\infty < x < \infty$$

Find:
a) The constants a and b
b) The probability density function of X

2.6. What is the probability that two randomly chosen numbers between 0 and 1 will have a sum no greater than 1 and a product no greater than $\frac{15}{64}$?

2.7. We know that the random variables X and Y have joint density $f(x,y)$. Assume that $\mathbf{P}(Y = 0) = 0$. Find the densities of the following variables:
a) $X + Y$
b) $X - Y$
c) XY
d) $\frac{X}{Y}$

2.8. Choose a point A at random in the interval $[0,1]$. Let L_1 (respectively L_2) be the length of the bigger (respectively smaller) segment determined by A on $[0,1]$. Calculate:
a) $\mathbf{P}(L_1 \leq x)$ for $x \in \mathbb{R}$.
b) $\mathbf{P}(L_2 \leq x)$ for $x \in \mathbb{R}$.

2.9. Two friends decide to meet at the Castle gate of Stevens Institute. They each arrive at that spot at some random time between a and $a+T$. They each wait for 15 minutes then leave if the other did not appear. What is the probability that they meet?

2.10. Let $X_1, X_2, \ldots, X_n$ be independent $U(0,1)$ random variables. Let $M = \max_{1 \le i \le n} X_i$. Calculate the distribution function of M.

2.11. The random variable whose probability density function is given by:

$$f(x) = \begin{cases} \frac{1}{2}\lambda e^{\lambda x} & , \quad \text{if } x \le 0 \\ \frac{1}{2}\lambda e^{-\lambda x} & , \quad \text{if } x > 0, \end{cases}$$

is said to have a Laplace, sometimes called a *double exponential*, distribution.
a) Verify that the density above defines a proper probability distribution.
b) Find the distribution function $F(x)$ for a Laplace random variable.

Now, let X and Y be independent exponential random variables with parameter λ. Let I be independent of X and Y and equally likely to be 1 or -1.

c) Show that $X - Y$ is a Laplace random variable.
d) Show that IX is a Laplace random variable.
e) Show that W is a Laplace random variable where:

$$W = \begin{cases} X & , \quad \text{if } I = 1 \\ -Y & , \quad \text{if } I = -1. \end{cases}$$

2.12. Give a proof of the lemma 2.2 on page 53.

Chapter 3
Integration Theory

In the previous chapter we learned about random variables and their distributions. This distribution completely characterizes a random variable. But in general distributions are very complex functions. The human brain cannot comprehend such things easily. So the human brain wants to talk about one typical value. For example, one can give a distribution for the random variable representing player salaries in the NBA. Here the variability (probability space) is represented by the specific player chosen. However, probably one is not interested in such a distribution. One simply wants to know what is the typical salary in the NBA. The person probably contemplates a career in sports and wants to find out if as an athlete should go for basketball or baseball, therefore he is much better serve by comparing only two numbers. Calculating such a number is hard (which number?). In this chapter we create a theory to calculate any numbers that the person wishes. Paradoxically, to calculate a simple number we need to understand a very complex theory.

3.1 Integral of measurable functions

Recall that the random variables are nothing more than measurable functions. Let $(\Omega, \mathscr{F}, P)$ be a probability space. We wish to define for any measurable function f an integral of f with respect to the measure P.

Notation. We shall use the following notations for this integral:

$$\int_\Omega f(\omega)\mathbf{P}(d\omega) = \int f d\mathbf{P}$$

$$\text{for } A \in \mathscr{F} \text{ we have } \int_A f(\omega)\mathbf{P}(d\omega) = \int_A f d\mathbf{P} = \int f \mathbf{1}_\mathbf{A} d\mathbf{P}$$

Recall the Dirac Delta we have defined previously? With its help summation is another kind of integral. Let $\{a_n\}$ be a sequence of real numbers. Let $\Omega = \mathbb{R}, \mathscr{F} = \mathscr{B}(\mathbb{R})$ and the measure on this set is $\delta(A) = \sum_{i=1}^{\infty} \delta_i(A)$.

Then the function $i \mapsto a_i$ is integrable if and only if $\sum a_i < \infty$ and in this case we have:

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \int_{-\infty}^{\infty} a_x d\delta_n(x) = \int_{-\infty}^{\infty} a_x \sum_{n=1}^{\infty} d\delta_n(x) = \int_{-\infty}^{\infty} a_x d\delta(x)$$

What is the point of this? The simple argument above shows that any "discrete" random variable (in the undergraduate text definition) may be treated as a "continuous" random variable. Not that there was any doubt after all the big fuss we made about it in the previous chapter.

Integral of Simple (Elementary) Functions

If $A \in \mathscr{F}$ we know that we can define a measurable function by its indicator $\mathbf{1}_A$. We define the integral of this measurable function $\int \mathbf{1}_A d\mathbf{P} = \mathbf{P}(A)$. We note that this variable has the same distribution as that of the Bernoulli random variable. The variable takes values 0 and 1 and we can easily calculate the probability that the variable is 1 as:

$$\mathbf{P} \circ \mathbf{1}_A^{-1}(\{1\}) = \mathbf{P}\{\omega : \mathbf{1}_A(\omega) = 1\} = \mathbf{P}(A).$$

Therefore the variable is distributed as a Bernoulli random variable with parameter $p = \mathbf{P}(A)$.

Definition 3.1 (Simple function). f is called a *simple* (elementary) function if and only if f can be written as a finite linear combination of indicators or, more specifically there exist sets $A_1, A_2, \ldots, A_n$ all in $\mathscr{F}$ and constants $a_1, a_2, \ldots, a_n$ in $\mathbb{R}$ such that:

$$f(\omega) = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}(\omega)$$

If the constants a_k are all positive, then f is a positive simple function.

Note that the sets A_i do not have to be disjoint but an easy exercise (Problem 3.1) shows that f could be written in terms of disjoint sets.

For any simple function f we define its integral:

$$\int f d\mathbf{P} = \sum_{k=1}^{n} a_k \mathbf{P}(A_k) < \infty$$

We adopt the conventions $0 * \infty = 0$ and $\infty * 0 = 0$ in the above summation.

We need to check that the above definition is proper. For there exist many representations of a simple function and we need to make sure that any such representation produces the same integral value. Furthermore, the linearity and monotonicity properties of the integral may be proven. We skip these results since they are simple to prove and do not bring any additional insight.

Integral of positive measurable functions

For every f positive measurable function $f : \Omega \longrightarrow [0,\infty)$ we define:

$$\int f d\mathbf{P} = \sup\left\{ \int h d\mathbf{P} : h \text{ is a simple function, } h \leq f \right\}$$

For a given positive measurable function can we find a sequence of simple functions that converge to it? The answer is yes and is provided by the next simple exercise:

Exercise 3.1. Let $f : \Omega \to [0,\infty]$ be a positive, measurable function. For all $n \geq 1$, we define:

$$f_n(\omega) := \sum_{k=0}^{n2^n-1} \frac{k}{2^n} \mathbf{1}_{\left\{\frac{k}{2^n} \leq f(\omega) < \frac{k+1}{2^n}\right\}}(\omega) + n\mathbf{1}_{\{f(\omega)\geq n\}} \tag{3.1}$$

1. Show that f_n is a simple function on $(\Omega, \mathscr{F})$, for all $n \geq 1$.
2. Show that the sets present in the indicators in equation (3.1) form a partition of Ω, for all $n \geq 1$.
3. Show that the sequence of simple functions is increasing $g_n \leq g_{n+1} \leq f$, for all $n \geq 1$.
4. Show that $g_n \uparrow f$ as $n \to \infty$. Note that this is not an a.s. statement, it is true for all $\omega \in \Omega$.

The solution to this exercise is not complicated and in fact it is an assigned problem (Problem 3.3).

The following lemma is a very easy to understand and useful tool.

Lemma 3.1. *If f is a positive measurable function and $\int f d\mathbf{P} = 0$ then $\mathbf{P}\{f > 0\} = 0$ (or $f = 0$ a.s.).*

Proof. We have $\{f > 0\} = \bigcup_{n\geq 0}\{f > \frac{1}{n}\}$. Since the events are increasing by the monotone convergence property of measure we must have $\mathbf{P}\{f > 0\} = \lim_{n\to\infty} \mathbf{P}\{f > \frac{1}{n}\}$. If we assume by absurd that $\mathbf{P}\{f > 0\} > 0$ then there must exist an n such that $\mathbf{P}\{f > \frac{1}{n}\} > 0$. However, in this case by the definition of the integral of positive measurable functions:

$$\int f d\mathbf{P} \geq \int \frac{1}{n} \mathbf{1}_{\{f>\frac{1}{n}\}} d\mathbf{P} > 0,$$

contradiction. □

The next theorem is one of the most useful in probability theory. In our immediate context it tells us that the integral for positive measurable functions is well defined.

Theorem 3.1 (Monotone Convergence Theorem). *If f is a sequence of measurable positive functions such that $f_n \uparrow f$ then:*

$$\int_\Omega f_n(\omega)\mathbf{P}(d\omega) \uparrow \int_\Omega f(\omega)\mathbf{P}(d\omega)$$

Note: This is all there is to integration theory. The proof of the monotone convergence theorem is not difficult, you may want to look at it.

Proof. **Ion: Write the proof**

Integral of measurable functions

Let f be any measurable function. Then we write $f = f^+ - f^-$ where:

$$f^+(s) = \max\{f(s), 0\}$$
$$f^-(s) = \max\{-f(s), 0\}$$

Then f^+ and f^- are positive measurable functions and $|f| = f^+ + f^-$. Since they are positive measurable their integrals are well defined by the previous part.

Definition 3.2. We define $L^1(\Omega, \mathscr{F}, P)$ as being the space of all functions f such that:

$$\int |f| d\mathbf{P} = \int f^+ d\mathbf{P} + \int f^- d\mathbf{P} < \infty$$

For any f in this space which we will shorten to $L^1(\Omega)$ or even simpler to L^1 we define:

$$\int f d\mathbf{P} = \int f^+ d\mathbf{P} - \int f^- d\mathbf{P}$$

Note: With the above it is trivial to show that $|\int f d\mathbf{P}| \leq \int |f| d\mathbf{P}$

Linearity:

If $f, g \in L^1(\Omega)$ with $a, b \in \mathbb{R}$, then:

$$af + bg \in L^1(\Omega)$$
$$\int (af + bg) d\mathbf{P} = a \int f d\mathbf{P} + b \int g d\mathbf{P}$$

Lemma 3.2 (Fatou's Lemma for measurable functions). *If one of the following is true:*

a) $\{f_n\}_n$ *is a sequence of positive measurable functions or*
b) $\{f_n\} \subset L^1(\Omega)$

then:

$$\int \liminf_n f_n d\mathbf{P} \leq \liminf_n \int f_n d\mathbf{P}$$

Proof. Note that $\liminf_n f_n = \lim_{m \to \infty} \inf_{n \geq m} f_n$, where $\lim_{m \to \infty} \inf_{n \geq m} f_n$ is an increasing sequence.

Let $g_m = \inf_{n \geq m} f_n$, and $n \geq m$:

$$f_n \geq \inf_{n\geq m} f_m = g_m \Rightarrow \int f_n d\mathbf{P} \geq \int g d\mathbf{P} \Rightarrow \int g_m d\mathbf{P} \leq \inf_{n\geq m} \int f_n d\mathbf{P}$$

Now g_m increases so we may use the Monotone Convergence Theorem and we get:

$$\int \lim_{m\to\infty} g_m d\mathbf{P} = \lim_{m\to\infty} \int g_m d\mathbf{P} \leq \lim_{m\to\infty} \inf_{n\geq m} \int f_n d\mathbf{P} = \liminf_n \int f_n d\mathbf{P}$$

Theorem 3.2 (Dominated Convergence Theorem). *If f_n, f are measurable, $f_n(\omega) \to f(\omega)$ for all $\omega \in \Omega$ and the sequence f_n is dominated by $g \in L^1(\Omega)$:*

$$|f_n(\omega)| \leq g(\omega), \qquad \forall \omega \in \Omega, \forall n \in \mathbb{N}$$

then:

$$f_n \to f \text{ in } L^1(\Omega) \qquad \left(\text{i.e. } \int |f_n - f| d\mathbf{P} \to 0 \right)$$

Thus $\int f_n d\mathbf{P} \to \int f d\mathbf{P}$ and $f \in L^1(\Omega)$.

The Standard Argument:

This argument is the most important argument in the probability theory. Suppose that we want to prove that some property holds for all functions h in some space such as $L^1(\Omega)$ or the space of measurable functions.

1. Show that the result is true for all indicator functions.
2. Use linearity to show the result holds true for all f simple functions.
3. Use the Monotone Convergence Theorem to obtain the result for measurable positive functions.
4. Finally from the previous step and writing $f = f^+ - f^-$ we show that the result is true for all measurable functions.

3.2 Expectations

Since a random variable is just a measurable function we just need to particularize the results of the previous section. An integral with respect to a probability measure is called an expectation. Let $(\Omega, \mathscr{F}, P)$ be a probability space.

Definition 3.3. For X a r.v. in $L^1(\Omega)$ define:

$$\mathbf{E}(X) = \int_\Omega X d\mathbf{P} = \int_\Omega X(\omega) d\mathbf{P}(\omega) = \int_\Omega X(\omega) \mathbf{P}(d\omega)$$

This expectation has the same properties of the integral defined before and some extra ones since the space has finite measure.

Convergence Theorems:

(i) *Monotone Convergence Theorem:* If $X_n \geq 0$, $X_n \in L^1$ and $X_n \uparrow X$ then $\mathbf{E}(X_n) \uparrow \mathbf{E}(X) \leq \infty$.
(ii) *Fatou:* $\mathbf{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbf{E}(X_n)$
(iii) *Dominated Convergence Theorem:* If $|X_n(\omega)| \leq Y(\omega)$ on Ω with $Y \in L^1(\Omega)$ and $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ then $\mathbf{E}(|X_n - X|) \to 0$.

Now let us present specific properties of the expectation. This is to be expected since the space has finite measure therefore we can obtain more specific properties.

Markov Inequality:

Let Z be a r.v. and let $g : \mathbb{R} \longrightarrow [0,\infty]$ be an *increasing* measurable function. Then:

$$\mathbf{E}[g(Z)] \geq \mathbf{E}\left[g(Z)\mathbf{1}_{\{Z\geq c\}}\right] \geq g(c)\mathbf{P}(Z \geq c)$$

Thus

$$\mathbf{P}(Z \geq c) \leq \frac{\mathbf{E}[g(Z)]}{g(c)}$$

for all g increasing functions and $c > 0$.

Example 3.1 (Special cases of the Markov inequality). If we take $g(x) = x$ an increasing function and X a positive random variable then we obtain:

$$\mathbf{P}(Z \geq c) \leq \frac{\mathbf{E}(Z)}{c}.$$

To get rid of the necessity that $X \geq 0$ take $Z = |X|$. Then we obtain the classical form of the Markov inequality:

$$\mathbf{P}(|X| \geq c) \leq \frac{\mathbf{E}(|X|)}{c}.$$

If we take $g(x) = x^2$, $Z = |X - \mathbf{E}(X)|$ and we use the variance definition (which we will see in a minute), we obtain the Chebyshev inequality:

$$\mathbf{P}(|X - \mathbf{E}(X)| \geq c) \leq \frac{Var(X)}{c^2}.$$

If we denote $\mathbf{E}(X) = \mu$ and $Var(X) = \sigma$ and we take $c = k\sigma$ in the previous inequality we will obtain the classical Chebyshev inequality presented in undergraduate courses:

$$\mathbf{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

If $g(x) = e^{\theta x}$, with $\theta > 0$ then

$$\mathbf{P}(Z \geq c) \leq e^{-\theta c}\mathbf{E}(e^{\theta z}),$$

This inequality states that the tail of the distribution decays exponentially in c if Z has finite exponential moments. With simple manipulations one can obtain Chernoff's inequality using it.

Jensen's Inequality for convex functions:

This is just a reminder.

Definition 3.4. A function $g : I \longrightarrow \mathbb{R}$ is called a convex function on I (where I is any open interval in $\mathbb{R}$, if its graph lies below any of its chords. Mathematically: for any $x, y \in I$ and for any $\alpha \in (0,1)$ we have

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

Some examples of convex functions on the whole $\mathbb{R}$: $|x|$, x^2 and $e^{\theta x}$, with $\theta > 0$.

Lemma 3.3 (Jensen's Inequality). *Let f be a convex function and let X be a r.v. in $L^1(\Omega)$. Assume that $\mathbf{E}(f(X)) \leq \infty$ then:*

$$f(\mathbf{E}(X)) \leq \mathbf{E}(f(X))$$

Proof. Skipped. The classical approach indicators $\rightarrow$ simple functions $\rightarrow$ positive measurable $\rightarrow$ measurable is a standard way to prove Jensen.

L^p spaces.

We generalize the L^1 notion presented earlier in the following way. For $1 \leq p \leq \infty$ we define the space:

$$L^p(\Omega, \mathscr{F}, P) = L^p(\Omega) = \left\{ X : \Omega \longrightarrow \mathbb{R} : \mathbf{E}\left[|X|^p\right] = \int |X|^p d\mathbf{P} < \infty \right\},$$

On this space we define a norm called the p-norm as:

$$||X||_p = \mathbf{E}\left[|X|^p\right]^{1/p}$$

Lemma 3.4 (Properties of L^p spaces).

(i) L^p is a vector space. (i.e., if $X, Y \in L^p$ and $a, b \in \mathbb{R}$ then $aX + bY \in L^p$).
(ii) L^p is complete (every Cauchy sequence in L^p is convergent)

Lemma 3.5 (Cauchy-Bunyakovsky-Schwarz inequality). *If $X, Y \in L^2(\Omega)$ then $X, Y \in L^1(\Omega)$ and*

$$|\mathbf{E}[XY]| \leq \mathbf{E}[|XY|] \leq ||X||_2 ||Y||_2$$

A historical remark. This inequality, one of the most famous and useful un any area of analysis (not only probability) is usually credited to Cauchy for sums and Schwartz for integrals and is usually known as the Cauchy-Schwartz inequality. However,the Russian mathematician Victor Yakovlevich Bunyakovsky (1804-1889) discovered and first published the inequality for integrals in 1859 (when Schwartz was 16). Unfortunately, he was born in eastern Europe... However, all who are born in eastern Europe (including myself) learn the inequality by its proper name.

Proof. The first inequality is clear by Jensen inequality. We need to show

$$\mathbf{E}[|XY|] \leq (\mathbf{E}[X^2])^{1/2}(\mathbf{E}[Y^2])^{1/2}$$

Let $W = |X|$ and $Z = |Y|$ then $W, Z \geq 0$.
Truncation:
Let $W_n = W \bigwedge n$ and $Z_n = Z \bigwedge n$ that is

$$W_n(\omega) = \begin{cases} W(\omega), & \text{if } W(\omega) < n \\ n, & \text{if } W(\omega) \geq n \end{cases}$$

Clearly, defined in this way W_n, Z_n are bounded. Let $a, b \in \mathbb{R}$ two constants. Then:

$$0 \leq \mathbf{E}[(aW_n + bZ_n)^2] = a^2\mathbf{E}(W_n^2) + 2ab\mathbf{E}(W_nZ_n) + b^2\mathbf{E}(Z_n^2)$$

If we let $a/b = c$ we get:

$$c^2\mathbf{E}(W_n^2) + 2c\mathbf{E}(W_nZ_n) + \mathbf{E}(Z_n^2) \geq 0 \quad \forall c \in \mathbb{R}$$

This means that the quadratic function in c has to be positive. But this is only possible if the determinant of the equation is negative and the leading coefficient $\mathbf{E}(W_n^2)$ is strictly positive, the later condition is obviously true. Thus we must have:

$$\begin{aligned} &4(\mathbf{E}(W_nZ_n))^2 - 4\mathbf{E}(W_n^2)\mathbf{E}(Z_n^2) \leq 0 \\ &\Rightarrow (\mathbf{E}(W_nZ_n))^2 \leq \mathbf{E}(W_n^2)\mathbf{E}(Z_n^2) \leq \mathbf{E}(W^2)\mathbf{E}(Z^2) \qquad \forall n \end{aligned}$$

If we let $n \uparrow \infty$ and use the monotone convergence theorem we get:

$$(\mathbf{E}(WZ))^2 \leq \mathbf{E}(W^2)\mathbf{E}(Z^2).$$

□

A more general inequality is:

Lemma 3.6 (Hölder inequality). *If* $1/p + 1/q = 1$, $X \in L^p(\Omega)$ *and* $Y \in L^q(\Omega)$ *then* $XY \in L^1(\Omega)$ *and:*

$$\mathbf{E}|XY| \leq \|X\|_p\|Y\|_q = (\mathbf{E}|X|^p)^{\frac{1}{p}}(\mathbf{E}|Y|^q)^{\frac{1}{q}}$$

Proof. The proof is simple and uses the following inequality (Young inequality): if a and b are positive real numbers and p, q are as in the theorem then:

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q},$$

with equality if and only if $a^p = b^q$.

Taking this inequality as given (not hard to prove) define:

$$f = \frac{|X|}{\|X\|_p}, \quad g = \frac{|Y|}{\|Y\|_p}.$$

Note that the Hölder inequality is equivalent with $\mathbf{E}[fg] \leq 1$ ($\|X\|_p$ and $\|Y\|_q$ are just numbers that can be taken in and out of integral by the linearity property). To prove this apply the Young inequality to $f \geq 0$ and $g \geq 0$ and then integrate to obtain:

$$\mathbf{E}[fg] \leq \frac{1}{p}\mathbf{E}[f^p] + \frac{1}{q}\mathbf{E}[g^q] = \frac{1}{p} + \frac{1}{q} = 1$$

$\mathbf{E}[f^p] = 1$ and similarly for g may be easily checked. Finally, the extreme cases ($p = 1$, $q = \infty$, etc.) may be treated separately. □

Lemma 3.7 (Minkowski Inequality). *If $X, Y \in L^p$ then $X + Y \in L^p$ and:*

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$$

Proof. We clearly have:

$$|X + Y|^p \leq 2^{p-1}(|X|^p + |Y|^p).$$

For example use the definition of convexity for the function x^p with $x = |X|$ and $y = |Y|$ and $\alpha = 1/2$. Now integrating implies that $X + Y \in L^p$. Now we can write:

$$\begin{aligned}
\|X+Y\|_p^p &= \mathbf{E}[|X+Y|^p] \leq \mathbf{E}\left[(|X|+|Y|)|X+Y|^{p-1}\right] \\
&= \mathbf{E}\left[|X||X+Y|^{p-1}\right] + \mathbf{E}\left[|Y||X+Y|^{p-1}\right] \\
&\overset{\text{Hölder}}{\leq} (\mathbf{E}[|X|^p])^{1/p}\left(\mathbf{E}\left[|X+Y|^{(p-1)q}\right]\right)^{1/q} + (\mathbf{E}[|Y|^p])^{1/p}\left(\mathbf{E}\left[|X+Y|^{(p-1)q}\right]\right)^{1/q} \\
&\overset{\left(q=\frac{p}{p-1}\right)}{=} (\|X\|_p + \|Y\|_p)(\mathbf{E}[|X+Y|^p])^{1-\frac{1}{p}} \\
&= (\|X\|_p + \|Y\|_p)\frac{\mathbf{E}[|X+Y|^p]}{\|X+Y\|_p}
\end{aligned}$$

Now, identifying the left and right hand after simplifications we obtain the result. □

Example 3.2 (due to Erdós). Suppose there are 17 fence posts around the perimeter of a field and exactly 5 of them are rotten. Show that irrespective of which of these

5 are rotten, there should exist a row of 7 consecutive posts of which at least 3 are rotten.

Proof (Solution). First we label the posts $1,2\cdots 17$. Now define :

$$I_k = \begin{cases} 1 & \text{if post } k \text{ is rotten} \\ 0 & \text{otherwise} \end{cases}$$

For any fixed k, let R_k denote the number of rotten posts among $k+1,\cdots,k+7$ (starting with the next one). Note that when any of $k+1,\cdots,k+7$ are larger than 17 we start again from 1 (i.e., modulo 17 +1).

Now pick a post at random this obviously can be done in 17 ways with equal probability. Then after we pick this post we calculate the number of rotten boards. We have:

$$\begin{aligned} \mathbf{E}(R_k) &= \sum_{k=1}^{17}(I_{k+1}+\cdots+I_{k+7})\frac{1}{17} \\ &= \frac{1}{17}\sum_{k=1}^{17}\sum_{j=1}^{7} I_{k+j} == \frac{1}{17}\sum_{j=1}^{7}\sum_{k=1}^{17} I_{j+k} \\ &= \frac{1}{17}\sum_{j=1}^{7} 5 \qquad \text{(the sum is 5 since we count all the rotten posts in the fence)} \\ &= \frac{35}{17} \end{aligned}$$

Now, $35/17 > 2$ which implies $\mathbf{E}(R_k) > 2$. Therefore, $\mathbf{P}(R_k > 2) > 0$ (otherwise the expectation is necessarily bounded by 2) and since R_k is integer valued $\mathbf{P}(R_k \geq 3) > 0$. So there exists some k such that $R_k \geq 3$.

Of course now that we see the proof we can play around with numbers and see that there exists a row of 4 consecutive posts in which at least two are rotten, or that there must exist a row of 11 consecutive posts in which at least 4 are rotten and so on (row of 14 containing all 5 rotten ones).

3.3 Variance and the correlation coefficient

Definition 3.5. The variance or the Dispersion of a random variable $X \in L^2(\Omega)$ is:

$$V(X) = \mathbf{E}[(X-\mu)^2] = \mathbf{E}(X^2) - \mu^2$$

Where $\mu = \mathbf{E}(X)$.

Definition 3.6. Given two random variables X, Y we call the covariance between X and Y the quantity:

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

Where $\mu_X = \mathbf{E}(X)$ and $\mu_Y = \mathbf{E}(Y)$.

Definition 3.7. Given random variables X, Y we call the correlation coefficient:

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\mathbf{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathbf{E}[(X-\mu_X)^2]\mathbf{E}[(Y-\mu_Y)^2]}}$$

From the Cauchy-Schwartz inequality applied to $X-\mu_X$ and $Y-\mu_Y$ we get $|\rho| < 1$ or $\rho \in [-1,1]$.

The variable X and Y are called **uncorrelated** if the covariance (or equivalently the correlation) between them is zero.

Proposition 3.1 (Properties of expectation). *The following are true:*

(i) If X and Y are integrable r.v.'s then for any constants α and β the r.v. $\alpha X + \beta Y$ is integrable and $\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}X + \beta \mathbf{E}Y$.
(ii) $V(aX+bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$
(iii) If X, Y are independent then $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$ and $Cov(X,Y) = 0$.
(iv) If $X(\omega) = c$ with probability 1 and $c \in \mathbb{R}$ a constant then $\mathbf{E}X = c$.
(v) If $X \geq Y$ a.s. then $\mathbf{E}X \geq \mathbf{E}Y$. Furthermore, if $X \geq Y$ a.s. and $\mathbf{E}X = \mathbf{E}Y$ then $X = Y$ a.s.

Proof. Exercise. Please note that the reverse of the part (iii) above is not true, if the two variables are uncorrelated this does not mean that they are independent. In fact in Problem 3.5 you are required to provide a counterexample.

3.4 Functions of random variables. The Transport Formula.

In Section 2.4 on page 49 we showed how to calculate distributions and in particular p.d.f.'s for continuous random variables. We have also promised a more general result. Well, here it is. This general result allows to construct random variables and in particular distributions in any space. This is the result that allows us to claim that studying random variables on $([0,1], \mathscr{B}([0,1]), \lambda)$ is enough. We had to postpone presenting the result until this point since we had to learn first how to integrate.

Theorem 3.3 (General Transport Formula). *Let $(\Omega, \mathbb{R}, P)$ be a probability space. Let f be a measurable function such that:*

$$(\Omega, \mathscr{F}) \xrightarrow{f} (S, \mathscr{G}) \xrightarrow{\varphi} (\mathbb{R}, \mathscr{B}(\mathbb{R})),$$

where $(S, \mathscr{G})$ is a measurable space. Assuming that at least one of the integrals exists we then have:

$$\int_\Omega \varphi \circ f d\mathbf{P} = \int_S \varphi d\mathbf{P} \circ f^{-1},$$

for all φ measurable functions.

Proof. We will use the standard argument technique discussed above.

1. Let φ be the indicator function. $\varphi = \mathbf{1}_A$ for $A \in \mathscr{G}$:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Then we get:

$$\begin{aligned} \int_\Omega \mathbf{1}_\mathbf{A} \circ f d\mathbf{P} &= \int_\Omega \mathbf{1}_A(f(\omega)) d\mathbf{P}(\omega) = \int_\Omega \mathbf{1}_{f^{-1}(A)}(\omega) d\mathbf{P}(\omega) \\ &= \mathbf{P}(f^{-1}(A)) = \mathbf{P} \circ f^{-1}(A) = \int_S \mathbf{1}_A d(\mathbf{P} \circ f^{-1}) \end{aligned}$$

recalling the definition of the integral of an indicator.

2. Let φ be a simple function $\varphi = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ where a_i's are constant and $A_i \in \mathscr{G}$.

$$\begin{aligned} \int_\Omega \varphi \circ f d\mathbf{P} &= \int_\Omega \left(\sum_{i=1}^n a_i \mathbf{1}_{A_i} \right) \circ f d\mathbf{P} \\ &= \int_\Omega \sum_{i=1}^n a_i (\mathbf{1}_{A_i} \circ f) d\mathbf{P} = \sum_{i=1}^n a_i \int_\Omega \mathbf{1}_{A_i} \circ f d\mathbf{P} \\ &\overset{\text{(part 1)}}{=} \sum_{i=1}^n a_i \int_S \mathbf{1}_{A_i} d\mathbf{P} \circ f^{-1} = \int_S \sum_{i=1}^n a_i \mathbf{1}_{A_i} d\mathbf{P} \circ f^{-1} = \int_S \varphi d\mathbf{P} \circ f^{-1} \end{aligned}$$

3. Let φ be a positive measurable function and let φ_n be a sequence of simple functions such that $\varphi_n \nearrow \varphi$ then:

$$\begin{aligned} \int_\Omega \varphi \circ f d\mathbf{P} &= \int_\Omega (\lim_{n\to\infty} \varphi_n) \circ f d\mathbf{P} \\ &= \int_\Omega \lim_{n\to\infty} (\varphi_n \circ f) d\mathbf{P} \overset{\text{monotone convergence}}{=} \lim_{n\to\infty} \int \varphi_n \circ f d\mathbf{P} \\ &\overset{\text{(part 2)}}{=} \lim_{n\to\infty} \int \varphi_n d\mathbf{P} \circ f^{-1} \overset{\text{monotone convergence}}{=} \int \lim_{n\to\infty} \varphi_n d\mathbf{P} \circ f^{-1} \\ &= \int_S \varphi d(\mathbf{P} \circ f^{-1}) \end{aligned}$$

4. Let φ be a measurable function then $\varphi^+ = \max(\varphi, 0)$, $\varphi^- = \max(-\varphi, 0)$. Which then gives us $\varphi = \varphi^+ - \varphi^-$. Since at least one integral is assumed to exist we get that $\int \varphi^+$ and $\int \varphi^-$ exist. Also note that:

$$\begin{aligned} \varphi^+ \circ f(\omega) &= \varphi^+(f^{-1}(\omega)) = \max(\varphi(f(\omega)), 0) \\ \max(\varphi \circ f(\omega), 0) &= (\varphi \circ f)^+(\omega) \end{aligned}$$

Then:

$$\int \varphi^+ d\mathbf{P} \circ f^{-1} = \int \varphi^+ \circ f d\mathbf{P} = \int (\varphi \circ f)^+ d\mathbf{P}$$
$$\int \varphi^- d\mathbf{P} \circ f^{-1} = \int \varphi^- \circ f d\mathbf{P} = \int (\varphi \circ f)^- d\mathbf{P}$$

These equalities follow from part 3 of the proof. After subtracting both:

$$\int \varphi d\mathbf{P} \circ f^{-1} = \int \varphi \circ f d\mathbf{P}$$

Exercise 3.2. If X and Y are independent random variables defined on $(\Omega, \mathbb{R}, P)$ with $X, Y \in L^1(\Omega)$ then $XY \in L^1(\Omega)$:

$$\int_\Omega XY d\mathbf{P} = \int_\Omega X d\mathbf{P} \int_\Omega Y d\mathbf{P} \qquad (\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y))$$

Proof (Solution). This is an exercise that you have seen before, here is presented to exercise the standard approach.

Example 3.3. Let us solve the previous exercise using the transport formula. Let us take $f : \Omega \to \mathbb{R}^2$, $f(\omega) = (X(\omega), Y(\omega))$; and $\varphi : \mathbb{R}^2 \to \mathbb{R}$, $\varphi(x,y) = xy$. Then we have from the transport formula:

$$\int_\Omega X(\omega)Y(\omega) dP(\omega) \overset{(T)}{=} \int_{\mathbb{R}^2} xy dP \circ (X,Y)^{-1}$$

The integral on the left is $\mathbf{E}(XY)$, while the integral on the right can be calculated as:

$$\int_{\mathbb{R}^2} xy d(P \circ X^{-1}, P \circ Y^{-1}) = \int_{\mathbb{R}} x dP \circ X^{-1} \int_{\mathbb{R}} y dP \circ Y^{-1}$$
$$\overset{(T)}{=} \int_\Omega X(\omega) dP(\omega) \int_\Omega Y(\omega) dP(\omega) = \mathbf{E}(X)\mathbf{E}(Y)$$

Example 3.4. Finally we conclude with an application of the transport formula which will produce one of the most useful formulas. Let X be a r.v. defined on the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ with distribution function $F(x)$. Show that:

$$\mathbf{E}(X) = \int_{\mathbb{R}} x dF(x),$$

where the integral is understood in Riemann-Stieltjes sense.

Proving the formula is immediate. Take $f : \Omega \to \mathbb{R}$, $f(\omega) = X(\omega)$ and $\varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(x) = x$. Then from the transport formula:

$$\mathbf{E}(X) = \int_\Omega X(\omega) d\mathbf{P}(\omega) = \int_\Omega x \circ X(\omega) d\mathbf{P}(\omega) \overset{(\mathrm{T})}{=} \int_{\mathbb{R}} x d\mathbf{P} \circ X^{-1}(x) = \int_{\mathbb{R}} x dF(x)$$

Clearly if the distribution function $F(x)$ is derivable with $\frac{dF}{dx}(x) = f(x)$ or $dF(x) = f(x)dx$ we obtain the lower level classes formula for calculating expec-

tation of a "continuous" random variable:

$$\mathbf{E}(X) = \int_{\mathbb{R}} x f(x) dx$$

3.5 Applications. Exercises in probability reasoning.

The next two theorems are presented to observe the proofs. They are both early exercises in probability. We will present later much stronger versions of these theorems (and we will also see that these convergence types have very precise definitions), but for now we lack the tools to give general proofs to these stronger versions.

Theorem 3.4 (Law of Large Numbers). *Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\{X_n\}_n$ be a sequence of i.i.d random variables with $\mathbf{E}(X_i) = \int_\Omega X_i d\mathbf{P} = \mu$. Assume that the fourth moment of these variables is finite and $\mathbf{E}(X_i^4) = K_4$ for all i. Then:*

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{a.s} \mu$$

Proof. Recall what it means for a statement to hold almost surely (a.s.). In our specific context if we denote $S_n = X_1 + \cdots + X_n$ then we need to show that $\mathbf{P}(S_n/n \to \mu) = 1$.

First step. Let us show that we can reduce to the case of $\mathbf{E}(X_i) = \mu = 0$. Take $Y_i = X_i - \mu$. If we prove that $\frac{Y_1 + \cdots + Y_n}{n} \to 0$ then substituting back we shall obtain $\frac{S_n - n\mu}{n} \to 0$, or $\frac{S_n}{n} \to \mu$. Which gives our result. Thus we assume that $\mathbf{E}(X_i) = \mu = 0$.

Second step. We want to show that $\frac{S_n}{n} \xrightarrow{\text{a.s}} 0$. We have:

$$\mathbf{E}\left(S_n^4\right) = \mathbf{E}\left((X_1 + \cdots + X_n)^4\right) = \mathbf{E}\left(\sum_{i,j,k,l} X_i X_j X_k X_l\right)$$

If any factor in the sum above appears with power one, from independence we will have $\mathbf{E}(X_i X_j X_k X_l) = \mathbf{E}(X_i)\mathbf{E}(X_j X_k X_l) = 0$. Thus, the only terms remaining in the sum above are those with power larger than one.

$$\begin{aligned}\mathbf{E}\left(\sum_{i,j,k,l} X_i X_j X_k X_l\right) &= \mathbf{E}\left(\sum_i X_i^4 + \sum_{i<j} \binom{4}{2} X_i^2 X_j^2\right) \\ &= \sum_i \mathbf{E}(X_i)^4 + 6 \sum_{i<j} \mathbf{E}(X_i^2 X_j^2)\end{aligned}$$

Using the Cauchy-Schwartz inequality we get:

$$\mathbf{E}(X_i^2 X_j^2) \leq \mathbf{E}(X_i^4)^{1/2} \mathbf{E}(X_j^4)^{1/2} = K_4 < \infty$$

Then:

$$\begin{aligned}\mathbf{E}(S_n^4) &= \sum_{i=1}^{n}\mathbf{E}(X_i)^4 + 6\sum_{i<j}\mathbf{E}(X_i^2X_j^2) \leq nK_4 + 6\binom{n}{2}\cdot K_4 \\ &= (n+3n(n-1))K_4 = (3n^2-2n)K_4 \leq 3n^2K_4\end{aligned}$$

Therefore:

$$\mathbf{E}\left(\sum_{n=1}^{\infty}\left(\frac{S_n}{n}\right)^4\right) = \sum_{n=1}^{\infty}\frac{\mathbf{E}(S_n^4)}{n^4} \leq \sum_{n=1}^{\infty}\frac{3n^2K}{n^4} = \sum_{n=1}^{\infty}\frac{3K}{n^2} < \infty$$

Since the expectation of the random variable is finite then we must have the random variable finite with the exception of a set of measure 0 (otherwise the expectation will be infinite). This implies:

$$\sum_{n}\left(\frac{S_n}{n}\right)^4 < \infty \quad \text{a.s.}$$

But a sum can only be convergent if the term under the sum converges to zero. Therefore:

$$\lim_{n\to\infty}\left(\frac{S_n}{n}\right)^4 = 0 \quad \text{a.s.}$$

and consequently:

$$\frac{S_n}{n} \xrightarrow{\text{a.s}} 0$$

□

Example 3.5. I cannot resist giving a simple application of this theorem. Let A be an event that appears with probability $\mathbf{P}(A) = p \in (0,1]$. For example, roll a fair six sided die and let A be the event roll a 1 or a 6 ($\mathbf{P}(\{1,6\}) = 1/3$). Let γ_n denote the number of times A appears in n *independent* repetitions of the experiment. Then :

$$\lim_{n\to\infty}\frac{\gamma_n}{n} = p$$

This is an important example for statistics. Suppose for instance that we do not know that the die is fair but we have our suspicions. How do we test? All we have to do is roll the die many times ($n \to \infty$) and look at the average number of times 1 or 6 appears. If this number stabilizes around a different value than $1/3$ then the die is tricked. The next theorem will also tell how many times to roll the dies to be confident in our assessment.

To prove the result we simply apply the previous theorem. Define X_i as:

$$X_i = \begin{cases} 1 \text{ if event } A \text{ appears in repetition } i \\ 0 \text{ otherwise} \end{cases}$$

Then $\mathbf{P}(X_i = 1) = p$ and $\mathbf{P}(X_i = 0) = 1 - p$ so that $\mathbf{E}(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p$. Clearly, the fourth moment is finite as well and applying the theorem: $\gamma_n = \sum_{i=1}^{n} X_i$. will converge to the stated value.

A Basic Central Limit Theorem: *The DeMoivre-Laplace Theorem:*

In order to prove the theorem we need:

Lemma 3.8 (Stirling's Formula). *For large n it can be shown that:*

$$n! \sim \sqrt{2\pi n} \cdot n^n e^{-n}$$

The proof of this theorem is only of marginal interest to us.

Theorem 3.5 (DeMoivre-Laplace). *Let $\xi_1 \cdots \xi_n$ be n independent r.v.'s each taking value 1 with probability p and 0 with probability $1 - p$ (Binomial(p) random variables). Let*

$$S_n = \sum_{i=1}^{n} \xi_i$$

and

$$S_n^* = \frac{S_n - \mathbf{E}(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}}$$

then for any $x_1, x_2 \in \mathbb{R}$, $x_1 < x_2$:

$$\begin{aligned} \lim_{n\to\infty} \mathbf{P}(x_1 \le S_n^* \le x_2) &= \Phi(x_2) - \Phi(x_1) \\ &= \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \end{aligned}$$

Note that Φ is the distribution function of a $N(0,1)$ random variable. This is exactly the statement of the regular Central Limit Theorem applied to Bernoulli random variables.

Proof. Notice that $S_n \sim$ Binomial(n, p) and $S_n^* = (S_n - np)/\sqrt{np(1-p)}$ is distributed equidistantly in the total interval $[\frac{-np}{\sqrt{np(1-p)}}, \frac{n-np}{\sqrt{np(1-p)}}]$. The length between two such consecutive values is $\Delta x = 1/\sqrt{np(1-p)}$.

For k large and $n - k$ large:

$$\begin{aligned}\mathbf{P}(S_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^k \\ &= \frac{\sqrt{2\pi n}\cdot n^n e^{-n}}{\sqrt{2\pi k}\cdot k^k e^{-k}\sqrt{2\pi(n-k)}\cdot (n-k)^{n-k} e^{-(n-k)}} p^k (1-p)^{n-k} \qquad (3.2) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}}_{\text{Term I}} \underbrace{\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}}_{\text{Term II}}\end{aligned}$$

(3.2) follows from Stirling's Formula. Remember that for $S_n = k$ the x value of $S_n^* = (S_n - np)/\sqrt{np(1-p)}$ is:

$$x = \frac{k-np}{\sqrt{np(1-p)}} \Rightarrow k = np + x\sqrt{np(1-p)}$$
$$\Rightarrow \frac{k}{np} = 1 + x\sqrt{\frac{1-p}{np}}$$

Likewise we may express:

$$n-k = n-np-x\sqrt{np(1-p)} \Rightarrow n-k = n(1-p) - x\sqrt{np(1-p)}$$
$$\Rightarrow \frac{n-k}{n(1-p)} = 1 - x\sqrt{\frac{p}{n(1-p)}}$$

Using these two expressions in the Term II of equation (3.2):

$$\log\left(\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}\right) = -k\log\frac{k}{np} - (n-k)\log\frac{n-k}{n(1-p)}$$
$$= -k\log\left(1 + x\sqrt{\frac{1-p}{np}}\right) - (n-k)\log\left(1 - x\sqrt{\frac{p}{n(1-p)}}\right)$$

If we approximate $\log(1+\alpha) \simeq \alpha - \frac{\alpha^2}{2}$ we continue:

$$\simeq -k\left(x\sqrt{\frac{1-p}{np}} - \frac{x^2}{2}\frac{1-p}{np}\right) - (n-k)\left(-x\sqrt{\frac{p}{n(1-p)}} - \frac{x^2}{2}\frac{p}{n(1-p)}\right) \qquad (3.3)$$

Finally, we substitute k and $n-k$ and after calculations (skipped) we obtain:

$$\lim_{n\to\infty} \log\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} = -\frac{x^2}{2}$$

Also note that:

$$\sqrt{\frac{n}{k(n-k)}} \simeq \sqrt{\frac{n}{np \cdot n(1-p)}} = \frac{1}{\sqrt{np(1-p)}}$$

Putting both terms together we obtain:

$$\lim_{n\to\infty} \mathbf{P}(S_n^* = x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Delta x$$

where $\Delta x = \frac{1}{\sqrt{np(1-p)}}$

Thus:

$$\lim_{n\to\infty} \mathbf{P}(x_1 \le S_n^* \le x_2) = \lim_{n\to\infty} \sum_{x_1 \le x \le x_2} \mathbf{P}(S_n^* = x) = \lim_{n\to\infty} \sum \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Delta x$$
$$= \frac{1}{\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-x^2/2} dx$$

□

Problems

3.1. It is well-known that 23 "random" people have a probability of about 1/2 of having at least 1 shared birthday. There are 365 x 24 x 60 = 525,600 minutes in a year. (We'll ignore leap days.) Suppose each person is labeled by the minute in which the person was born, so that there are 525,600 possible labels. Assume that a "random" person is equally likely to have any of the 525,600 labels, and that different "random" people have independent labels.

a) About how many random people are needed to have a probability greater than 1/2 of at least one shared birth-minute? (A numerical value is required.)
b) About how many random people are needed to have a probability greater than 1/2 of at least one birth-minute shared by three or more people? (Again, a numerical value is required. You can use heuristic reasoning, but explain your thinking.)

3.2. Show that any simple function f can be written as $\sum_i b_i \mathbf{1}_{B_i}$ with B_i disjoint sets (i.e. $B_i \cap B_j = \emptyset$, if $i \neq j$).

3.3. Prove the 4 assertions in Exercise 3.1 on page 61.

3.4. Give an example of two variables X and Y which are uncorrelated but not independent.

3.5. Prove the properties (i)-(v) of the expectation in Proposition 3.1 on page 69.

Chapter 4
Product spaces. Conditional Distribution and Conditional Expectation

In this chapter we look at the following type of problems: If we know something extra about the experiment, how does that change our probability calculations. An important part of statistics (Bayesian statistics) is build on conditional distributions. However, what about the more complex and abstract notion of conditional expectation?

Why do we need conditional expectation?

Conditional expectation is a fundamental concept in the theory of stochastic processes. The simple idea is the following: suppose we have no information about a certain variable then our best guess about it would be some sort of regular expectation. However, in real life it often happens that we have some partial information about the random variable (or in time we come to know more about it). Then what we should do is every time there is new information the sample space Ω or the σ-algebra $\mathscr{F}$ is changing so they need to be recalculated. That will in turn change the probability $\mathbf{P}$ which will change the expectation of the variable. The conditional expectation provides a way to recalculate the expectation of the random variable given any new "consistent" information without going through the trouble of recalculating $(\Omega, \mathscr{F}, \mathbf{P})$ every time.

It is also easy to reason that since we calculate with respect to more precise information it will be depending on this more precise information, thus it is going to be a random variable itself, "adapted" to this information.

4.1 Product Spaces

Let $(\Omega_1, \mathscr{F}_1, \mu_1)$ and $(\Omega_2, \mathscr{F}_2, \mu_2)$ be two σ-finite measure spaces. Define:

$$\Omega = \Omega_1 \times \Omega_2 \text{ the cartesian product}$$
$$\mathscr{F} = \sigma(\{B_1 \times B_2 : B_1 \in \mathscr{F}_1, B_2 \in \mathscr{F}_2\})$$

Let $f : \Omega \to \mathbb{R}$ be $\mathscr{F}$ measurable such that

$$\forall \omega_1 \in \Omega_1 \ f(\omega_1, \cdot) \text{ is } \mathscr{F}_2 \text{ measurable on } \Omega_2$$
$$\forall \omega_2 \in \Omega_2 \ f(\cdot, \omega_2) \text{ is } \mathscr{F}_1 \text{ measurable on } \Omega_1$$

Then we define:

$$I_1^f(\omega_1) = \int_{\Omega_2} f(\omega_1, \omega_2)\mu_2(d\omega_2)$$
$$I_2^f(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2)\mu_1(d\omega_1)$$

which are a kind of partial integrals, well defined by the measurability of the restrictions.

Theorem 4.1 (Fubini's theorem). *Define a measure:*

$$\mu(F) = \int_{\Omega_1}\int_{\Omega_2} 1_F(\omega_1, \omega_2)\mu_1(d\omega_1)\mu_2(d\omega_2).$$

Then μ is the unique measure defined on $(\Omega, \mathscr{F})$ called the product measure *with the property:*

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2) \qquad \forall A_i \in \mathscr{F}_i,$$

and as a consequence:

$$\int_\Omega f d\mu = \int_{\Omega_1} I_1^f(\omega_1)\mu(d\omega_1) = \int_{\Omega_2} I_2^f(\omega_2)\mu(d\omega_2)$$

Proof. Skipped. Apply the standard argument. Note that the first step is already given.

Example 4.1 (Application of Fubini's Theorem). Let X be a positive r.v. on $(\Omega, \mathscr{F}, P)$. Consider $P \times \lambda$ on $(\Omega, \mathscr{F}) \times ([0,\infty), \mathscr{B}((0,\infty])),$ where λ is the Lebesgue measure. Let $A := \{(\omega, x) : 0 \le x < X(\omega)\}$. Note that A is the region under the graph of the random variable X. Let the indicator of this set be denoted with $h = 1_A$. Then:

$$I_1^h(\omega) = \int_{[0,\infty)} \mathbf{1}_A(\omega, x)d\lambda(x) = \int_0^\infty \mathbf{1}_{\{0 \le x < X(\omega)\}}(x)d\lambda(x) = \int_0^{X(\omega)} d\lambda(x) = X(\omega)$$
$$I_2^h(x) = \int_\Omega \mathbf{1}_A(\omega, x)d\mathbf{P}(\omega) = \int_\Omega \mathbf{1}_{\{0 \le x < X(\omega)\}}(\omega)d\mathbf{P}(\omega) = \mathbf{P}\{\omega : X(\omega) > x\},$$

since X is a positive r.v.

We now apply Fubini's Theorem and we get :

$$\begin{aligned}\mu(A) &= \int_\Omega \int_{[0,\infty)} 1_A(x, \omega)d\mu(x)d\mathbf{P}(\omega) \\ &= \int_\Omega X(\omega)d\mathbf{P}(\omega) = \int_0^\infty \mathbf{P}(X > x)dx\end{aligned}$$

Thus reading the last line above:

$$\mathbf{E}(X) = \int_0^\infty \mathbf{P}(X > x)dx$$

This result is actually so useful that we will state it separately.

Corollary 4.1. *If X is a **positive** random variable with distribution function $F(x)$ and we denote $\overline{F}(x) = 1 - F(x)$, we have:*

$$\mathbf{E}(X) = \int_0^\infty \overline{F}(x)dx$$

4.2 Conditional distribution and expectation. Calculation in simple cases

We shall give a general formulation of conditional expectation that will be most useful in the second part of this textbook. But, until then we will present the cases that actually allow the explicit calculation of conditional distribution and expectation.

Let X and Y be two discrete variables on $(\Omega, \mathscr{F}, P)$.

Definition 4.1 (Discrete Conditional Distribution). The conditional distribution of Y given $X = x$: $F_{Y|X}(\cdot|x)$ is:

$$F_{Y|X}(y|x) = \mathbf{P}(Y \leq y|X = x)$$

The conditional probability mass function of $Y|X$ is:

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

Note: In the case when $\mathbf{P}(X = x) = 0$ we cannot define the conditional probability.

Definition 4.2 (Discrete Conditional Expectation). Let $\psi(x) = \mathbf{E}(Y|X = x)$ then $\psi(X) = \mathbf{E}[Y|X]$ is called the conditional expectation.

Remark 4.1. The conditional expectation is a random variable.

Definition 4.3 (Continuous Conditional Distribution). Let X, Y be two continuous random variables. The conditional distribution is defined as:

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

The function $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$ is the conditional probability density function.

Definition 4.4 (Continuous Conditional Expectation). The conditional expectation for two continuous random variables is $\psi(X) = \mathbf{E}[Y|X]$ where the function ψ is calculated:

$$\psi(x) = \mathbf{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

Example 4.2. A point is picked uniformly from the surface of the unit sphere. Let L =longitude angle θ and let l = latitude angle ϕ. Let us find the distribution functions of $\theta|\phi$ and $\phi|\theta$.

Let C be a set on the sphere (or generally in $\mathbb{R}^3$). The surface area of the sphere is $4\pi r^2 = 4\pi$. The set of points from which we sample is $S(0,1) = \{(x,y,z) : x^2+y^2+z^2 = 1\}$. Then, since we pick the points uniformly the position of a point chosen has distribution:

$$\mathbf{P}((x,y,z) \in C) = \int_C \frac{1}{4\pi} \mathbf{1}_{\{x^2+y^2+z^2=1\}}(x,y,z)dxdydz$$

Since we are interested in longitude and latitude we change to polar coordinates to obtain the distribution of these variables. We take the transformation: $X = r\cos\theta\cos\phi$, $Y = r\sin\theta\cos\phi$ and $Z = r\sin\phi$. To obtain the distribution we calculate the new integral. The Jacobian of the transformation is:

$$\begin{aligned} J &= \begin{vmatrix} -r\sin\theta\cos\phi & -r\cos\theta\sin\phi & \cos\theta\cos\phi \\ r\cos\theta\cos\phi & -r\sin\theta\sin\phi & \sin\theta\cos\phi \\ 0 & r\cos\phi & \sin\phi \end{vmatrix} \\ &= r^2\cos^3\phi + r^2\sin^2\phi\cos\phi = r^2\cos\phi \end{aligned}$$

Note that the indicator is 1 if and only if $r = 1$. We conclude that

$$\mathbf{P}((x,y,z) \in C) = \int_{\mathrm{Im}\, C} \frac{1}{4\pi}|\cos\phi| d\theta d\phi,$$

where *ImC* is the set of *polar* coordinates that make the set C. Therefore, the joint distribution function is

$$f(\theta,\phi) = \frac{1}{4\pi}|\cos\phi|, \qquad \phi \in [-\pi/2, \pi/2], \theta \in [0, 2\pi].$$

Now, we get the marginal of ϕ:

$$f_\phi(\phi) = \int_0^{2\pi} \frac{1}{4\pi}|\cos\phi| d\theta = \frac{|\cos\phi|}{2},$$

and the marginal of θ:

$$f_\theta(\theta) = \int_{-\pi/2}^{\pi/2} \frac{1}{4\pi}|\cos\phi| d\phi = \int_{-\pi/2}^{\pi/2} \frac{1}{4\pi}\cos\phi d\phi = \frac{1}{2\pi}$$

Thus, we calculate immediately the conditional distributions:

$$f_{\theta|\phi}(\theta|\phi) = \frac{1}{2\pi}, \qquad \theta \in [0, 2\pi]$$
$$f_{\phi|\theta}(\phi|\theta) = \frac{\cos\phi}{2}, \qquad \phi \in [-\pi/2, \pi/2]$$

We note that θ and ϕ are independent (the product of marginals is equal to the joint distribution) but the conditionals are different due to the parameterizations (this particular example is known as *the Borel paradox*). Also note that the conditional expectations are equal to the regular expectations, this is of course because the variables are independent. We will obtain this property in general in the following section.

Example 4.3. Many clustering algorithms are based on random projections. For simplicity we consider the direction of the first coordinate unit vector $\overrightarrow{e}_1$ as the best possible projection. However, the probability of finding this direction exactly is zero so we consider a tolerance angle α_e and we say that a projection is "good enough" if it makes an angle less than α_e with $\overrightarrow{e}_1$.

We want to calculate the probability that a random direction makes an angle less than α with $\overrightarrow{e}_1$.

The example is in $\mathbb{R}^3$ but we can easily generalize it to any dimension. We assume that $0 < \alpha_e < \pi/2$, otherwise the problem becomes trivial.

Directions in $\mathbb{R}^3$ are equivalent to points on the unit sphere. Therefore, the probability to be calculated is twice the probability that a point chosen at random on the sphere belongs to the cone of angle α_e centered at the origin. Why twice? Because we do not care if the angle formed by the random direction is with $\overrightarrow{e}_1$ or $-\overrightarrow{e}_1$. Thus, we calculate the probability by taking the ratio of the area of the intersection of the said cone and the sphere and the total surface area of the sphere.

The area of the unit sphere in $\mathbb{R}^d$ is readily calculated as $\frac{2\pi^{d/2}}{\Gamma(d/2)}$ (e.g., Kendall (2004), $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function). In the particular case when $d = 3$ ($\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$) we obtain the well known area 4π.

To compute the support area of the cone we switch to polar coordinates:

$$\begin{aligned} x_1 &= r\cos\theta_1 \\ x_2 &= r\sin\theta_1\cos\theta_2 \\ x_3 &= r\sin\theta_1\sin\theta_2 \end{aligned}$$

where$r \in [0,\infty), \theta_1 \in [0,\pi], \theta_2 \in [0, 2\pi]$.

The points of interest can be found when $r = 1$ and $\theta_1 \in [0, \alpha_e]$, and we need to double the final area found to take into account symmetric angles with respect to $\overrightarrow{e}_1$.

One can check immediately, that the Jacobian of this change of variables is $r^2\sin\theta_1$ and that the probability needed is easily calculated as:

$$2\sin^2\frac{\alpha_e}{2}$$

If we now consider K projections then the probability that at least one is a "good enough projection" is:

$$1-\left(1-2\sin^2\frac{\alpha_e}{2}\right)^K$$

Note that the example is extendable to the more interesting R^d case but in that case we do not obtain an exact formula instead only bounds. See **Ion: give citation once it exists.**

4.3 Conditional expectation. General definition

To summarize the previous section, if X and Y are two random variables we have defined the conditional distribution and conditional expectation of one **with respect to the other**. In fact, we have defined more: the conditional expectation of one **with respect to the information contained in the other**.

More precisely, in the previous subsection we defined the expectation of X conditioned by the σ-algebra generated by Y: $\sigma(Y)$. Thus, we may write without a problem:

$$\mathbf{E}[X|Y]=\mathbf{E}[X|\sigma(Y)].$$

This notion may be generalized to define conditional expectation with respect to any kind of of information (σ-algebra). As definition we shall use the following theorem. We will skip the proof.

Theorem 4.2. *Let $(\Omega,\mathscr{F},\mathbf{P})$ be a probability space, and let $\mathscr{K}\subseteq\mathscr{F}$ a sub-σ-algebra. Let X be a random variable on $(\Omega,\mathscr{F},\mathbf{P})$ such that either X is positive or $X\in L^1(\Omega)$. Then, there exists a random variable Y, measurable with respect to $\mathscr{K}$ with the property:*

$$\int_A YdP=\int_A XdP \quad ,\forall A\in\mathscr{K}$$

This Y is defined to be the conditional expectation of X with respect to $\mathscr{K}$ and is denoted $\mathbf{E}[X|\mathscr{K}]$.

Remark 4.2. We note that the conditional expectation, unlike the regular expectation is a random variable measurable with respect to the sigma algebra under which is conditioned. In simple language it has adapted itself to the information contained in the σ-algebra $\mathscr{K}$. In the simple cases presented in the previous section the conditional expectation is measurable with respect to $\sigma(Y)$. But since this is a very simple sigma algebra then it has to be in fact a function of Y.

Note: We will take this theorem as a definition.

Proposition 4.1 (Properties of the Conditional Expectation). *Let $(\Omega,\mathscr{F},\mathbf{P})$ a probability space, and let $\mathscr{K},\mathscr{K}_1,\mathscr{K}_2$ sub-σ-algebras. Let X and Y be random variables of the probability space. Then we have:*

(1) If $\mathscr{K}=\{\varnothing,\Omega\}$ then $\mathbf{E}[X|\mathscr{K}]=\mathbf{E}X=const.$
(2) $\mathbf{E}[\alpha X+\beta Y|\mathscr{K}]=\alpha\mathbf{E}[X|\mathscr{K}]+\beta\mathbf{E}[Y|\mathscr{K}]$ for α,β real constants.
(3) If $X\le Y$ a.s. then $\mathbf{E}[X|\mathscr{K}]\le\mathbf{E}[Y|\mathscr{K}]$ a.s.
(4) $\mathbf{E}\left[\mathbf{E}[X|\mathscr{K}]\right]=\mathbf{E}X$
(5) If $\mathscr{K}_1\subseteq\mathscr{K}_2$ then

$$\mathbf{E}\left[\mathbf{E}[X|\mathscr{K}_1]|\,\mathscr{K}_2\right]=\mathbf{E}\left[\mathbf{E}[X|\mathscr{K}_2]|\,\mathscr{K}_1\right]=\mathbf{E}[X|\mathscr{K}_1]$$

(6) If X is independent of $\mathscr{K}$ then

$$\mathbf{E}[X|\mathscr{K}]=\mathbf{E}[X]$$

(7) If Y is measurable with respect to $\mathscr{K}$ then

$$\mathbf{E}[XY|\mathscr{K}]=Y\mathbf{E}[X|\mathscr{K}]$$

After proving these properties (see Problem 4.2) they will become essential in working with conditional expectation. In fact the definition is never used anymore.

Example 4.4. Let us obtain a weak form of the Wald's equation (an equation that serves a fundamental role in the theory of stochastic processes) right now by a simple argument. Let $X_1,X_2,\ldots,X_n,\ldots$ be i.i.d. with finite mean μ and let N be a random variable taking values in strictly positive integers and independent of X_i for all i. For example, X_i's may be the results of random experiments and N may be some stopping strategy established in advance. Let $S_N=X_1+X_2+\cdots+X_N$. Find $\mathbf{E}(S_N)$.

Let

$$\begin{aligned}\varphi(n)&=\mathbf{E}[S_N|N=n]=\mathbf{E}[X_1+X_2+\cdots+X_N|N=n]\\&=\sum_{i=1}^{n}\mathbf{E}[X_i|N=n]=\sum_{i=1}^{n}\mathbf{E}[X_i]=n\mu\end{aligned}$$

by independence. Therefore, $\mathbf{E}[S_N|N]=\varphi(N)=N\mu$. Finally, using the properties of conditional expectation:

$$\mathbf{E}(S_N)=\mathbf{E}[\mathbf{E}[S_N|N]]=\mathbf{E}[N\mu]=\mu\mathbf{E}[N].$$

Note that we have not used any distribution form only the properties of the conditional expectation.

Problems

4.1. Prove the Fubini's Theorem 4.1 on page 80.

4.2. Using the Theorem-Definition 4.2 on page 84 prove the seven properties of the conditional expectation in Proposition 4.1.

4.3. Let X be a random variable on the probability space $(\Omega, \mathscr{F}, \mathbf{P})$. Let a set $A \in \mathscr{F}$ and the sigma algebra generated by the set denoted $\sigma(A)$. What is $\mathbf{E}[X|\sigma(A)]$? Let $\mathbf{1}_A$ denote the indicator of A. What is $\mathbf{E}[X|\mathbf{1}_A]$?

4.4. Let X, Y, Z be three random variables with joint distribution

$$P(X = k, Y = m, Z = n) = p^3 q^{n-3}$$

for integers k, m, n satisfying $1 \leq k < m < n$, where $0 < p < 1$, $p + q = 1$. Find $E\{Z|X, Y\}$.

4.5. A circular dartboard has a radius of 1 foot. Thom throws 3 darts at the board until all 3 darts are sticking in the board. The locations of the 3 darts are independent and uniformly distributed on the surface of the board. Let T_1, T_2, and T_3 be the distances from the center to the closest dart, the next closest dart, and the farthest dart, respectively, so that $T_1 \leq T_2 \leq T_3$. Find $\mathbf{E}[T_2]$.

4.6. Let $X_1, X_2, \ldots, X_{1000}$ be i.i.d. each taking on both 0 and 1 with probability $\frac{1}{2}$. Put $S_n = X_1 + \cdots + X_n$. Find $\mathbf{E}\left[(S_{1000} - S_{300})\mathbf{1}_{\{S_{700}=400\}}\right]$ and $\mathbf{E}\left[(S_{1000} - S_{300})^2\mathbf{1}_{\{S_{700}=400\}}\right]$.