Introductory Course on Probability Theory
European University Institute
September 2006

Tobias Broer

[1]

[1] I thank David Scherrer and Joel Van der Weele for detailed comments on an earlier version.

Contents

1 Introduction

1.1 Examples

1.1.1 Example 1: Betting on a coin toss

A friend offers you a bet: you pay 1 euro, and get 2 euros when exactly one head occurs in two tosses. How do you decide wether to accept or reject?

The situation you face is random, i.e. there is uncertainty about what is going to happen. In order to decide wether to accept the bet, we need to describe the structure of this uncertainty. So you probably would like to know

- The set, and number, of possible "outcomes", here $\{(H,H),(T,H),(H,T),(T,T)\}$.
- The set of outcomes where the statement "exactly one head" is true, which we will call the "Event" "exactly one head"; and the set of outcomes where is it not, or the event "not exactly one head".
- The probability of any one of the outcomes where "exactly one head" is true to occur - i.e. the "probabilistic size" of the set of outcomes where the statement is true relative to that of all possible outcomes. With a fair coin (i.e. all outcomes are equally likely) this is equal to the number of outcomes in both sets divided by the number of all possible outcomes 4.

This example shows that to analyse simple probability mechanisms, we need to formalise three related objects:

1. The set of possible outcomes (or the "sample space" of a random experiment).
2. A "coherent" family of subsets to the set of all outcomes defined by possible statements such as "Exactly one H", or alternatively "At least one tail", "Three heads", etc. (The subsets of outcomes described by these statements are also called "events", the set of events under consideration a "sigma algebra" of the sample space.)
3. The concept of "size" of these sets as the probability of any one of the outcomes they contain to occur.

These are the three main elements of "Probability set theory", which we will look at in section II of this course.

1.1.2 Example 2: Ringing a family's doorbell

You ring the bell of a house where a couple lives with their two children.
a) What is the probability that a boy opens the door?
b) The door opens, and a boy says hello to you. What is the probability that the other child is also a boy?

Answer 1
a: The set of possible "children outcomes", ordering the younger child first, is $\{(B,B),(B,G),(G,B),(G,G)\}$. Supposing that the probability of the younger child opening the door is 1/2, and the probability of a child being born as a boy is also 1/2, the probability of a boy opening the door is $P(Y\ opens \cap Y\ is\ a\ boy) + P(O\ opens \cap O\ is\ a\ boy) = 2 * (1/2 * 1/2) = 1/2$.

b: One is tempted to say one half, as of course the probability that any other child is born as a boy is one half. But now we know that there is at least one boy in the family. So we can discard the outcome (G,G), and the set of possible outcomes is reduced to $\{(B,B),(B,G),(G,B)\}$. The statement is true in only one case, $\{B,B\}$, all outcomes are equally likely with probability 1/3 once we discarded the event (G,G) so the probability of "the second child is a boy" is 1/3. (Alternatively to get the "conditional" probability of $\{B,B\}$, given that only one of the outcomes $\{(B,B),(B,G),(G,B)\}$ will occur, one can "scale" the original, or "unconditional" probability of $\{B,B\}$ and $\{(B,B),(B,G),(G,B)\}$ occuring together (which is simply the probability of $\{B,B\}$, so 1/4) by the original probability of the remaining set of possible outcomes once we discard (G,G) (which is 3/4). So the probability is $1*1/4/(3/4) = 1/3$. This is "Bayes' Law".)

Answer 2
Now we try to answer the question by distinguishing wether the younger or the older child opens the door.
a: The set of possible outcomes, where we order the younger child first, is $\{(B,B,younger\ child\ opens),(B,G,younger\ child\ opens),(G,B,younger\ child\ opens),(G,G,younger\ child\ opens),(B,B,older\ child\ opens),(B,G,older\ child\ opens),(G,B,older\ child\ opens),(G,G,older\ child\ opens)\}$ The statement is true in 4 of these 8 outcomes. Supposing that the probability of having a boy is 1/2 and that both children answer the door equally likely, we get the probability as $4 * 1/8 = 1/2$.

b: Once a boy opens, the set of possible outcomes is reduced to $\{(B, B, \textit{younger child opens}), (B, G, \textit{younger child opens}), (B, B, \textit{older child opens}), (G, B, \textit{older child opens})$. The statement is true in 2 cases, but the probability of the reduced set of outcomes is 1/2 as we have seen before, so we scale according to Bayes' law, and get the probability $\frac{2 \cdot 1/8}{1/2} = 1/2$. Why?

Answer3
Instead of looking at the set of possible sequences of boys and girls, and the subsets in which the statements are true, one can simply look at the number of boys in the family. This is an integer in the interval $[0, 2]$, and we can get the probabilities of these "numbers" by summing up the probabilities of the events that yield those numbers. Denote the number of boys as a function X of the outcomes of the random experiment. So for X="number of boys" we get X(G,G)=0, X(B,B)=2, X(B,G)=X(G,B)=1. The probability of the number X is simply the probability of the outcomes that yield that number, so we get P(0)=P(2)=1/4, P(1)=1/2. In this case, the probability of a the second child also being a boy is simply the probability of there being two boys, given that there is at least one, which again is 1/3.

This example shows first of all, probabilities are mind-twisting things sometimes, especially if we have to incorporate new information, and thus look at "conditional probabilities" (as we will do in section III).
In answer 3, we have introduced a function X from the set of all possible outcomes to a subset of the real line, here the set of values $\{0, 1, 2\}$. Such a function is called a "random variable". Section IV of the course defines the concept of a "random variable" more closely, and shows how we can describe the relative probabilities of its values by a "probability distribution function".

1.1.3 Example 3: Betting on your post-PhD salary

A friend from your Masters course offers the following bet: You get 1000 euros if your salary is higher than hers in 5 years. Otherwise you pay her 1000 euros. How do you decide whether doing a PhD increases your chances of winning the bet?
First of all, there are a lot of things that can happen in those 5 years to you and your friend. But no matter what happens, there is always a salary attached to it (if we identify no salary with 0 salary). So you can view your salary after 5

years as a function from the space of possible outcomes onto the positive real line, which is the definition of a random variable. But when looking at the events associated to amounts of salary, there is a problem: First, the the set of possible salaries, or the induced "sample space", is more difficult to handle, as your salary can be any amount between 0 and a very large real number. So the number of possible outcomes is approximately infinite. In other words, X is an approximately *continuous random variable* here. Also, this means that the probability of any salary, e.g. "$Prob(X = exactly\ 50.000)$" is very small, or zero.

This example shows that defining the sets of possible events for a random experiment with infinite, or continuous, outcomes needs special care, as we see in section II. When we interpret the salary as a random variable that can take any real non-negative number, we will have to define events as subsets of the real line that one can use to define the probability distribution associated to a continuous random variable X, a concept introduced in section IV.
Section V looks at how to summarise the shape of probability functions in terms of "moments" of location (what is the mean salary I can expect when doing a PhD?), spread (what is a likely range for the salary to fall into ...), skewness, etc.

Apart from the continuous character of the set of outcomes, example 3 also shows a completely new problem. In the earlier examples, we knew fairly well the characteristics of the random experiment, such as the number of outcomes and their probabilities, and the difficulty was to find a formalisation in order to analyse them. But here, we don't know what the random experiment at work exactly is. In other words we would like to somehow get an idea about the probability distribution of the random variables X: "salary of a Masters student in 5 years with and without a PhD".
One obvious way to do this is to ask a number of Masters students for their salary 5 years after the end of the course, and compare the average of those doing a PhD with those not doing one. I.e. we calculate on the basis of "data", or observations on a random variable, a number which is a function of the data, or a "statistic", here the mean. Clearly, as a function of a random variable, any statistic is a random variable in itself and so has a distribution. Under certain assumptions about the process generating the data, e.g. independence of observations, exogeneity of conditioning variables (here the PhD, a strong assumption), etc., one can derive the sampling disbribution of a statistic such

as the mean of the data as a function of the distribution of the random variable X itself. So given our data, we can try best guesses about features, generally moments, of the underlying distribution of our original random variable (estimation). Or we can assess how likely certain statements are to be true, e.g. "A PhD increases my salary" (Hypothesis testing). This will be done in the main part of the course.

1.1.4 Example 4: A simple Macroeconomic model of Optimal Growth

Take the period 0 problem of an infinitely-lived representative consumer that maximises expected discounted utility. She has access to a technology F that takes capital as input, and owns initial capital k_0. We can summarise her decision problems as $max_{c_t,k_{t+1}} E_0 \sum_{t=0}^{\infty} \beta^t U(c_t)$ subject to the budget constraint $k_{t+1} = F(k_t) - c_t$, and given k_0, z_0. Now suppose technology is stochastic with $F(k_t) = z_t f(k_t)$, where z_t is a random variable, and that $f()$ and $U()$ have the usual properties (increasing, twice continuously differentiable, strictly (quasi)concave, etc.). Without solving the problem, we can see that probability theory can help us address several potential problems contained in this setup.

- What is the dependence structure of the z_t across different periods? Are they independent, or do they perhaps have some kind of "recursive" structure?
- What is the set of possible outcomes of z_t and thus that of outcomes for the sequence $\{z_t\}_{t=0}^{\infty}$.
- Is expected utility defined for all possible $\{z_t\}_{t=0}^{\infty}$, and $U()$, or do we have to place restrictions on these to have a well-defined expectation?

1.2 Aim of the course

The course is meant to

- Give students the background in probability theory they need for the compulsory courses.
- Enable students to master more advanced texts in probability and measure theory.

More particulary, the course wants to

- provide tools to formally describe and analyse random experiments.

- introduce the concepts of both discrete and continuous random variables and their probability distributions.
- provide an overview of the main univariate and multivariate probability distributions and their characteristics.
- show students how probability theory is a special case of the more general theory of measures, measurable functions, etc., and how this can help to formalise probability (without being indispensable for a basic understanding of the concepts).
- provide enough possibilities for practice.

1.3 Probability Theory vs. Statistics vs. Econometrics

Probability Theory aims to derive characteristics of probability mechanisms on the basis of a limited number of definitions and axioms. For example what is the variance of a random variable X="number of heads in a triple coin toss"?
Statistics takes observations on a particular random experiment, or "data", together with some maintained assumptions on the underlying probability mechanism (e.g. independence of observations; general structure such as linearity, normality, etc.), and tries to "estimate" the parameters of the probability mechanism, or assess "hypotheses" about them. For example, given a sequence of observations on the number of heads in triple coin tosses, what is the best estimate for the parameter of the probability mechanism, the probability of "Head" in a single toss. Can we reject the hypothesis that the coin is "fair" with sufficient confidence?
Econometrics applies statistics to assess the likelihood of economic models and theories to be true, bearing in mind that unlike in other fields, experiments are often not possible. Or in Greene's (2003) words, "Econometrics is the field of economics that concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory."

1.4 Outline

The first part of the course follows roughly the problems raised by the examples above. In **Section II**, we discuss some basic definitions, and characterise formally a consistent set of event, and the "probability set function" as a mapping from this set to the interval [0,1]. We then show that these concepts are just

special cases of some more general notions in measure theory (i.e. the set of all possible events is a "Sigma-algebra" of the sample space, on which the probability set function provides a "measure", etc.). The section then discusses the properties of these probability set functions including independence of events and Bayes' Law. Section III discusses two things: First, the incorporation of new information into probabilities in the form of "conditional probability", and the concept of independence. Second, it discusses some rules for counting selections of outcomes from the sample space - this is very useful when all outcomes are equally likely and the probability of an event is determined by the number of elementary outcomes it comprises relative to the total number of possible outcomes.

Section II only considers probability functions defined on arbitrary events, or subsets of possible outcomes of any random experiment. But often we want to look at numerical characteristics of random experiments, such as "the number of heads" in a coin toss. Thus, **Section IV** introduces and discusses the concept of a "random variable" that maps outcomes into real numbers, and shows that a random variable is simply a special case of a "measurable function". We then define the "distribution function" as a convenient way to summarise the probabilities associated to values of both continuous and discrete random variables. The section closes by discussing how to change between two random variables when one is a function of the other.

To summarise distribution functions, we often use "moments" such as expected value or variance. **Section V**, after a brief discussion of integration theory, moves on to define these moments in terms of integrals and weighted sums of functions of random variables.

Section VI presents some common univariate distributions. **Section VII** discusses multivariate random variables, their joint distribution as compared to marginal or conditional distributions, and the concept of independence of random variables. **Section VIII** gives some common multivariate distributions.

1.5 References

The first chapters of *Goldberger (1991)* give an intuitive and easy-to-read introduction to probability theory and statistics. It contains very little formal mathematics, but gives most of the things one needs to survive. *Hogg and Craig (various editions)* is a classic, thorough yet readable introduction to mathematical statistics, including plenty of examples and exercises with answers. The appendix to *Hansens's lecture notes* is a very concise reference, giving all basic

definitions, distributions, etc. with some limited intuition and proofs, and without too much maths. See http://www.ssc.wisc.edu/ bhansen/notes/notes.htm. *Benoît Champagne's class notes* are another very good introduction with plenty of examples, that takes you quite far in terms of probability and measure theory. See http://www.ece.mcgill.ca/ info305a/class%20notes.html. *Wilde's script* provides on some 70 pages a very thorough introduction to probability on the basis of measure and integration theory. Recommended to those who like maths. See http://www.mth.kcl.ac.uk/ iwilde/notes/mip/mip.pdf. *Spanos (1986)* is also a very rigorous introduction to axiomatic probability theory, without stressing the mathematical concepts as much. If you struggle with some individual concepts, http://mathworld.wolfram.com/ is a useful reference.

2 Probability spaces

A situation is random, as opposed to deterministic, when something may or may not occur, or when the follow-on situation is uncertain. Probability tries to introduce regularity into randomness.
More particularly, probability aims to attach to possible statements about future situations a number that describes their likelihood to be true once the situation arises. But these statements are not necessarily themselves single situations: "exactly one Head" in the repeated coin toss describes two distinct situations: $\{H, T\}$ and $\{T, H\}$. We call these statements events. This section describes how to attach probabilities to events.
There are different approaches to probability theory. The **classical approach** (Laplace 1812) takes the probability of an event A describing a subset of the set of possible outcomes S as the ratio of N_A, the number of outcomes in A, to N, the number of outcomes in S. Or $P(A) = \frac{N_A}{N}$. The **frequency approach** (Van Mises 1919) takes probability as the limit of relative frequency, i.e. the limit of the ratio of trials where an event occurs to the total number of trials as the latter goes to infinity. Or $P(A) = lim_{N \longrightarrow \infty} \frac{N(A\ occurs)}{N}$. Both these approaches have difficulties. We base our exposition on the so-called **axiomatic approach** to probability theory, where probability is defined as a function on a set of events that fulfills a number of axioms. But first we need some definitions.

2.1 Random Experiment, Events, and Sigma-Algebras

2.1.1 Random Experiment Ξ

Definition: A random experiment Ξ is a situation with different possible *outcomes* (follow-on situations), such that

1. There is always exactly one outcome.
2. All possible outcomes are known a priori.
3. In a particular trial, the outcome is not known a priori.
4. The situation is repeatable.

A particular realisation of a random experiment, yielding a particular outcome, is call a **trial**.
Example 1: Repeated coin toss
Example 2: Rainfall in Florence in August (Although condition 4 is debatable here.)

2.1.2 Sample Space S of a random experiment

Definition: The set of all possible outcomes of a random experiment Ξ is called the "Sample Space", which we denote S. Elements of S are called outcomes or "elementary events".

Example 1: Repeated coin toss: $S = \{\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}\}$.
Example 2: Rainfall in Florence in August: S = R^+.

2.1.3 Event

An event in a random experiment Ξ is "any proposition associated with Ξ which may occur or not at each trial" (Spanos). Or more formally:
Definition: Any collection of outcomes, or subset of the sample space S is an event, including sure (S) and impossible event ($\emptyset$).
An event "occurs" if one of the outcomes it comprises occurs.

Note: Given the definition of an event on the basis of subsets, we can apply the algebra of set operations to events, e.g. complements, unions, intersections, etc. In particular, and quite intuitively, for two events A_1 and A_2, the following are also events:

- "not A_1", which is the complementary set of A_1 relative to S, or A_1^c.
- "A_1 and/or A_2", which is the set equal to the union $A_1 \cup A_2$.
- "A_1 and A_2", which is the set equal to the intersection $A_1 \cap A_2 = (A_1^c \cup A_2^c)^c$.
- "A_1 but not A_2", which is the set $A_1 \backslash A_2 = (A_1^c \cup A_2)^c$.

Note that we have reduced the four set operations on events to union and complementarisation. This will be useful in the following.

Exercise: Write down the subsets of S that correspond to the following events
Example 1: At least one head in 2 coin tosses.
Example 2: Rainfall in Florence in August of more than 20 liters (per square meter).

Definitions:
If for two events A and B $A \cap B = \emptyset$, then A and B are called "disjoint", or

"mutually exclusive".
If all events $A_1, A_2, ...$ are pairwise disjoint, and "collectively exhaustive", i.e. $\bigcup_i A_i = S$, the collection $A_1, A_2, ...$ is called a "partition of S".

2.1.4 Sets of events Ω

Ultimately, we would like to associate probabilities to events. But first we need to define a consistent set of events that we can use as the domain for probability. With a coin toss, a dice throw, etc. this is straightforward in principle, as there is a finite number of outcomes. So we can just use the set of all possible subsets of S, which like S itself has a finite number of elements. But the rainfall example shows that with approximately continuous sample spaces, defining the set of possible events is more difficult, as there is an infinite number of outcomes that can be "combined" into events.
More generally, we need a concept to describe a set of events as a family of subsets of S that is consistent with its members being "events", and thus implying other events by set operations. From the definition of an event, and the fact that we can reduce four set operations to two, any set of events Ω needs to satisfy the following: For every event A_i in Ω

1. A_i may not occur, so $A_i^c = S \setminus A_i$ must be in Ω.
2. The event "A_1, and/or A_2, and/or ...." is an event, so must be in Ω.
3. The event "A_1, and/or A_1^c" is the "sure event" equal to the set of all possible outcomes, so $S \in \Omega$.
4. From 1 and 3, the "impossible event" $S^c = \emptyset \in \Omega$.

2.1.5 Sigma-algebra $\Im$ of S

A consistent set of events Ω is also called a "Sigma-algebra" (or "Sigma-field") on the sample space S, as the following definition shows.

Definition: A family $\Im$ of subsets of a set S is called a "Sigma-algebra" of S, if

1. For every $A \in \Im$, $A^c = \{s \in S : s \notin A\} \in \Im$. ($\Im$ is closed under complementation.)
2. For every sequence of $A_i \in \Im$, i = 1,2,..., $\bigcup_{i=1}^{\infty} A_i \in \Im$. ($\Im$ is closed under countable union.)

This implies:

3. $S \in \Im$ (since $A \cup A^c \cup \emptyset \cup \emptyset \cup ... = S \in \Im$)
4. $\emptyset \in \Im$ (since $S^c = \emptyset \in \Im$)
5. $(\bigcup_{i=1}^{\infty} A_i)^c = \bigcap_{i=1}^{\infty} A_i^c \in \Im$

The pair (S,$\Im$) is called a **"measurable space"**.

Note: For S with a finite, or countably infinite, number of elements, we can use as Sigma-Algebra the "**power set**" P of S, written P(S).
Definition: The set of all subsets of S is called the power set of S.

Example 1: Toss coin once, i.e. $S = \{H, T\}$
One Sigma-algebra is the power set $P = \{\emptyset, S, \{H\}, \{T\}\}$
But $\Omega_1 = \{\emptyset, S\}$ is also a Sigma-algebra for S, called the "trivial Sigma-algebra".
Example 2: Toss coin twice, i.e. $S = \{\{H,H\},\{H,T\},\{T,H\},\{T,T\}\}$
$P = \{\emptyset, S, \{(H,T)\}, \{(H,H)\}, \{(T,T)\}, \{(T,H)\},$
$\{(H,T),(H,H)\}, \{(T,T),(H,H)\}, \{(T,H),(H,H)\}, \{(H,T),(T,H)\},$
$\{(T,T),(T,H)\}, \{(H,T),(T,T)\},$
$\{(H,T),(T,H),(T,T)\},\ \{(H,T),(T,H),(H,H)\},\ \{(H,H),(T,T),(T,H)\},$
$\{(H,H),(T,T),(TH,T)\}\}$

Exercise: Consider $S = \{\{H,H\},\{H,T\},\{T,H\},\{T,T\}\}$, the repeated coin toss. Are the following valid Sigma-algebras for S:

1. $\Omega = \{\ \emptyset, S, \{(H,H)\}, \{(T,T)\}, \{(H,T),(T,H),(T,T)\}, \{(H,H),(H,T),T,H)\},$
   $\{(H,H),(T,T)\}\ \}$
2. $\Omega = \{\ \emptyset, S, \{(H,H)\}, \{(T,T)\}, \{(H,T),(T,H),(T,T)\},$
   $\{(H,H),(H,T),(T,H)\}, \{(H,H),(T,T)\}\{(H,T),(T,H)\}\ \}$
3. $\Omega = \{\ \emptyset, S, \{(H,T),(H,H)\}, \{(T,H),(T,T)\}\ \}$

2.1.6 Sigma algebra generated by family of subsets $C \subseteq P(S)$

For finite, or countably infinite, sample spaces we can always use their power set as a Sigma-algebra of possible events. But even in the simple examples above the power set can get very big (infact its number of elements is 2^n for n

elements in S). However, in the rainfall example, the power set is much larger. In fact, for any S with an uncountably infinite number of elements, the power set becomes so big that it is inconsistent with non-zero probabilities of its elements. This leads to inconsistencies when we try to define the probability of events.
So we need to find smaller Sigma-Algebras, but want these to include a certain class of subsets of S, call it "C". The smallest Sigma-Algebra of S that contains all elements of C is called the "Sigma-algebra generated by C".

Proposition: For $\Im_\alpha$, a collection of Sigma-algebras of a set S, $\bigcap_\alpha \Im_\alpha$ is also a Sigma-algebra of S.
Proof: We need to show that the empty set and S, as well as all complements and unions of any element $A \in \bigcap_\alpha \Im_\alpha$ are also in this intersection.
First, all $\Im \in \Im_\alpha$ contain $\emptyset$, and S, so $\emptyset, S \in \bigcap_\alpha \Im_\alpha$. Also, for any $A \in \bigcap_\alpha \Im_\alpha$ all $\Im \in \Im_\alpha$ contain A. As they are all Sigma-algebras, they all contain A^c, so $A^c \in \bigcap_\alpha \Im_\alpha$. The same logic applies to unions of sets $A_1, A_2, ... \in \bigcap_\alpha \Im_\alpha$.

With this property of Sigma-algebras in hand, we can define the Sigma-algebra generated by C as follows.
Definition: For C a collection of subsets of S, $\Im(C) \doteq \bigcap_{\Im \supseteq C} \Im$, the intersection of all Sigma algebras containing C, is called "Sigma-algebra generated by C".

Example: Consider $S = \{(H,H),(H,T),(T,H),(T,T)\}$, the repeated coin toss. The Sigma-Algebra generated by $C = \{\{(HH),(TT)\}\}$ is
$\Im(C) = \{\ \emptyset, S, \{(H,T),T,H)\}, \{(H,H),(T,T)\}\ \}$
Exercise: Consider $S = \{(H,H),(H,T),T,H),(T,T)\}$, the repeated coin toss. What is the Sigma-Algebra generated by $C = \{\{(HH)\},\{(TT)\}\}$?

2.1.7 Borel algebra

With uncountably infinite sample spaces the power set is not a useful Sigma-algebra. Here we use the concept of a Sigma-algebra generated by a family of subsets of S to define a useful Sigma-algebra for the n-dimensional space of real numbers.

Definition: For $S = \mathbb{R}^n$ the n dimensional Euclidian Space, the Borel Algebra $\mathbb{B}^n$ is defined as the smallest Sigma-algebra containing all open sets in

$\mathbb{R}^n$. Moreover, any $B \in \mathbb{B}^n$ is a "Borel set".

Proposition: $\mathbb{B}$, the Borel Algebra for the one-dimensional Euclidian space contains

- all open intervals $(-\infty, b)$,(a, ∞),(a, b),$(-\infty, \infty)$ (by definition of $\mathbb{B}$)
- all closed and half-closed intervals $(-\infty, b]$,$[a, \infty)$,$[a, b]$, *etc.* (by complementation and intersection of open sets)
- $\mathbb{R}$ (by countably infinite union of open sets)
- $\emptyset$ (by complementation of $\mathbb{R}$)

Proposition: $\Im(open) = \Im(closed) = \Im(compact) = \mathbb{B}(\mathbb{R})$ I.e. the Sigma-Algebra generated by open sets in $\mathbb{R}^n$ is the same as that generated by closed and compact sets.

Proof:

1. Any open set is the complement to a closed set, so in $\Im(closed)$. Since $\Im(open)$ is the smallest Sigma-algebra containing all open sets, we have $\Im(closed) \supseteq \Im(open)$. However, all closed sets are complements to open sets so in $\Im(open)$. Thus, by the same logic $\Im(open) \supseteq \Im(closed)$, and $\Im(open) = \Im(closed)$ follows.

2. Compact sets are closed, so $\Im(compact) \subseteq \Im(closed)$. But all closed sets F can be written as $F = \bigcup_{n=1}^{\infty}([-n, n] \cap F)$. As every [-n,n] is bounded, so is $([-n, n] \cap F)$. Thus F is a countable union of compact sets and therefore in $\Im(compact)$, so $\Im(closed) \subseteq \Im(compact)$, and equality follows.

Proposition: The Borel Algebra of the one-dimensional Euclidian Space $\mathbb{B}(\mathbb{R})$ can be generated by any of the families C_i of subsets of $\mathbb{R}$ defined by the following intervals, where $a, b \in \mathbb{R},\ a < b$:

1. (a, b)
2. $(-\infty, a)$
3. (a, ∞)
4. $[a, b]$
5. $(-\infty, a]$

6. $[a, \infty)$
7. $(a, b]$
8. $[a, b)$
9. *any closed subset of* $\mathbb{R}$

Proof Sketch (for details, see Wilde chapter 1): First note that all 9 kinds of intervals are in $\mathbb{B}(\mathbb{R})$, thus $\Im(C_i) \subseteq \mathbb{B}(\mathbb{R})$. Second note that any open interval of the form (a,b) can be constructed by complementation and countable union from intervals of any of the other 8 kinds, so $(a, b) \in \Im(C_i), i = 2, ..., 9$ and thus $\Im(C_1) \subseteq \Im(C_i), i = 2, ..., 9$. The proof proceeds by showing that any compact set is equal to the countable intersection of the union of certain open sets of the form (a,b), i.e.
$(K = \bigcap_{n=1}^{\infty} \bigcup_{x \in K} (x - 1/n, x + 1/n$ for all compact sets K. Thus all compact sets are in $\Im(C_1)$. Given $\Im(compact) = \mathbb{B}(\mathbb{R})$ as proven before, we get $\Im(compact) = \mathbb{B}(\mathbb{R}) \subseteq \Im(C_1) \subseteq \Im(C_i)$. This establishes the equality.

Intuition: This proposition is important. It means that when dealing with Borel subsets of $\mathbb{R}$ as long as we can be sure that what we do also applies to complements and unions of the sets we deal with, we can limit ourselves to any of the 9 intervals above. This feature can be easily generalised to more dimensional spaces. It will be important when we analyse probability functions for Euclidian sample spaces (e.g. $S = \mathbb{R}$), as then any probability function defined on one family of "interval-subsets" will be defined on the others as well, which makes our task much easier.

2.2 Additional exercises

- Banerjee Exercise sheet 1, exercise 5

2.3 Probability functions as measures

In the previous section we defined a consistent set of events that we can attach probabilities to. This section gives the axiomatic definition of probability.

2.3.1 Probability of an event A

Our intuition tells us that the concept "probability" should at least have the following features

1. It is defined for all elements of a consistent set of events (i.e. for all elements in a Sigma-algebra of the set of possible outcomes S).

2. The probability of every event is greater or equal to 0 and less or equal to 1.

3. For any two events A and B that do not share any outcomes (rainfall tomorrow vs. a dry day), we want the probability of either of the two occuring to be the sum of their individual probabilities.

4. The probability that anything occurs, or the probability of the sample space S is 1 (i.e. there is always some outcome). Together with 2, this implies $P(A) + P(A^c) = 1$.

The axiomatic definition of probability formalises these features.

Definition: Probability set function
"Probability of A" is a set function $P(\cdot) : \Im \longrightarrow [0, 1]$ s.t.

- $P(A) \geq 0, \ \forall A \in \Im$
- $P(S) = 1$
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for all sequences of disjoint events $\{A_i\}$, i.e. $A_i \cap A_j = \emptyset, \forall i \neq j$

Note: By the definition of the probability set function we can write the probability of any event as the sum of the probabilities of its elementary events, as these are by definition disjoint.

2.3.2 Measure μ on $\Im$

$P(\cdot)$ is a special case of a "measure", defined on the set of all possible events Ω. The more general concept "measure" assigns a "size" to subsets of a set S in an internally consistent way, for example as "length", "weight", "volume", etc. Or more formally:

Definition: For a given measurable space $(S, \Im)$, a measure $\mu(\cdot)$ is an extended real-valued set function $\mu : \Im \longrightarrow \mathbb{R}^+ \cup \infty$, s.t. if $\{A_n\}_{n=1}^{\infty}$ is a countable, disjoint sequence of subsets in $\Im$, then $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ ("countable additivity" or "σ additivity").

- The Triple $(S, \Im, \mu)$ is called a "**measure space**".

- If $\mu(A)$ is finite for all $A \in \Im$, then μ is called a "**finite measure**".

- If $\mu(S) = 1$, then $\mu(\cdot)$ is called a "**probability measure**", and $(S, \Im, \mu)$ is called a "**probability space**".

2.3.3 Proposition: Properties of finite measures

1. $\mu(\emptyset) = 0$ (by noting that the union of empty sets is empty, i.e. $\emptyset \cup \emptyset \cup ... = \emptyset$, so $\mu(\emptyset \cup \emptyset \cup ...) = \sum_{i=1}^{\infty} \mu(\emptyset) = \mu(\emptyset)$, which only holds if $\mu(\emptyset) = 0$)
2. $\mu(A) \geq 0, \ \forall A \in \Im$ (by the definiton of the range of μ).
3. $\mu(A_1 \cup A_2 \cup .... \cup A_n) = \sum_{i=1}^{n} \mu(A_i)$ for any finite sequence of disjoint subsets $\{A_i\}$ of $\Im$ (by setting $A_{n+1}, A_{n+2}, ... = \emptyset$)
4. $\mu(A) \leq \mu(B)$ *if* $A \subseteq B$ *and* $A, B \in \Im$ (by noting that $\mu(B) = \mu(A \cup B \setminus A) = \mu(A) + \mu(B \setminus A)$ *and* $\mu(B \setminus A) \geq 0$)

Example: "Length" L defined on $\mathbb{B}(\mathbb{R})$ as

- $L(A) = a - b$ for finite unions and complements of all open and closed intervals in $\mathbb{R}$, i.e. $A = (a, b), [a, b]$, etc., with $a \geq b$
- L(A) $= \infty$ for $A = (\infty, a), [a, \infty)$, etc.
- $L(\emptyset) = 0$
- $L(\bigcup_{i=1}^{N}(a_i, b_i)) = \sum_{i=1}^{N}(b_i - a_i)$, for all disjoint intervals (a_i, b_i)

is a measure on $\mathbb{B}(\mathbb{R})$.
Similarly, area, volume, etc. are measures on higher dimensional Borel sets.

Exercise Consider the set of "stones on a beach", and as Sigma algebra $\Im$ its power set. For $A \in \Im$, are the following valid measures on $\Im$: Weight, Number of stones in A. Length of stones in A put in a row. Volume of stones in a (gigantic) measuring beaker?

2.3.4 Measures on countable measure spaces

Consider a countable set $S = \{s_1, s_2, ..., s_n\}$ (i.e. a set with a finite or countably infinite number of elements). We can define a finite measure on its power set P(S) using *any* sequence of non-negative numbers $\{p_i\}$ with $\sum_i p_i$ finite, as $\mu(A) = \sum_{i \in I_A} p_i,\ for A \in P(S)\ and\ I_A = \{i : s_i \in A\}$.

2.3.5 Properties of probability functions

The properties of probability functions follow from the fact that a probability set function is a special case of a measure. Another way of easily visualising them is by using a **Venn diagram**. But note that this is not a prove. To prove the following properties, we would have to show that they are implied by the 3 assumed properties of probabilities. E.g.

- $P(A^c) = 1 - P(A)$
  Proof: Since $A \cup A^c$ is the union of disjoint sets, we have $P(A \cup A^c) = P(A) + P(A^c)$. Also, $A \cup A^c = S$, and P(S)=1 by the properties of probability measures. The rest follows. The other proofs are left to the reader.
- $P(\emptyset) = 0$
- $P(A) \leq 1$
- $P(B \cap A^c) = P(B) - P(B \cap A)$
- For $A_1 \subseteq A_2$, $A_1, A_2 \in \Omega$, $P(A_1) \leq P(A_2)$
- $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
- If $\{A_i\}_{i=1}^N$ is a monotone sequence of events in Ω, then $P(lim(\{A_i\})) = lim(P(A_i))$.

Exercise: Prove

- $P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1$ ("Bonferroni's inequality")
- $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$ ("Boole's inequality)

2.4 Additional exercises

- Banerjee Exercise sheet 1, exercise 3

3 Conditional Probability, Independence and Combinatorics

3.1 Conditional Probability

What is the probability of sunshine tomorrow once we know that it is not going to rain?
Often, we want to know the probability of an event A, knowing that some other event B has been observed. Thus, we want to calculate the "conditional probability" of "A given B", written $P(A|B)$. For this we need to know the "joint probability" of A and B, the unconditional probability that both events A and B occur together. The conditional probability is then simply the joint probability "scaled" by the probability of B occuring.
In other words, the unconditional probability of A is its probability size relative to that of the whole sample space, while the conditional probability of A given B is the probability size of "A and B" relative to that of B. This is most obvious by drawing a **Venn diagram**.

Definition: More formally, if for a probability space $(S, \Im, P())$ $A, B \in \Im$ and $P(B) \geq 0$, the conditional probability of event A, given event B, is $P(A|B) = P(A \cap B)/P(B)$.
This implies immediately the **"Law of Multiplication"** $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$ which allows to compute $P(A \cap B)$ e.g. if $P(A|B)$ and $P(B)$ are easier to calculate.

Proposition: $P(\cdot|B)$ is a probability set function where we replace S with B, and $\Im$ with $\Im_B$, the Sigma-algebra generated by $\{A_i \cap B : \ A_i \in \Im\}$. That is:

1. $P(A|B) \geq 0$ for any $A \in \Im$
2. $P(B|B) = 1$
3. $P(\bigcup_i A_i|B) = \sum_i P(A_i|B)$ for any sequence of disjoint events $A_i \in \Im$

Proof:
For all $A \in \Im$

1. $P(A \cap B), P(B) \geq 0$, so $P(A \cap B)/P(B) \geq 0$.
2. $P(B \cap B) = P(B)$, so $P(B \cap B)/P(B) = 1$.

3. For any sequence of disjoint events $A_1, A_2, ...$ $(A_1 \cap B)$, $(A_2 \cap B)$,... are also disjoint events. So $P(A_1 \cap B \cup A_2 \cap B) = P(\ (A_1 \cap B)\ ) + P(\ (A_2 \cap B)\ )$ from the definition of probability. The rest follows immediately.

3.1.1 Law of total probability

The definition of conditional probability implies that for a sequence $\{C_i\}_{i=1}^N$ of mutually exclusive and collectively exhaustive events (a partition of the sample space), i.e. $C_i \cap C_j = \emptyset, \forall\ j \neq i$ and $\bigcup_{i=1}^N C_i = S$, the probability of any event C is given by
$P(C) = \sum_{i=1}^N P(C|C_i) \cdot P(C_i)$. (Exercise: show this formally.)

Note: This implies

1. If C comprises K mutually exclusive and collectively exhaustive events, e.g. K different outcomes, that have probability $P(C_i)$, $P(C) = \sum_{i=1}^K 1 \cdot P(C_i)$, as the conditional probability is 1 for $C_i, i = 1, 2, ...K$ and 0 otherwise.

2. If all C_i are "equi-likely" with probability p=1/N, e.g. elementary events, $P(C) = \sum_{i=1}^K 1 \cdot p = K/N$. This means all we have to do is to count the C_is to get K, and the elements of the sample space to get N. There are rules how to do this, summarised under the heading of "combinatorics" (see section 2.4).

From the law of total probability, the definition of conditional probability, and the law or multiplication, we can immediately deduce Bayes' Rule:

3.1.2 Bayes' Rule

For any set B and any partition $A_1, A_2, ...$ of the sample space S,
$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^\infty P(B|A_j) \cdot P(A_j)}$.

Example (Champagne 2003) An urn contains 10 white balls and 5 black balls. We draw two balls from the urn at random, without replacement. Given the second ball is white, what is the probability that the first one was also white?

Solution: Define events W_1="First ball is white", B_1="First ball is black", W_2="Second ball is white", and use Bayes'law noting that W_1 and B_1 partition the sample space. Thus
$P(W_1|W_2) = \frac{P(W_2|W_1)P(W_1)}{P(W_2|W_1)P(W_1)+P(W_2|B_1)P(B_1)}$

$= \frac{9/14*2/3}{9/14*2/3+10/14*1/3} = 9/14.$

Exercise (Champagne 2003) A car rental agency has a fleet of 1000 Ford vehicules: 400 Escorts, 400 Taurus and 200 Explorers. These are equipped with either Firestone or Goodyear tires in the following proportions:

	Firestone	Goodyear
Escort	35%	65%
Taurus	55 %	45 %
Explorer	40 %	60 %

A customer selects a car at random: given that the car is equipped with Firestone tires, what is the probability that it is an Explorer?

3.2 Independence

If the occurrence of event B does not cause us to revise our probabilistic assessment of event A occurring, we say that A and B are independent.

Definition: More formally, two events A and B are "statistically independent", if $P(A \cap B) = P(A)P(B)$, implying $P(A|B) = P(A)$.
A collection of events $A_1, ..., A_n$ is "mutually independent" when for any of its subsets,
$P(\bigcap_{i \in I_A} A_i) = \prod_{i \in I_A} P(A_i)$ for $I_A \subseteq \{1, ..., n\}$.

Example: The probability to get (H) in second toss is 0.5, no matter the outcome of first toss.

Exercise: Show that if A and B are two independent events, then so are A and B^c, A^c and B, A^c and B^c.
Hint: Use the law of total probability and the fact the (B, B^c) is a partition of the sample space.

3.3 Combinatorics

We have seen that for an event C comprised of K out of N equilikely, mutually exclusive and collectively exhaustive events $\{C_i\}$, the probability of C is K/N. Particularly, for any event C that consists of equilikely elementary outcomes, we can define $P(C) = \frac{Number\ of\ "good"\ elementary\ outcomes\ where\ C\ occurs}{Number\ of\ all\ elementary\ outcomes\ N}$. To determine P(C), all we need are rules to determine the number of "good" outcomes, and the total number of outcomes N.

The most general way of describing any situation with equilikely outcomes is as a *draw from an urn with balls numbered from 1 to N*. This includes a coin toss, a dice throw, a lottery, etc.
Non-repeated experiments are generally easy - the probability of drawing a number smaller or equal to K is K/N. But it becomes more difficult with "selections" of balls, i.e. repeated draws from the urn.

Example What is the probability of 2 aces when drawing twice from a deck of 52 cards?
Let us first determine N, the number of possible pairs of cards drawn from a poker deck, noting that all cards are different (by kind or by colour). So there are 52 possibilities in the first draw, and 51 the second, yielding 2652. But for the pair (king of spades, 10 of clubs) we don't distinguish which was drawn first. So we divide this number by 2 to get N=1326. To get the number of "good" outcomes, where there are actually two aces, we need to count the different ways this can happen. As there are 4 aces in total, we can draw 4*3/2=6 different pairs, where again we divide by two to correct for the fact that drawing the ace of spaces first or second is the same thing in our example.
So the probability is 6/1326, or less then half a percent.
Note that this example is equivalent to drawing 2 balls from an urn of 52, where balls have colours and numbers, and the draw does not take note of the order in which the ball are taken out of the urn.

Counting rules
More generally, when counting the number of selections from a set (draws from the balls in an urn), we need to make 2 distinctions:

1. ordered (i.e. it matters at which draw we get a particular ball) vs. unordered

2. and with vs. without replacement

Definition Define $n! = n(n-1)(n-2)...1$ and $0! = 1$
When choosing r objects from a set A of n objects, we get the following numbers of selections

- Ordered, with replacement: n^r
- Ordered, without replacement (also called "r-element permutation of A"): P(n,r)=$n!/(n-r)!$
- Unordered, with replacement: $(n+r-1)!/(r! * (n-1)!)$
- Unordered, without replacement (also called "r-element combination of A"): C(n,r)=$n!/(r! * (n-r)!) =$ "$noverr$"

Exercise 1: What is the probability of winning the first prize of a lottery of 6 from 49 numbers, i.e. getting all 6 numbers right, where the order of the draws does not matter?
Exercise 2 What is the probability of winning the second prize of a lottery of 6 from 49 numbers, i.e. of getting all but 1 number correct?

3.4 Additional exercises

- Banerjee Exercise sheet 1, exercise 1 and 2

4 Univariate Random Variables

So far, we have looked at probabilities of arbitrary events, or at probability measures of subsets of an arbitrary sample space S. In other words, $P(\cdot)$ is a set function from a Sigma-algebra Ω to [0,1].
But there are cases, where we are much more interested in some number that is attached to every outcome of the experiment than the outcome itself. For example, I might be more interested in the probability of future salaries, then that of individual situations - state of the labour market, performance of my firm relative to others, etc. - that determine it.

Also, the sample space S, and its Sigma-Algebra Ω, the domain of $P(\cdot)$, are different for every random experiment. Often, one has to tabulate all elementary outcomes, and events together with their probabilities, which can be cumbersome, as our simple toin coss example has shown.

In other words, we would like to introduce as a more standard domain for probability functions the set of real numbers, because often that is what we are interested in anyway, and it makes things a lot easier. So we are looking for functions that replace the original sample space with real numbers.

This section first discusses the necessary properties of functions, or "random variables" that achieve this replacement of outcomes by real numbers while maintaining the probability structure of the random experiment. We then look at ways to summarise the relative probabilities of the values that this function takes in a "probability distribution function".

4.1 Definition: Random Variable

Suppose we have a mapping $X : S \longrightarrow \mathbb{R}$ that assigns a real number to every elementary event in the sample space. Often there is a natural candidate for X, such as "the sum of numbers on a dice", or "the number of heads". But in order to assign probabilities to numbers that X may take, or subsets of the real line, the mapping X has to preserve the event and probability structure of the measurable space (S, Ω). This requires that there is a well-defined event E in Ω that corresponds to any subset M of the range of X, as well as their unions, intersections and complements. The probability of M is then simply the probability of E. If X has this property, we call it "random variable".

Definition: For a given probability space $(S, \Omega, P(\cdot))$ a Random Variable (RV) is a real-valued function X from S to $\mathbb{R}$, which satisfies the condition that for every half-closed interval $I_x = (-\infty, x], x \in \mathbb{R}$, the inverse image $X^{-1}(I_x) \doteq \{s \in S : X(s) \leq x\}$ is an event in Ω. In other words, every half-closed interval $I_x = (-\infty, x], x \in \mathbb{R}$ has a corresponding subset of S in Ω, given by the set of elements in S that X maps into I_x.

Note:To check that X is a valid random variable, we have to show that for every half-closed interval I in $\mathbb{R}$, there is an event in Ω that X maps into I.

Example: Is $X \doteq \ Number\ of\ Heads$, a valid random variable for S = {H,T}, a single coin toss and the power set of S as its Sigma-algebra Ω?
Answer:

- X(H) = 1; X(T)=0
- So $X^{-1}(1) = H;\ X^{-1}(0) = T;\ X^{-1}(a) \ = \ \emptyset,\ for\ all\ a \notin \{1, 0\}.$
- So for every $B = (-\infty, a]$ we have
  - If $a \leq 0,\ X^{-1}(B) = \emptyset \in \Omega.$
  - If $0 \leq a < 1,\ X^{-1}(B) = \{T\} \in \Omega.$
  - If $1 \leq a,\ X^{-1}(B) = \{T, H\} \in \Omega.$
- So X is a random variable.

Sometimes you rind a different definition of a random variable, on the basis of all Borel sets (e.g. Spanos 1986), not just the half-closed intervals. The following proposition shows that these definitions are actually equivalent.
Proposition $X : S \longrightarrow \mathbb{R}$ is a random variable on $(S, \Omega, P(\cdot))$, if and only if $X^{-1}(B) = \{s \in S : X(s) \in B\}$ is an event in Ω for all Borel sets $B \in \mathbb{B}$.

Proof:

- **IF**: The half-closed intervals are Borel-sets, i.e. $I_x = (-\infty, x] \in \mathbb{B}, \forall x \in \mathbb{R}$. So if $X^{-1}(B) \in \Omega, \forall B \in \mathbb{B}$ then $X^{-1}(I_x) \in \Omega, \forall I_x \subseteq \mathbb{R}$.
- **ONLY IF**: Suppose $X^{-1}(I_x) \in \Omega, \forall I_x \subseteq \mathbb{R}$. Denote the set of all subsets of $\mathbb{R}$ that have a corresponding event, or inverse image in Ω, $\mathbb{C} = \{E \subseteq \mathbb{R} : X^{-1}(E) \in \Omega\}$. This is a Sigma-Algebra, as
  - $\mathbb{R} \in \mathbb{C}$, as $\{s \in S : X(s) \in \mathbb{R}\} = S \in \Omega$.

- Equally $\{s \in S : X(s) \in \emptyset\} = \emptyset \in \Omega$.
- $(X^{-1}(E^c) = \{s \in S : X(s) \in E^c\} = (X^{-1}(E))^c \in \Omega$, so $E^c \in \mathbb{C}$ also. In other words, given that $X(\cdot)$ is a function, the inverse image of the complement of a set E is the complement of the inverse image of E, which are both in Ω. So both E and E^c are in $\mathbb{C}$, which is thus closed under complementation.
- $X^{-1}(E_1 \cup E_2 \cup ...) = \{s \in S : X(s) \in (E_1 \cup E_2 \cup ...)\} = \{s \in S : X(s) \in (E_1)\} \cup \{s \in S : X(s) \in (E_2)\} \cup ....$ In other words the inverse image of a union of Es is the union of the inverse images, which is in Ω. So the union of Es is in $\mathbb{C}$, which is thus closed under countable union.
- Since $X^{-1}(I_x) \in \Omega, \forall I_x \subseteq \mathbb{R}$, we have $I_x \in \mathbb{C}$. Thus, given $\mathbb{C}$ is a Sigma-algebra, $\Im(I_x) = \mathbb{B}(\mathbb{R}) \subseteq \mathbb{C}$, which says that all $X^{-1}(B) \in \Omega, \forall\ B \in \mathbb{B}(\mathbb{R})$.

Note:

- A random variable is always defined with respect to some Sigma-Algebra Ω.
- Distinguish the random variable X from x, the value it takes in a particular trial of a random experiment .
- To decide whether $X(\cdot) : S \longrightarrow \mathbb{R}$ is a random variable, one needs to proceed from the half-closed intervals in $\mathbb{R}$ to the elements of Ω, the Sigma-Algebra of S, not the other way.
- A random variable is a function, neither "random", nor "variable".

Exercise: Consider S = {(H,H),(H,T),(T,H),(T,T)}, the repeated coin toss, and the two Sigma-algebras

1. $\Omega = \{\ \emptyset, S, \{(H,H)\}, \{(T,T)\}, \{(H,H),(T,H),(H,T)\},$
   $\{(T,H),(H,T),(T,T)\}, \{(H,H),(T,T)\}\{(H,T),(T,H)\}\ \}$
2. $\Omega = \{\ \emptyset, S, \{(H,T),(H,H)\}, \{(T,H),(T,T)\}\ \}$

- Take the Random Variable X = "number of heads". Write down the sets $\{s \in \Omega : X(s) = a\}$ where $a \in \{0,1,2\}$.
- Now consider the half-open intervals defined by $(-\infty, a]$, $a \in \mathbb{R}$. Write down the sets $\{s \in \Omega : X(s) \leq a\}$ where $a \in \{0,1,2\}$.
- Using this, show that X is a random variable with respect to the first but not the second Sigma-Algebra.

- Consider the random variable Y defined by $Y(\{H,H\}) = Y(\{H,T\}) = 1, Y(\{T,T\}) = Y(\{T,H\}) = 0$. With respect to which of the two Sigma-Algebras is Y a random variable?

4.2 Measurable functions

This section shows that random variables are simply a special example of measurable functions.

Definition: Given a measurable space $(S, \Im)$ a real-valued function $g(\cdot) : S \longrightarrow \mathbb{R}$ is "Borel measurable with respect to $\Im$" if for all open sets $A \subseteq \mathbb{R}$ the sets $\{s \in S : g(s) \in A\}$ are in $\Im$.

4.2.1 Random variables are measurable functions on a probability space

To show that a RV is nothing but a measurable function on a probability space, we only have to show that measurable functions associate not only open sets, but all Borel sets to events in $\Im$. The following proposition does this.

Proposition
$g(\cdot) : S \longrightarrow \mathbb{R}$ is a measurable function on $(S, \Im)$, if and only if $X^{-1}(B) \doteq \{s \in S : X(s) \in B\}$ is in $\Im$ for all Borel sets $B \in \mathbb{B}$. That is a measurable function gives for every Borel set an element of the Sigma-Algebra of its domain.

Proof:
The proof is equivalent to that of the preceding proposition.

- **IF**: The open sets $\{I^o\}$ are Borel-sets, i.e. $I^o \in \mathbb{B}, \forall I^o \in \mathbb{R}$. So if $X^{-1}(B) \in \Im, \forall B \in \mathbb{B}$ then $X^{-1}(I^o) \in \Im, \forall I^o \subseteq \mathbb{R}$.
- **ONLY IF**: Suppose $X^{-1}(I^o) \in \Omega, \forall I^o \subseteq \mathbb{R}$. Denote the set $\mathbb{C} = \{E \subseteq \mathbb{R} : X^{-1}(E) \in \Im\}$. This is a Sigma-Algebra, as shown above. So given $X^{-1}(I^o) \in \Im, \forall I^o \subseteq \mathbb{R}$, we have $\Im(I_x) = \mathbb{B}(\mathbb{R}) \subseteq \mathbb{C}$, which says that all $X^{-1}(B) \in \Im, \forall\ B \in \mathbb{B}(\mathbb{R})$.

This establishes that a random variable is simply a measurable function on a probability space.

Note: The same logic applies to all other subsets of $\mathbb{R}$ that generate the Borel-Algebra (especially the 9 kinds of intervals mentioned in the section on Borel Algebras). So $g(\cdot) : S \longrightarrow \mathbb{R}$ is a measurable function on $(S, \Im, P(\cdot))$, if and only if $X^{-1}(C_i) \in \Im, \forall C_i \subseteq \mathbb{R}$, where C_i denotes any collection of subsets of $\mathbb{R}$

that generate the Borel-Algebra of the 9 intervals.

Exercise: For $S = \{0, 1\}$, consider $\Im_1 = \{\emptyset, \{1\}, \{0\}, S\}$ and $\Im_2 = \{\emptyset, S\}$. Show that all functions on S are $\Im_1$-measurable, but only constant functions are $\Im_2$-measurable.

4.2.2 Proposition: Properties of measurable functions

- All monotone functions $g(\cdot) : (a, b) \longrightarrow \mathbb{R}$ are measurable on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$.
  Proof: For increasing (decreasing) functions $x \leq x'$ implies $g(x) \leq (\geq) g(x')$, for all $x \in (a, b)$. So to every open interval $\{y \in Y : y < \overline{y}\}$ corresponds one of the following sets:
  - $\emptyset$ if $\overline{y} \leq g(a)$
  - an open interval $\{x \in X : a < x < g^{-1}(\overline{y})\}$ if $\overline{y} \leq g(b)$
  - $\{x \in X : a < x < b\}$ if $\overline{y} > g(b)$

  That is the inverse image of every open interval is a subset of $\mathbb{R}$, and so in $\mathbb{B}(\mathbb{R}$.

- All continuous functions $g(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$, are measurable on $(\mathbb{R}, \mathbb{B}(\mathbb{R})$.
  Proof: Inverse Images of open sets are open for continuous functions. I.e. $g^{-1}(A) = \textit{open}$, for all open sets $A \subseteq \mathbb{R}$. But all open sets are in the Borel-algebra, so for every open set in $\mathbb{R}$, there is a Borel set in $\mathbb{B}(\mathbb{R})$.

- For Borel-measurable functions $f : S \longrightarrow \mathbb{R}$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ the composite function $g \circ f : \S \longrightarrow \mathbb{R}$ is Borel-measurable on $(S, \Im)$.
  Proof: Let A be any open set. $g^{-1}(A) \in \mathbb{B}(\mathbb{R})$ by measurability. Also, $f^{-1}(B) \in \Im, \forall B \in \mathbb{B}(\mathbb{R})$ by measurability, so $f^{-1}(g^{-1}(A)) \in \Im$ for all open sets A.

- For $(S, \Im)$ a measurable space and $f(\cdot) : S \longrightarrow \mathbb{R}$, $g(\cdot) : S \longrightarrow \mathbb{R}$ two Borel-measurable functions, the following are Borel functions

  1. $af + b,\ a, b \in \mathbb{R}$
     Proof: For any $c \in \mathbb{R}$, the set $\{x \in S : af + b \leq c\}$ equals
     - $\{x \in S : f \leq \frac{c-b}{a}\}$ for a¿0
     - $\{x \in S : f \geq \frac{c-b}{a}\}$ for a¡0
     - X for a=0 and $c \geq b$, and
     - $\emptyset$ for a=0 and $c < b$.

     All these sets are in $\mathbb{B}(\mathbb{R})$, so $af + c$ is measurable.

2. $f + g$
3. $|f|^\alpha,\ \forall \alpha \geq 0$
4. if f never vanishes, 1/f
5. f*g
6. max{f,g}, min{f,g}

Remaining proofs: see Wilde, p. 6-8

Note: This proposition says that the space of measurable functions is closed under addition, multiplication, scalar multiplication, etc.

4.2.3 Properties of random variables

Again, properties of measurable functions translate of course to those of random variables, i.e.

- A measurable function $g : \mathbb{R} \longrightarrow \mathbb{R}$ of a random variable $X : S \longrightarrow \mathbb{R}$ is itself a random variable.
- The sum of n random variables is a random variable, so is their mean.
- etc.

4.3 Distribution functions of random variables

4.3.1 Defining probability measures for random variables

A Random variable X "replaces an arbitrary sample space S with $\mathbb{R}$" for a particular random experiment such that its event and probability structure is preserved. But we still need to assign probabilities to values of random variables in $\mathbb{R}$.
A natural candidate for this is of course $P(x) = P(s : X(s) = x)$ for all values $x \in \mathbb{R}$, i.e. the probability of any value x of a random variable is simply the probability of those outcomes that X maps into x. And this indeed works well for random experiments with a finite number of outcomes, where the function P(x) is called the "probability mass function". But for random experiments with continuous, or infinite sample spaces, this is more difficult: the probability of every individual number to occur may well be zero, so the sum of any number of these probabilities remains zero.
But note that by proposition **??**, for any random variable there is an event

associated with every Borel set B in $\mathbb{R}$. So we can define the "Probability of $X \in B$" to equal the probability of that event in Ω that corresponds to B according to the mapping X (i.e. that comprises all those elementary outcomes s that X maps into B). So we get a function $P_X(B) : \mathbb{B}(\mathbb{R}) \longrightarrow [0,1]$ defined by $P_X(B) = P_\Omega(X^{-1}(B))$. This function is sigma-additive and assigns a value $P \in [0,1]$ to all Borel sets, so $P(B) = P(\{s : X(s) \in B\}$ defines a probability measure on the Borel sets of $\mathbb{R}$.
However, there are very many of these sets as we saw before. So we would like to summarise their probabilities somehow. The next proposition shows that we can do that.

Proposition
For a RV $X : S \longrightarrow \mathbb{R}$ on a probability space $(S, \Omega, P(\cdot))$, the probability measure v on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ given by $v(B) = P(\{s \in S : X(s) \in B\}), \forall B \in \mathbb{B}$ is uniquely determined by the probabilities attached to any collection C_i of subsets of $\mathbb{R}$ that generates the Borel-algebra.
Proof: The first part of the proof is equivalent to proposition **??**, where we showed that the definition of a RV, to have inverse images of half-closed intervals in Ω, implies that the inverse images of all Borel-sets are events in Ω. This holds for all collection of subsets of $\mathbb{R}$ that generate the Borel algebra.
The rest is trivial: since the mapping $P(\cdot) : \Omega \longrightarrow [0,1]$ is unique for all $E \in \Omega$, there is a unique probability for all complements and unions of subsets $C \in C_i$, that is for all Borel-sets. So to completely characterise the probabilities assigned to values of a random variable in Borel subsets of $\mathbb{R}$, we only need to consider for example the probabilities of X lying in the half-closed intervals $(-\infty, x]$, i.e. $P_X(X \in (-\infty, x]) = P_\Omega(\{s \in S : X(s) \in (\infty, x]\})$.
This probability measure has already a far more convenient domain. But note that it is still a set function. However, the half open intervals are entirely characterized by their upper bound x. So we can simplify the probability function even further by defining the point function $F_X(x) \doteq Prob(X \leq x)$. This is the "Cumulative distribution function of X".

Note: Using the concept of a random variable we have gone from a probability measure defined on *subsets* of an arbitrary sample space S to a much simpler point function $F_X(\cdot) : \mathbb{R} \longrightarrow [0,1]$.

4.3.2 Cumulative Distribution function (CDF)

The function $F_X(x) \doteq Prob(X \leq x)$ is called the "Cumulative Distribution function of random variable X" or simply its "Distribution function".

Proposition: A function $F(\cdot)$ is a CDF **if and only if** it satisfies the following three properties:

1. $0 \leq F(x) \leq 1$
2. F(x) is nondecreasing in x
3. $lim_{x \longrightarrow -\infty} = 0$, $lim_{x \longrightarrow \infty} = 1$
4. F(x) is right-continuous

Intuition: The proposition states two things. First, all distribution functions satisfy the three given properties. But also, *any* function that satisfies the properties is the distribution for some random variable, i.e. for every function $F(\cdot)$ satisfying the four properties there is a probability space (S, Ω, P) and a RV $X : S \longrightarrow \mathbb{R}$, such that $F(\cdot)$ is its distribution.

Proof:

- **IF**
  Consider a probability space (S, Ω, P) and a random variable X.
  1. This follows immediately from the next two properties.
  2. Define the events $A = \{s \in S : X(s) \in (-\infty, x]\}$, and $B = \{s \in S : X(s) \in (-\infty, y]\}$, $x \leq y$, so $A \subseteq B$. This yields $P(A) = F(x) \leq P(B) = F(y)$.
  3. Note that for $A_n = (-\infty, n]$ $A_1 \subseteq A_2 \subseteq A_3 ...$ is an increasing sequence in $\mathbb{B}$ with $limA_n = \mathbb{R}$. So $\{E_i\}$, with $E_i = X^{-1}(A_i)$ is an increasing sequence in Ω with $limE_i = S$. Thus $lim_{x \longrightarrow \infty} F(x) = limP(E_i) = P(lim(E_i)) = 1$. Equivalently, noting that $\bigcap A_n = \emptyset$ for a decreasing sequence $\{A_n\}$ with $A_n = (-\infty, -n]$ and that $\{s \in S : X(s) \in \emptyset\} = \emptyset$, we get for $E_i = X^{-1}(A_i)$ $lim_{x \longrightarrow -\infty} F(x) = limP(E_i) = P(lim(E_i)) = P(\emptyset) = 0$.
  4. For any $y \in \mathbb{R}$ define a decreasing sequence as $\{B_n\}$, *with* $B_n = (-\infty, y + \frac{1}{n}]$. Note that $lim \bigcap B_n = (-\infty, y]$. Denote $E_i = X^{-1}(B_i)$ So $lim_{x \longrightarrow y^+} F(y) = limP(E_i) = P(lim(E_i)) = F(y)$.
- **ONLY IF**
  We want to show that for every function $F(\cdot)$ with the 4 properties there is a probability space (S, Ω, P) and a RV $X : S \longrightarrow \mathbb{R}$ such that $F(\cdot)$ is its distribution, i.e. $F(x) = P(\{s \in S : X(s) \in (-\infty, x]\})$ for some RV X and all $x \in \mathbb{R}$. Consider the probability space $(S, \Omega, P) = (\mathbb{R}, \mathbb{B}(\mathbb{R}), v)$, where v is defined as

1. $v(\ (-\infty, a)\ ) = F(a^-)$
2. $v(\ (a, b)\ ) = F(b^-) - F(a)$
3. $v(\ (a, b]\ ) = F(b) - F(a)$
4. $v(\ [a, b]\ ) = F(b) - F(a^-)$
5. $v(\ [a, b)\ ) = F(b^-) - F(a^-)$

with $F(a^-)$ the left limit of F at a.
For all $F*(\cdot)$ the measure on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ defined by 1 to 5 is called the "**Lebesgue-Stieltjes measure given by** $F(\cdot)$".
Now consider $S = \mathbb{R}$, $\Omega = \mathbb{B}(R)$ and $P = v(\cdot)$, where v is the Lebesgue-Stieltjes measure given by F, and take the random variable $X(s) = s$, for all $s \in S = \mathbb{R}$. The inverse image $X^{-1}(\cdot)$ maps all Borel sets in the range of X into Events in its domain, or elements of the Sigma-algebra of our probability space (which itself is the Borel Algebra). So it is a random variable. Its CDF is given by $F_f(x) = v(\ \{s \in \mathbb{R} : X(s) \in (-\infty, x]\}\ ) = v((-\infty, X^{inv}(x)]) = F(X^{inv}(x)) = F(x)$. So $F(\cdot)$ is indeed the CDF of the random variable $X(s) = s$ for the probability space $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ and the Lebesgue-Stieltjes measure given by $F(\cdot)$.

Note: The Lebesgue-Stieltjes measure with respect to the particular $F(\cdot)$
$F(x) = \frac{x-a}{b-a},\ \forall\ x \in [a, b], b \geq a$,
$F(x) = 0\ \forall x \leq a$,
$F(x) = b - a\ \forall x \geq b$,
is the "**Lebesgue measure**" on $[a, b]$. It weights intervals in [a,b] by their length.

Corollary: We have shown that every function $F(\cdot)$ with the Properties of a CDF is the distribution function of a random variable $X : \mathbb{R} \longrightarrow \mathbb{R}$ with $X(s) = s$ on $(\mathbb{R}, \mathbb{B}, v)$ where v is the Lebesgue-Stieltjes measure with respect to F. Thus, any random variable Y on an arbitrary probability space (S, Ω, P) induces a probability measure μ on $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ with $\mu = v$, the Lebesgue-Stieltjes measure with respect to the CDF of Y.

4.3.3 Discrete random variables and their distribution

Definition: A random variable X is called "discrete" if its CDF is a step function.
For X a discrete random variable, the "**probability mass function**" of X is $p_X(x) = P(X = x)$, with the following properties:

1. $p_X(x) > 0$ at all points of discontinuity of its CDF or on $\{X_i : F(x_i^-) - F(x_i) > 0\}$, $p_X(x) = 0$ otherwise

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

Note:

- Any random variable on a discrete (i.e. finite or countably infinite) sample space is discrete.
- More generally, any random variable with discrete range is discrete.
- The "support of X" D_X is defined for a discrete random variable as the set of "jumps", i.e. the points with positive mass, i.e. $D = \{x \in \mathbb{R} : P_X(x) > 0\}$.

Example The random variable "number of heads in the repeate coin toss" with the sigma algebra the Power set, has the following CDF and PMF.

(Insert picture.)

Exercise: (to be added)

4.3.4 Continuous random variables and their distribution

Definition: A random variable X is called "continuous" if its CDF is continuous for all $x \in \mathbb{R}$.

In this case, P(X=x)=0, for all $x \in \mathbb{R}$. We limit our attention to cases where $F(\cdot)$ is "absolutely continuous". So there is a "**probability density function**" (pdf) $f_X(x)$ that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(x)dx$, such that $f(x) = \frac{d}{dx}F_X(x)$, and $Prob(a \leq x \leq b) = F_X(b) - F_x(a) = \int_a^b f_X(x)dx$.

A function $f_X(x)$ is a pdf if and only if

1. $f_X(x) \geq 0$ for all $x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$

The "support of X" D is defined for a continuous random variable X as the set of points with positive density, i.e. $D = x \in \mathbb{R} : f_X(x) > 0$.

Example (to be added)

Exercise: (to be added)

Note:
If a random variable is neither discrete, nor continuous, it is called a "mixed random variable".

Graph of a mixed random variable

4.3.5 Summarizing distribution functions - location and spread

A first step in order to summarise the characteristics of a distribution of a random variable X, is to describe its *location* on the real line. Three different location measures are

- Its expectation, or mean, E(X) (see next section)
- The "middle observation" or "median" m_x defined by $\int_{-\infty}^{m_x} f_X(x)dx = 1/2$ for continuous variables
- The "mode" defined as the most frequent observation, or value with highest density

A second step is to describe the "*spread*" of X around its location measures. This can be done for example by its "Variance", or "Standard deviation".

Mean and Variance are example of "moments" of the random variable X, which we will define more closely in the next section.

4.4 Additional exercises

- Banerjee Exercise sheet 1, exercise 6, 7 and 9

4.5 Distributions of functions of random variables and the change of variables formula

Sometimes we know the probability density function of a random variable $X : S \longrightarrow \mathbb{R}$, but want to calculate the pdf of a new random variable $Y \in \mathbb{R}$ which is a function g of X, i.e. Y=g(x). If $g(\cdot)$ is one-to-one (i.e. it maps every point in the range of X into exactly one point in $\mathbb{R}$), differentiable and invertible, then the following procedure can be applied.

- If g(X) is an increasing function, then $g(X) \leq y$ if and only if $X \leq h(y)$, where $h(\cdot)$ is the inverse of g. So

$F_Y(y) = Prob(Y \leq y) = Prob(g(X) \leq y) = Prob(X \leq h(y)) = F_X(h(y)$.
We are looking for a probability density $f(y)$, such that $\int_{-\infty}^{y} f_Y(s)ds = F_X(h(y)$. By differentiating both sides with respect to y and applying Leibniz rule, we get
$f(y) = \frac{d}{dy}F_X(h(y)) = f_X(h(y))\frac{d}{dy}h(y)$.

- If g(X) is a decreasing function, then $g(X) \leq y$ if and only if $X \geq h(y)$. So $F_Y(y) = Prob(Y \leq y) = Prob(g(X) \leq y) = 1 - Prob(X \leq h(y)) = 1 - F_X(h(y)$, we get
$f(y) = -\frac{d}{dy}F_X(h(y) = -f_X(h(y)\frac{d}{dy}h(y)$.
Given that in the first case $\frac{d}{dy}h(y)$ is positive, in the second negative, we can combine the two cases as
$f(y) = f_X(h(y)|\frac{d}{dy}h(y)|$.
Note: Some care needs to be taken when transforming the domain of X D_X into the Domain of Y D_Y, e.g. when taking expectations.

Example:
Suppose X is distributed uniformly on $[0, 1]$, and define $Y = -ln(x)$ which maps values in $D_X = [0, 1]$ into values in the Domain of Y $D_Y = [0, \infty]$. Then we have $h(y) = e^{-y}$, and $| \frac{\delta}{\delta y}(h(y)) | = e^{-y}$. Note that $f_X(x) = 1$ for $x \in [0, 1]$, so $f_X(h(y)) = 1$ for $y \in [0, \infty]$, which yields $f_Y(y) = e^{-y}$, $y \in [0, \infty]$, an exponential density. **Exercise:** (to be added).

4.6 Additional exercises

- Banerjee Exercise sheet 3, exercise 2 and 3

5 Integration theory, mathematical expectation, and moments of random variables

5.1 Reader's digest integration theory

5.1.1 Integral

The integral of a real-valued function $f(\cdot)$ gives the area under its graph. It is calculated as the limit of a weighted sum of functional values, where the weights are measures of subsets of the Domain D associated to these functional values, written $\int_D f d\mu = lim \sum_{i=1}^n f_i d\mu_i$.

5.1.2 Riemann integral

The Riemann integral is defined for real-valued functions on a bounded interval (a, b). It is the usual integral encountered in high-school calculus, calculated with respect to the "Jordan measure". It is defined as the sum of the length (area,volume, ...) of subintervals (rectangles, boxes, ...) of the domain multiplied by an arbitrary function value on that subinterval, when the width of these intervals goes to zero. Or formally,
$\int_D f d\mu = lim_{max\Delta x_i \longrightarrow 0} \sum_{i=1}^n f_i^* \Delta x_i$, where f_i^* is an arbitrary functional value on the ith interval. $f(\cdot)$ is called "Riemann integrable" if the limit converges, i.e. if
$lim_{max\Delta x_i \longrightarrow 0} \sum_{i=1}^n sup\{f(x) : x \in [x_i, x_i + \Delta x_i]\}^* \Delta x_i$
$= \ lim_{max\Delta x_i \longrightarrow 0} \sum_{i=1}^n inf\{f(x) : x \in [x_i, x_i + \Delta x_i]\} \cdot \Delta x_i$
$\leq \ \infty$.

5.1.3 (Lebesgue) integral

A more general integral with respect to any measure μ is defined using the concept of "simple functions" and their integrals.

Definition A "simple function" $s(\cdot) : S \longrightarrow \mathbb{R}$ is a function that takes on a finite number of values, i.e. $s = \sum_{i=1}^n \alpha_i * 1_{A_i}$, for 1_{A_i} the indicator function that takes value 1 on A_i and 0 otherwise, for a sequence of disjoint sets $\{A_i\} \in S$.

Definition For an arbitrary finite measure space $(S, \Im, \mu)$ the "integral of a simple function $s(\cdot) : S \longrightarrow \mathbb{R}$ over $E \in \Im$ with respect to measure μ is defined as
$\int_E s d\mu = \sum_{i=1}^n \alpha_i * \mu(A_i \cap E)$.

That is the integral is simply the function value times the measure of the set of points where it applies.

Using these two definitions, the integral of a non-negative function $f(\cdot)$ is calculated as the supremum of the integrals over simple functions under $f(\cdot)$. Or formally,
$\int_D f d\mu = sup(\int_D s d\mu,\ 0 \leq s(\cdot) \leq f(\cdot)$.
$f(\cdot)$ is said to be "integrable" if the value on the rhs is finite.
For general real-valued functions, integrate separately the positive and negative parts and calculate the sum
$\int_D f d\mu = \int_D f^+ d\mu - \int_D f^- d\mu$
This is also called "Lebesgue integral".

Note: From the above it follows that a function is "**integrable** with respect to μ" if it is measurable with respect to μ and has a finite integral.

Note:

- The Riemann integral takes an ever finer grid on the "x-axis" to calculate the limit of the sum. The Lebesgue integral takes an ever finer grid on the "y-axis".
- The Lebesgue integral is defined for a wider class of functions than the Riemann integral.
- For bounded real-valued functions Riemann integrability implies Lebesgue integrability. Also, the value of the integral for Riemann integrable functions is the same as the value of the Lebesgue integral.

Example: The Dirichlet function is 0 where its argument is irrational and 1 otherwise. It has no Riemann-integral on $[0, 1]$, as for any subinterval there is always at least one irrational and one rational number in it, so the two sums in the definition do not converge in the limit. However, from the definition of the Lebesgue integral, denoting 1_Q as the indicator function of the rationals and μ as Lebesgue measure, we see that $\int_D f d\mu = sup(\int_D 1_Q d\mu) = 1 * \mu(Q \cap [0, 1]) = 0$, since Q is a countable set.

Intuition: An intuitive way of thinking about these integrals is the following. Suppose you want to calculate the volume of a mountain. The Riemann integral chops the mountain into towers of rectangular base. It then calculates

the volumes of 2 kinds of rectangular blocks: of those that just fit into the towers, and of those that the tower just fits into. It adds these separately, and if, moving to ever smaller bases of towers, the two sums converge, we have the mountain's volume.
The Lebesgue integral looks at the map of the mountain, and measures the area between its contours. It then multiplies this measure by the height of the lower contour for every contour and adds them up. The limit for ever finer contours is the volume of the mountain.

5.2 Mathematical expectation

The expectation of a random variable on a probability space is an average of its functional values, weighted by the probability of the events associated to a particular value.

Definition The expectation of a random variable X(s) on a probability space (S, Ω, P) is defined as

$E(X) = \int_S X(s)dP,$

the "(Lebesgue) integral of X over the sample space S with respect to the probability measure P". X is said to have finite expectation if the integral exists.
We have seen that measurable functions $g(\cdot)$ of random variables are a random variable. So more generally,$E(g(X)) = \int_S g(X(s))dP$, which weights functional values g(X) by the probability of events in Ω that correspond to the Borel sets in the Domain of g.

Can we simplify this expression somehow? We have seen that a random variable X on an arbitrary probability space always induces a new probability measure $v(\cdot)$ on $(\mathbb{R}, \mathbb{B})$, with v equal to the Lebesgue-Stieltjes measure with respect to F, the CDF of X. So weighting the distinct values of g(X(s)) by the probability of the sets of s's that are in the inverse image of distinct Xs, is intuitively the same as weighting them by the measure given to values of X by the Lebesgue-Stieltjes measure. This is formalised in the following proposition.

5.2.1 Proposition

Suppose that X has finite expectation. Then $E(X) = \int_S X dP = \int_{\mathbb{R}} x dF_X$, where F_X denotes the CDF of X.
Proof: See Wilde, p. 43.

However, ideally we would express expectations in terms of the usual Riemann integral we know from high-school calculus. The first step is to transform the expression further to one that weighs intervals by their length.

5.2.2 Proposition

Suppose X has finite expectation, and F_x is absolutely continuous, i.e. there is a function $f(\cdot) \geq 0$ such that $F_x(x) = \int_{\infty}^{x} f(x)dx$ with dx the Lebesgue measure. Then we have $E(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx$.
Proof: See Wilde p. 47.

In fact, we have now an expression very close to that from high-school calculus, and it turns out that whenever g(x) is Riemann-integrable, then the value of the integral is the same wether we consider it as a Riemann or Lebesgue integral.

In the following we will only consider Riemann-integrable functions.

Note:

- The integral is well-defined if
  1. $g(\cdot)$ is "measurable" with respect to Ω and
  2. finite, i.e. $\int_{-\infty}^{\infty} |g(x)| f_X(x)dx < \infty$.
  If the integral is not bounded, evaluate $I^+ = \int_{g(x)>0} g(x)f_X(x)dx$ and $I^- = -\int_{g(x)<0} g(x)f_X(x)dx$. If $I^+ = \infty$ and $I^- < \infty$, $E(g(x)) = \infty$. If $I^+ < \infty$ and $I^- = \infty$, $E(g(x)) = -\infty$. If both integrals are ∞, the expectation is not defined.

- A random variable X for which condition 2 holds, has an expected value, as it is measurable by definition. Thus, the "expected value of X" exists.

- For any random variable X and any bounded measurable function $g(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$ the composite function $f(\cdot) : S \longrightarrow \mathbb{R}$ defined as $f(s) = g(X(s))$ is measurable and bounded, so has an expectation.

Note: Not all random variables satisfy condition 2. E.g. the "Cauchy Distribution" defined by the pdf $f(x) = \frac{1}{\pi(1+x^2)}$ has no expectation.

5.2.3 Expectation of functions of a discrete random variable

Definition: For a probability space (S, Ω, P) with finite S, the expectation of a function $g(\cdot)$ of a discrete random variable X with support $D = x_1, ..., x_n$ is defined as
$E(g(x)) = \sum_{i=1}^{n} g(x_i)P_X(x_i)$, where P_X is the pmf of X.

5.3 Moments of Random Variables

5.4 Mean and other raw moments

The "expectation" or "mean" of a random variable X, defined as $E(x) = \int_{-\infty}^{\infty} x f_X(x)dx$, or $E(g(x)) = \sum_i g(x_i)p_X(x_i)$ for discrete random variables, is also called the "first moment about 0", "first raw moment", or simply "first moment" of X.

Properties of the mean:

- E(a) = a
- By the linearity of the sum or integral operators, we have E(a+b(g(X)) = a+bE(g(X)).
- $Pr(X \geq \lambda E(X)) \leq 1/\lambda$ (Markov inequality)

The mean of X is often denoted μ_x.

More generally, for any positive m the **mth (raw) moment** of X is defined as $E(x^m)$.

5.5 Variance and other central moments

The mth "central" moment of X is defined as $E((x - \mu_x)^m)$.

The "variance of X" is the second central moment, denoted as $Var(X)$ or σ_x^2.
It can be expressed in terms of raw moments as
$Var(X) = E((x - \mu_x)^2 = E(x^2 - 2x\mu_x + \mu_x^2) = E(x^2) - \mu_x^2$.

Properties of the variance:

- $Var(a) = 0$
- $Var(a + bX) = E([a + bX \ - \ (a + b\mu_x)]^2) = b^2 Var(X)$.
- $Pr(|X - E(X)| \geq h) \leq \frac{Var(X)}{h^2}$ (Chebyshev's inequality)

The square root of the variance, σ, is called the standard deviation of x. If X has units, the standard deviation has the same units as X.
The third central moment is called "skewness", the fourth "kurtosis".

5.6 Moment generating function

The moment generating function (MGF) of a random variable X is defined for both discrete and continuous distributions as $M(\lambda) = E(e^{\lambda X})$.
The MGF is useful because, while it does not exist for all random variables, if $\exists h > 0 \ s.t. for \ -h < \lambda < h \ E(e^{\lambda X})$ is defined and finite, then $\frac{d^m}{d^m \lambda} M(\lambda)|_{\lambda=0} = E(X^m)$. Thus, the mth derivative of the moment generating function evaluated at 0 gives the mth raw moment of the random variable X.
Proposition: Two random variables share the same MGF if and only if they have the same CDF.
Proof: (to be completed)

Exercise 1: Using the MGF, show that the expectation and variance of the **Poisson distribution** $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$ are equal to λ.
Hint: Use the definition of the exponential function by its Maclaurin Series $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.
Exercise 2: Using the MGF, show that the expectation and variance of the **exponential distribution** $f(x) = \frac{1}{\theta} e^{-\theta x}$ are equal to θ and θ^2 respectively.

5.7 Additional exercises

- Banerjee Exercise sheet 1, exercise 8
- Banerjee Exercise sheet 2, exercise 2

6 Common univariate distribution functions

When looking at distribution functions, we consider the following criteria:

- What is its support, i.e. what are the values of the underlying random variable X with positive probability density or mass? These may be the integers, the positive real numbers, etc.
- What parameters characterise the distribution, or its pdf / pmf?
- What are its moments?

6.1 Discrete univariate distributions

6.1.1 Bernoulli Distribution

The Bernoulli distribution is defined for random variables that take value 1 ("success") and 0 ("failure") only. It is defined by one parameter $0 \leq p \leq 1$ such that $Prob(X = 1) = p$, and $Prob(X = 0) = 1 - p$, or more compactly $P(X = x) = p^x(1-p)^{(1-x)},\ x \in \{0, 1\}$

Moments

E(x)=p

Var(x)=p(1-p)

6.1.2 Binomial Distribution

The random variable X calculated as the sum of N independent Bernoulli distributed random variables with identical p follows the binomial distribution. It is characterised by the two parameters N and p.

The range of X are the integers from 0 to N, with probability mass function $Prob(X = x) = Prob(\text{"}x\ times\ 1\ and\ N - x\ times\ 0\text{"}) = (\frac{N!}{x! * (N-x)!}) * p^x * (1 - p)^{(N-x)}$.

Exercise: Show that the moments of the Binomial distribution are given as $E(X) = n \cdot p$, Var(X)=np(1-p).

6.1.3 Poisson Distribution

The Poisson distribution has probability mass function
$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!},\ for\ x = 1, 2, 3, ...$ It is characterised entirely by the parameter lambda.

Exercise: Using the definition of the exponential function, show that $E(X) = Var(X) = \lambda$.

6.2 Continuous univariate distributions

6.2.1 Uniform Distribution

The pdf of a uniform distribution is $f(x) = \frac{1}{b-a},\ for\ x \in [a,b]$. It is characterised by the two parameters a,b.

Exercise: Show that the moments of the uniform distribution +are $E(X) = \frac{b-a}{2},\ Var(X) = \frac{(b-a)^2}{12}$.

6.2.2 Exponential Distribution

The exponential distribution has pdf $f(x) = \frac{1}{\theta}e^{\frac{x}{\theta}}$, for $\infty > x \geq 0$ and $\theta > 0$ its only parameter.

Exercise

Using the moment generating function show that $E(x) = \theta$ and $Var(x) = \theta^2$.

6.2.3 Normal Distribution

The standard normal distribution X has pdf $f(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}},\ x \in \mathbb{R}$. Its CDF has no closed form solution. As it is symmetric around 0, all odd central moments are 0. Also, $E(X^2) = Var(X) = 1$, so one writes often $X \sim N(0,1)$

Exercise: Show, using the change of variables technique, that $Y = \mu + \sigma X$ has $f(y) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}$, and using iterated integration by parts, that $E(X) = \mu$, and $Var(X) = \sigma^2$.

Y is a general univariate normal distribution, also written as $Y \sim N(\mu, \sigma^2)$. It is entirely characterised by its mean and variance.

7 Multivariate Random Variables

So far we have looked at random variables that take values on the real line. A natural extension is to look at functions that map every outcome of a random experiment into an ordered list, or vector, of n real numbers. Their joint distribution function maps from the n-dimensional half-closed balls into the interval $[0, 1]$, and is a natural extension of the univariate case. But with more than one random variable on the same probability space we can ask some new questions. First, we can derive new distribution functions of a k-dimensional subvector X_I, for $X = (X_I, X_J)$ by fixing the (n-k) remaining random variables at a particular value $x_J = (x_{k+1}, ..., x_n)$ ("conditional distribution of X_I"), or by summing the probability mass, or density, over all values of X_J ("marginal distribution of X_I"). Furthermore, we can ask if there is any dependence, or likely comovement between random variables, summarised for example in the "covariance" of two random variables, or the relationship between the conditional and marginal distribution functions of a subvector X_I.

7.1 Bivariate Random Variables

Definition: A pair of random variables (X,Y) on the same probability space make a measurable function from the sample space into $\mathbb{R}^2$, and are called a "bivariate random variable".

7.1.1 Joint distribution function

The event $\{s \in S : (X(s), Y(s)) \in B^2\}$ is a well-defined event for all Borel-sets $B^2 \in \mathbb{B}^2$. So we can define the joint probability function of a bivariate random variables on the 2-dimensional Borel sets as
$P(X \in B_1, Y \in B_2) = P(s \in S : X(s) \in B_1 \; and \; Y(s) \in B_2)$ for all Borel sets $B^2 = (B_1, B_2) \in \mathbb{B}^2$. Again using the fact that all Borel sets can be constructed as unions, intersections or complements of the half-open intervals $([-\infty, x], [-\infty, y])$ with $(x, y) \in \mathbb{R}^2$, we can summarize the probability distribution of a bivariate random variable by its **joint cumulative distribution function** $F(x, y) = P(X \leq x, Y \leq y)$.

Properties of the joint cumulative distribution function
The properties of the joint cumulative distribution function of a bivariate random variable mirror those of the univariate distribution function and are stated without proof.

1. $F(\cdot)$ is a non-decreasing function in both of its arguments.
2. $F(\cdot)$ is a right-continuous function in both of its arguments.
3. $F(-\infty, y) = F(x, -\infty) = F(-\infty, -\infty) = 0$
4. $F(\infty, \infty) = 1$

Joint probability mass function

If X and Y are both discrete random variables on the same probability space, or "jointly discrete", with positive mass on the support $D_X = \{x_1, x_2, ...\}, D_Y = \{y_1, y_2, ...\}$, we can define the bivariate probability mass function as $p(x, y) = P(X = x, Y = y)$, with properties similar to those of univariate PMFs, i.e.

1. $1 \geq p(x, y) \geq 0$
2. $p(x, y) = 0 \; \forall \; x \notin D_X, y \notin D_Y$
3. $\sum_{x \in D_x} \sum_{y \in D_y} p(x, y) = 1.$

Joint probability density function

If X and Y are both continuous random variables on the same probability space, or "jointly continuous", we can define the bivariate probability density function $f(x, y)$ for the (usual) case where $F(\cdot)$ is differentiable with respect to all of its arguments, by $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y) dx dy$ with the properties

1. $f(x, y) \geq 0$
2. $\int \int_{\mathbb{R}^2} f(x, y) dx dy = 1$

7.1.2 Marginal distribution function

For every Borel set A, $X \in A$ is a well-defined event with probability $P(X \in A) = P(s \in S : X(s) \in A) = P(s \in S : X(s) \in A \; and \; Y \in \mathbb{R})$. The marginal distribution of X is obtained from the joint distribution by summing, for any given Borel subset of X-values, the joint probability over all Y values.

For jointly discrete X and Y this yields the **marginal PMF of X** $p(x) = P(X = x) = \sum_{y \in D_Y} p(x, y)$ and marginal CDF of X $F_X(x) = \sum_{s_x \in D_X : s_x \leq x} \sum_{y \in D_Y} p(s_x, y)$, where D_X, D_Y are the supports of X and Y.

For jointly continuous X and Y, the **marginal probability density of X** is $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$, and the marginal CDF of X $F_X(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y) dx dy = lim_{y \longrightarrow \infty} F(x, y)$.

7.1.3 Conditional distribution function

For a bivariate random variable (X,Y) and a given Borel set C, the Probability $P(X \in A, Y \in C)$ is well defined for every Borel set A. Thus we can calculate the conditional Probablity of $X \in A$ given $Y \in C$ as $P(X \in A \mid Y \in C) = \frac{P(X \in A, Y \in C)}{P(Y \in C)}$.

For jointly discrete random variables X and Y, we can calculate the **conditional PMF of X, given Y=y** $p_{X|Y}(x \mid y) = P(X = x, \mid Y = y) = \frac{p(x,y)}{p_Y(y)}$. The **conditional CDF** is simply $F(x \mid y) = \sum_{s_x \in D_X : s_x \leq x} p_{X|Y}(s_x \mid Y = y)$.

For jointly continuous random variables, the conditional CDF $F_{X|Y}(x \mid Y = y)$ runs into the problem that $P_Y(Y = y) = 0$, so the notation from the simple definition of conditional probability is unavailable.

Proposition: Using limits, we can write $F_{X|Y}(x \mid Y = y) = lim_{\epsilon \longrightarrow 0^+} P(X \leq x \mid y - \epsilon < Y < y + \epsilon)$. The **conditional PDF** is then $f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}$
Proof:
$f_{X|Y}(x \mid y) = \frac{\delta}{\delta x} F_{X|Y}(x \mid y)$
$= \frac{\delta}{\delta x} lim_{\epsilon \longrightarrow 0^+} \frac{P(X \leq x, y-\epsilon < Y < y+\epsilon)}{P(y-\epsilon < Y < y+\epsilon)}$
$= \frac{\delta}{\delta x} lim_{\epsilon \longrightarrow 0^+} \frac{F(x,y+\epsilon) - F(x,y-\epsilon)}{F_Y(y+\epsilon) - F_Y(y-\epsilon)}$
$= \frac{\delta}{\delta x} \frac{\frac{\delta}{\delta y} F(x,y)}{f_Y(y)}$
$. = \frac{f(x,y)}{f_Y(y)}$

Example: Uniform distribution on the plane
Two jointly distributed continuous random variables have the bivariate uniform distribution on $(0, a) \times (0, b)$ if their joint density is a constant, i.e. $f(x, y) = c, c \in \mathbb{R}$. We can get the value of c depending on the support of $f(\cdot)$ by setting
$\int_0^a \int_0^b c dy dx = 1$, so $c = \frac{1}{ab}$.
For the CDF of x and y, $F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{ab} dx dy$, we need to distinguish between 5 different cases:

- $F(x, y) = 0$ for $x, y : x < 0$ *or* $y < 0$
- $F(x, y) = \frac{1}{ab} xy$ for $\{x, y : 0 \leq x \leq a, 0 \leq y \leq b\}$
- $F(x, y) = \frac{1}{a} x =$ for $\{x, y : 0 \leq x \leq a, b < y\}$
- $F(x, y) = \frac{1}{b} y =$ for $\{x, y : a < x, o \leq y \leq b\}$

- $F(x, y) = 1$ for $\{x, y : a < x, b < y\}$

The marginal pdf of X is $f(x) = \frac{1}{a}$ on $\{x : 0 \leq x \leq a\}$ and 0 otherwise. So the marginal CDF of X is simply $F(x, y) = \frac{1}{a}x =$ for $\{x : 0 \leq x \leq a\}$, 0 for $x < 0$ and 1 for $x > a$.
The conditional pdf of X given Y=y is $f_{X|Y}(x \mid Y = y) = \frac{f(x,y)}{f(y)} = \frac{1}{a}$. The conditional CDF follows as above. **Exercise (Champagne 2003)**:
A rope of length L is cut into three pieces in the following way:

- The first piece of length X is obtained by cutting the rope at random (with uniform probability for all points $x \in [0, L]$.
- The second piece of length Y is obtained by cutting the remaining segment of length $L - X$ at random.
- The third piece is obtained as the remaining segment of length $L - X - Y$.

1. Find $f_Y \mid X(y \mid x)$, the conditional PDF of Y given $X = x$, $(0 < x < L)$.
2. Find f(x, y), the Joint PDF of X and Y , and illustrate the region of the plane where it takes on non-zero values.
3. What is the probability that both X and Y be less than $L = 2$?

7.1.4 Expectations and moments of bivariate random variables

The expectation of a bivariate random variable is simply the expectations of the individual random variables written as a vector. But often we are interested in the expectation of a measurable function $g(\cdot) : \mathbb{R}^2 \longrightarrow \mathbb{R}$ of two jointly distributed random variables X and Y. As it turns out, to calculate the expectation $E(g(x, y))$, we do not have to calculate the distribution function of the new random variable $Z = g(x, y)$, which is often difficult.
For jointly discrete random variables with support $D_X = \{x_1, x_2, ...\}, D_Y = \{y_1, y_2, ...\}$ we have $E(g(x, y) = \sum_{D_X} \sum_{D_Y} g(x, y) f(x, y)$.
For jointly continuous random variables we have $E(g(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy$.

Definition: Covariance
For two jointly distributed random variables X and Y with unconditional means μ_X and μ_Y, the covariance of X and Y is defined as $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

Properties of the covariance

1. $Cov(X, X) = Var(X) = \sigma_X^2$ (by the definition of variance and covariance)
2. $Cov(X, Y) = Cov(Y, X)$ (by the commutativity of the product operator)
3. $Cov(aX + b, cY + d) = acCov(X, Y)$ (by the linearity of the expectations operator)
4. $Cov(X, Y) = E(XY) - E(X)E(Y)$ (as $E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y) = E(XY) - 2E(X)E(Y) + E(X)E(Y)$
5. $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$ (as $E((X + Y - \mu_X - \mu_Y)^2) = E((X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)) = Var(X) + Var(Y) + 2Cov(X, Y))$

Definition: Correlation coefficient ρ
The correlation coefficient ρ of two jointly distributed random variables is defined on the basis of their covariance as $\rho(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, where σ_X and σ_Y are the standard deviations of X and Y.

Properties of the correlation coefficient

1. ρ is dimensionless even for random variables that have units.
2. $-1 \leq \rho \leq 1$ (as $\mid Cov(X, Y) \mid \leq \sqrt{Var(X)Var(Y)} = \sigma_X \sigma_Y$)
3. $\rho(X, X) = 1$ (as $Cov(X, X) = Var(X) = \sigma_X \sigma_X$)
4. $\rho(X, Y) = \rho(Y, X)$ (as $Cov(X, Y) = Cov(Y, X)$)
5. $\rho(aX+b, cY+d) = sign(ac)\rho(X, Y)$ (as $Cov(aX+b, cY+d) = acCov(X, Y)$, for $Z = aX + b$ $\sigma_Z = \mid a \mid \sigma_X$, etc.)
6. $\rho(X, Y) = 1 \iff Y = aX + b$ for any $a > 0$, *and any* $b \in \mathbb{R}$
   $\rho(X, Y) = -1 \iff Y = aX + b$ for any $a < 0$, *and any* $b \in \mathbb{R}$ (by a similar argument)

Note: The correlation coefficient is a measure of linear association between two random variables. From the last property, it equals 1 or -1 whenever one random variable is a linear affine function of the other, e.g $Y = a + bX$.

7.1.5 Independence

The random variables X and Y are defined to be independent if the events $X \in A$ and $Y \in B$ are independent for any pair of Borel sets A and B, i.e. $P(X \in A, Y \in B) = P(X \in A)P(Y \in B),\ \forall A, B \in \mathbb{B}$. Of course this also holds for the half-open intervals, so $\mathrm{F}(x, y) = P(X \leq x, Y \leq y) = F_X(x)F_Y(y)$. By the definition of the joint PMF, this implies for independent jointly discrete random variables $p(x, y) = p_X(x)p_Y(y)$. For jointly continuous random variables we equally have $f(x, y) = f_X(x)f_Y(y)$.

Properties of independent random variables
For any 2 independent jointly distributed random variables X,Y

1. The random variables defined by measurable function g(x) and h(y) are also independent.

2. Conditional distributions are equal to marginal distributions, e.g. $F_{X|Y}(x \mid y) = F_X(x)$, and equivalently for PMF, PDF.

3. $E(g(x)h(y)) = E(g(x))E(h(y))$ for all measurable functions g, h. Particularly, $E(xy) = E(x)E(y)$, i.e. for independent random variables the expectation of the product is the product of the expectation.

4. 3. implies that the moment generating function for the sum of two random variables is the product of the individual moment generating functions.

5. Cov(X,Y)=corr(X,Y)=0. But not all bivariate random variables where this holds are independent, e.g. for any random variable X that is symmetric around 0, or where $E(X) = E(X^3) = 0$, we have $Cov(X, X^2) = 0$.

Exercise: Prove properties 1-4 for 2 jointly continuous random variables X and Y.

7.1.6 Conditional expectation and variance

With more-dimensional random variables, when taking expectations, we have to specify which random variable and which distribution we are taking expectations over, and which values of the remaining random variables we "condition" the expectation on.
With a bivariate jointly continuous random variable, we have

1. The **unconditional expectation of X** is a real number defined by $E(X) = \int_X \int_Y xf(x,y)dxdy = \int_X xf_X(x)dx = \mu_X$.

2. The **conditional expectation of X given** $Y = y$ is a function $\mu_{X|Y}(\cdot) : D_Y \longrightarrow D_X$ that maps every value y into the conditional expectation of X given Y=y. $\mu_{X|Y} = E_{X|Y}(X \mid y) = \int_X xf_{X|Y}(x \mid y)dx$.

3. Equally, the **conditional Variance of X given Y=y** is defined as $\sigma^2_{X|Y} = E_{X|Y}((X - E_{X|Y}(X \mid y))^2 \mid y) = E_{X|Y}(X^2 \mid y) - (E_{X|Y}(X \mid y))^2$.

Both the conditional expectation and conditional variance are measurable functions of random variables, and thus random variables themselves.

7.1.7 Law of iterated expectations and Decomposition of Variance

The simple law of iterated expectations states that the unconditional expectation of X equals the "expectation over Y of the conditional expectation function of X given Y=y", or $E_Y(E_{X|Y}(X \mid y)) = E(X)$.

Exercise 1: Prove the simple law of iterated expectations for X,Y two jointly continuous random variables.
Exercise 2: Decomposition of Variance
The Variance decomposes into $Var(X) = Var(X \mid Y = y) + Var(E_{X|}(X \mid y)$.
Prove this for X,Y two jointly continuous random variables.
Exercise 3: Conditioning Theorem
Prove for X,Y two jointly continuous random variables the conditioning theorem
$E_{X|Y}(g(y)X \mid y) = g(y)E_{X|Y}(X \mid y)$.

Exercise 4 (Based on Champagne 2003):
Let RVs X and Y be jointly uniform over the region $D = \{(x,y)\} : 0 < x < y < 1$.

1. Draw the region D.

2. Find the real number u such that x,y are jointly uniform on D with pdf $f(x,y) = u,\ \forall x,y \in D$, and $f(x,y) = 0$ otherwise.

3. Find $E(X),\ E(Y),\ E(X^2),\ E(XY)$

7.2 Additional exercises

- Banerjee Exercise sheet 2, exercise 4
- Banerjee Exercise sheet 1, exercise 10

7.3 Multivariate Random Variables

7.3.1 Definition: Random Vector

A list of n random variables (X,Y) on the same probability space make a measurable function from the sample space into $\mathbb{R}^n$, and is called an "n-dimensional random vector".

7.3.2 Joint distribution function

The event $\{s \in S : (x_1, x_2, ...x_n) \in B\}$ is a well-defined event for all n-dimensional Borel-sets $B = (B_1, B_2, ..., B_n) \in \mathbb{B}^n$, so the joint probability $P(x_1 \in B_1, x_2 \in B_2, ..., x_n \in B_n)$ is well defined. As in the bivariate case, it can be summarized by the **joint CDF** of $X1, ..., Xn$, defined as $F(x_1, x_2, ...x_n) = P(X1 \leq x1, X2 \leq x2, ..., Xn \leq xn)$.

Properties of the Joint CDF
These are equivalent to those of the bivariate case, i.e.

1. $F(\cdot)$ is a non-decreasing function in all of its arguments.
2. $lim_{x_i \longrightarrow -\infty} F(x_1, x_2, ...x_n) = 0,\ i \in \{1, ..., n\}$
3. $lim_{x_1 \longrightarrow \infty, x_2 \longrightarrow \infty, \ ...} F(x_1, x_2, ...x_n) = 1$

If $X_1, X_2, ...$ are all "jointly continuous" we can define the multivariate probability density function $f(x_1, x_2, ...x_n)$ by $f(x_1, x_2, ...x_n) = \frac{\delta^n}{\delta x} F(x_1, x_2, ...x_n)$ with $\delta x = (\delta x_1, \delta x_2, ..., \delta x_n)$. The definition of the joint PMF for jointly discrete random variables is equivalent to the bivariate case and omitted here.

7.3.3 Marginal distribution function

The **marginal CDF** of a subvector X_I, $I \subseteq \{1, 2, ..., n\}$ is obtained by letting all $x_j \longrightarrow \infty$, for $X_j \in X_J, J = \{1, 2, ..., n\} \backslash I$.
For n jointly continuous random variables, the **marginal PDF** of a subvector X_I is obtained from the joint PDF by integrating over all $X_j \in X_J, J =$

$\{1,2,...,n\}\backslash I$, i.e. $f_{X_I} = \int_J f(x_1, x_2, ..., x_n)dx_J$.
Equally, the marginal PMF of a subvector X_I is obtained by summing over all possible values of $X_j \in X_J, J = \{1,2,...,n\}\backslash I$, i.e. $p_{X_I} = \sum_J p(x_1 1, x_2, ..., x_n)$.

7.3.4 Conditional distribution function

In analogy to the bivariate case we define the conditional CDF of subvectors X_I given values x_j of subvector X_J as $F_{X_I|X_J}(x_I \mid x_J) = P(X_I \leq x_I \mid X_J = x_J) = lim_{\epsilon \longrightarrow 0^+} \frac{P(X_I \leq x_I, x_J - \epsilon \leq X_J \leq x_J + \epsilon)}{P(x_J - \epsilon \leq X_J \leq x_J + \epsilon)}$. So the conditional PMF of n jointly discrete random variables is $p_{X_I|X_J}(x_J \mid x_J) = \frac{p(x_I, x_J)}{p_{X_J}(x_J)}$, wherever $p_{X_J}(x_J) > 0$. And the conditional PDF of n jointly continuous random variables is $f_{X_I|X_J}(x_I \mid x_J) = \frac{f(x_I, x_J)}{f_{X_J}(x_J)}$.

7.3.5 Expectation and covariance of n random variables

The expectation of a function $g(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}$ of an n-dimensional vector is
$E(g(X_1, X_2, ..., X_n)) = \int_{X_1} \int_{X_2} ... \int_{X_n} g(x_1, x_2, ..., x_n) f((x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n$.
Or for the special case of jointly discrete random variables
$E(g(X_1, X_2, ..., X_n)) = \sum_{X_1} \sum_{X_2} ... \sum_{X_n} g(x_1, x_2, ..., x_n) p((x_1, x_2, ..., x_n)$.
The covariance being a bivariate concept we can extend it to the n-variable case by defining the **Variance-Covariance Matrix of the multivariate random variable** $X = (X_1, X_2, ..., X_n)$ as

$$\Sigma = \begin{bmatrix} Var_{X_1} & Cov_{21} & ... & Cov_{n1} \\ Cov_{21} & Var_{22} & ... & Cov_{1n} \\ ... & & ... & \\ Cov_{1n} & Cov_{12} & ... & Var_{nn} \end{bmatrix}$$

7.3.6 Independence of n random variables

The random variables $X_1, X_2, ..., X_n$ are called independent if for all Borel sets $B_i \in \mathbb{B}, i = 1, ..., n$, the events $(X_1 \in B_1), (X_2 \in B_2), ..., (X_n \in B_n)$ are mutually independent. This implies $P((X_1 \in B_1), (X_2 \in B_2), ..., (X_n \in B_n)) = P(X_1 \in B_1)P(X_2 \in B_2)...P(X_n \in B_n)$.
Thus for jointly discrete random variables the joint PMF is $p(x_1, x_2, ..., x_n) = p_{X_1}(x_1)p_{X_2}(x_2)...p_{X_n}(x_n)$, and for jointly continuous random variables, the joint

PDF is $f(x_1, x_2, ..., x_n) = f_{X_1}(x_1)f_{X_2}(x_2)...f_{X_n}(x_n)$.
The VCM of n independent random variables is diagonal.

7.3.7 General law of iterated expectations

A more general version of the law of iterated expectations is $E_Y \mid Z(E_{X|Y,Z}(X \mid y, z)) = E_{X|Z}(X \mid z)$.

7.3.8 Multivariate Transformations

The transformation theorem, or "change-of-variables formula", extends to functions of n-dimensional random vectors, $g(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ that are one-to-one, differentiable and invertible. Defining a new random vector Y as $Y = g(x)$, and the inverse function $h(y) = g(x)^{-1}$, and using the definition of the Jacobian matrix as the determinant of the matrix of first partial derivatives of a vector valued funtion, i.e.
$J(y) = det(\frac{\delta}{\delta y'}h(y))$, the probability distribution function of Y is then
$f_Y(y) = f_X(h(y)) \mid J \mid$.
Proof: (Difficult).
Example: Let random variables X and Y be jointly uniform over the region $D = \{(x, y)\} : 0 < x < y < 1$ as before.
Let R and Q be defined by $R = 2 + 3x - 2y$, and $Q = 1 - 0.5x + y$. What is the joint pdf of R and Q?
First, note that $g(x, y) = (2+3x-2y, 1-0.5x+y)$, which yields $h(r, q) = (y, x) = (0.5q+1/12r-2/3, 1/2r+q-1)$. So the Jacobian is $det(\begin{bmatrix} 1/2 & 1 \\ 1/12 & 1/2 \end{bmatrix}) = 1/6$.
Thus we get $f(r, q) = 1/3$, so r and q are uniformly distributed with parameter 1/3. It is easy to check that f(r,q) integrates to 1 over its domain (to be completed).

7.4 Multivariate Normal Distribution

For non-singular Σ_X, the PDF of an n-dimensional normal random variable X is
$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} det(\Sigma_X)^{1/2}} e^{-\frac{(x-\mu_X)'\Sigma_X^{-1}(x-\mu_X)}{2}}$.

Its domain is $x \in \mathbb{R}^n$. Its expectation μ_X and the $(n+1)n/2$ distinct elements

of the variance-covariance matrix Σ_X are the parameters that completely characterise its distribution, so we write $X \sim N(\mu_X, \Sigma_X)$.

Note:

The following are equivalent definitions of an n-dimensional normal random vector X, allowing for singular Σ_X:

- Every linear combination $Y = a_1X_1 + ... + a_nX_n$ is normally distributed.
- There is an m-dimensional random vector Z whose elements are independent normal random variables, a vector of real numbers $\mu = (\mu_1, ..., \mu_n)$, and an nxm matrix A such that $X = \mu + Az$.
- There is a vector of real numbers $\mu = (\mu_1, ..., \mu_n)$, and a positive semi-definite matrix Σ_X, such that the Moment generating function of X is $e^{\lambda'\mu + \lambda'\Sigma_X\lambda/2}$.

7.4.1 Conditional and marginal distributions of subvector $x_I = (x_1, ..., x_k),\ k \leq n$

If we partition X into a kx1 and an (n-k)x1 vector as $X = (X_I, X_J)$, and accordingly $\mu_X = (\mu_I, \mu_J)$ and

$\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ with Σ_{11} kxk, Σ_{21} kx(n-k), etc., the conditional distribution of X_I given X_J is multivariate normal $X_I \mid X_J = x_J \sim N(\mu_{I|J}, \Sigma_{I|J})$, where $\mu_{I|J} = \mu_I + \Sigma_{12}\Sigma_{22}^{-1}(x_J - \mu_J)$ and $\Sigma_{I|J} = \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Knowing the value of x_J alters mean and variance of X_I.

On the other hand, the marginal distribution of X_I is simply $N(\mu_I, \Sigma_{11}$.

Exercise 1: Using the fact that the covariances of independent random variables are zero, show that the joint pdf of n indepent jointly normal variables is the product of the univariate normal pdfs. This also shows, that for the particular case of normal random variables, independence implies 0 covariance AND vice versa.

Exercise 2: Using the multivariate change-of-variables formula, show for an n-dimensional normal vector X, and $Y = A + Bx$, that linear affine transformations of multivariate normal variables are also normally distributed, with

$f_Y(y) = \frac{1}{2\pi^{\frac{n}{2}} det(B\Sigma_X B')^{1/2}} e^{-\frac{(y-B\mu_X)'(B\Sigma_X B')^{-1}(x-B\mu_X)}{2}}$

$= \frac{1}{2\pi^{\frac{n}{2}} det(\Sigma_Y)^{1/2}} e^{-\frac{(x-\mu_Y)'\Sigma_Y^{-1}(x-\mu_Y)}{2}}$.

7.5 Additional exercises

- Banerjee Exercise sheet 2, exercise 3