Conditional Exqəɬtations

by Khallil Ebrahim Benyattou

少(すこ)しずつ

January 20, 2026

This page intentionally left almost blank.

Contents

Chapter 1: Introduction
1.1 Describing/characterising a set of measurements 9
Visualising Data.

Chapter 2: Mathematical Framework of Experiments
2.1 Experiments 11
2.2 Probability 11
Frequentist Interpretation; Kolmogorov's Axiomatic Framework.
2.3 Collections of Events, $\mathcal{F}$ 12
Generating (σ-)Algebras.
2.4 The Extended Real Line, $\overline{\mathbb{R}}$ 14
2.5 Measures on $\mathcal{F}$ 16
Sub-σ-algebras and Subspace Measures; Completion of Measure.
2.6 Why Not Always $\mathcal{F} = 2^X$? 21
A Non-Measurable Set (Vitali Set); Damage Control.
2.7 Probability Measures 23
Defining Ω and Counting Subsets.
2.8 Independence 25
2.9 (Naïve) Conditional Probability 26

Chapter 3: Constructing Measures
3.1 Terminology 29
3.2 Chapter Roadmap 29
3.3 Approximation by Covering 30
Outer Measure.
3.4 Outer Measurability and Carathéodory Extension 33
3.5 Refined Carathéodory Extension 36
Extension From $\mathcal{S}$ to Alg($\mathcal{S}$); Extension From Alg($\mathcal{S}$) to $\sigma(\mathcal{S})$; Uniqueness Of Our Extension.
3.6 Defining the Lebesgue Measure 47
3.7 Product Measures 48

Chapter 4: Measurable Functions
4.1 Random Variables 49
4.2 Properties of Measurable Functions 51
4.3 Probability Distribution of X 55
4.4 Support of Probability Distribution 56
Borel Spaces; Measurable Classification of Borel Spaces; Topological Support of Probability Measure.

Chapter 5: Measuring Functions (Integral)
5.1 Historical Shortcomings 59
5.2 Definition of the Lebesgue Integral 60
5.3 MCT, Fatou, DCT, and Fubini 63
5.4 Absolute Continuity & Radon-Nikodým Derivative 64
5.5 Pushforward Measure & Change of Variables 65

5.6 Types of Random Variables 67
Absolutely Continuous; Discrete.

Chapter 6: Random Vectors
6.1 Probability Distribution of $\mathbf{X}$ 72
Joint CDF.
6.2 Types of Random Vectors 74
6.3 Marginal Distributions 74
6.4 Conditional Distributions 76
Jointly Discrete; Jointly Absolutely Continuous.
6.5 Independence of Random Variables 77

Chapter 7: Averages, Dispersion, and Correlation
7.1 Variance 79
7.2 Covariance 80
7.3 Correlation 81
7.4 Calculating Expectations and Variances of Linear Combinations 82

Chapter 8: Discrete Probability Distributions
8.1 Uniform 84
8.2 Bernoulli 84
8.3 An Important Point! 85
8.4 Binomial 86
8.5 (Discrete) Geometric 87
Memorylessness.
8.6 Negative Binomial 89
8.7 Hypergeometric 90
8.8 Hypergeometric Approximates Binomial 91

Chapter 9: Point Processes (Random Scatters)
9.1 Point Processes 92
9.2 Poisson Point Processes 93
9.3 Poisson Processes 97
A Bridge to Absolutely Continuous Distributions.

Chapter 10: Absolutely Continuous Distributions
10.1 Uniform Distribution 100
10.2 Normal Distribution 100
Standard Normal; Link: Normal Approximates Binomial.
10.3 Gamma Distribution 102
Shape; Scale Parameter; Link: Poisson and Gamma.
10.4 Gamma($\nu/2, \beta = 2$), The Chi-Squared Distribution ($\nu \in \mathbb{N}$) 106
10.5 Gamma($\alpha = 1, \beta$), The Exponential Distribution 107
Memorylessness; Link: Exponential and Geometric.
10.6 Beta Distribution 108
Link: Beta and Binomial.
10.7 Chebyshev's Theorem 109
10.8 Expectations of Discontinuous Functions and Mixed Probability Distributions 109
Expectation of a Mixed Random Variable.
10.9 Summary 110
10.10 Location-Scale Families 110

Chapter 11: Moment-Generating Functions
11.1 Technical Points 113
11.2 Generating Moments 114
11.3 Alternative Derivation 115

Chapter 12: Multivariable Distributions
12.1 Multinomial Distribution 116
12.2 Bivariable Normal Distribution 116

Chapter 13: Population and Sampling
13.1 What Really Is A Population? 117
13.2 Inadequacies of a Single Space 119
13.3 The Two-Space Framework 121
13.4 Simple Random Sampling With Replacement (SRSWR) 122
13.5 Canonical Randomness 124
Constructive Proof For h (Π at most countable); Non-trivial Example (SRSWR).
13.6 Simple Random Sampling 127

Chapter 14: Functions of Random Variables
14.1 Statistics, Estimators and Estimates 130
How "Good" Is An Estimator?.
14.2 The 3 Methods 131
14.3 Method of Distribution Functions 132
$Z = X + Y$; Leibniz's Integral Rule; $Z = X - Y$; $Z = X - Y$ (Non-Negative X, Y); $R = \sqrt{X^2 + Y^2}$; $Z = X/Y$; $Z = \max(X, Y)$, $W = \min(X, Y)$; $Z = \max(X, Y)$; $W = \min(X, Y)$; $Z = \max(X, Y)/\min(X, Y)$.
14.4 Method of Transformations 144
14.5 Method of Moment-Generating Functions 146
14.6 Multivariable/variate Transformations Using Jacobians 148
14.7 Order Statistics 148

Chapter 15: Sampling Distributions
15.1 Sampling From A Normally Distributed Population 155
Multivariable-Multivariate Transformations.
15.2 The t-distribution 161
The Density of Student's t-distribution; Properties of the t_ν-distribution.
15.3 The F-Distribution 166
The Density of the F-distribution; Properties of the F_{ν_1,ν_2}-distribution.

Chapter 16: More Measurability
16.1 Theorem A.42 171
16.2 Application: Defining a Statistic 174

Chapter 17: Conditional Expectation
17.1 Prequel to Abstract Conditional Expectation 175
17.2 Abstract Conditional Expectation ($X \in L^1$) 180
17.3 Properties of Conditional Expectation ($X \in L^1$) 181
An Illustrative Example of Conditional Expectation.
17.4 The Conditional Expectation ($X \geqslant 0$) 188
Properties of $\mathbb{E}[X \,|\, \mathcal{G}]$ for $X \geqslant 0$.
17.5 Conditional Expectation As Projection ($X \in L^2$) 189
Application to Conditional Expectation.

17.6 More Properties of Conditional Expectation ($X \geqslant 0$ or $X \in L^1$) 190

Chapter 18: Conditional Probability
18.1 Conditional Probability of A given $\mathcal{G}$ 191
18.2 Regular Conditional Probability 193
Regular conditional distribution of X given $\mathcal{G}$; $\mathcal{G}$ gen. by partition; Defining conditional expectation via r.c.p on $\mathcal{F}$ given $\mathcal{G}$-arbitrary.
18.3 I'm Disintegrating 199
Extending the Disintegration Formula; Conditional law of X given Y; Push-forward of Markov kernels; Example: Conditional Density Formula; So many random variables.

Chapter 19: Disintegration (Separable Case)
19.1 Atoms 211
19.2 Decomposition Theorem 213

Chapter 20: Approximation
20.1 Approximation 216
20.2 Convergence in Probability 217
Consistency; Jensen's Inequality.
20.3 Almost Sure Convergence 219
20.4 Convergence in Distribution 220
20.5 An Approximation for $\overline{X}$ — Classical Central Limit Theorem 222
20.6 The Normal Approximation to the Binomial Distribution 225

Chapter 21: Goodness of Estimators, Point and Interval
21.1 Point Estimation 226
21.2 Interval Estimation 226
21.3 Bias and MSE of Point Estimators 227
21.4 Evaluating the Goodness of a Point Estimator 228
21.5 Pivotal Method for Interval Estimation 229
21.6 Selecting the Sample Size 231
21.7 Large-Sample Confidence Intervals 232
21.8 Small-Sample Confidence Intervals 233
μ; $\mu_1 - \mu_2$.
21.9 Confidence Intervals for σ^2 234
21.10 Summary 235
21.11 Properties of Point Estimators 236
Relative Efficiency; Consistency.

Chapter 22: Data Reduction
22.1 Sufficiency 239
Factorisation Theorem.

Chapter 23: Temp
23.1 2025-09-21, Decomposing Spaces 245
Application to Conditioning.
23.2 2025-10-13, Lebesgue-Stieltjes Measure 246
23.3 2025-11-11, Hypothesis Testing 247

Chapter A: Rings $\overset{?}{\leftrightarrow}$ Algebras

Chapter B: Extending Properties
B.1 π-λ Theorem 250

Chapter C: The Big Three
C.1 Monotone Convergence Theorem 252
C.2 Dominated Convergence Theorem 254
Fatou's Lemma.
C.3 Non-Example for MCT/DCT 257
C.4 Fubini's Theorem 258
C.5 Non-Example for Fubini's Theorem 258

Chapter D: Conditional Independence
D.1 Relating Conditional Expectation and Independence 261

Chapter E: Future Topics
E.1 Hypothesis Tests 264

Foreword

I'm using Wackerly (the 6^{th} and 7^{th} editions — [6] and [7] respectively) as a spine for my learning. The tangents I take from the book are often big (but necessary) and I use several other books.

Key:

- **Keywords** are in this colour.
- **Words to look up later** are in this colour.
- **Things** I want to
- This environment denotes something I'm uneasy about or have left as a gap to fill in later.
- **link together** are in this colour. These will (almost always) be on the *same page* and occur in pairs so the link is clear.

CHAPTER 1

Introduction

What is statistics to me?

Statistics is a field of study that involves:

- designing experiments and surveys to collect sample data on a population,
- analysing said samples,
- and making inferences about the wider *population* from the sample.

A **population** is a set Π of similar objects or interests — be it real or conceptual. For example, the collection of objects that a company may manufacture at some point in the future can be considered to be a hypothetical population. A **sample** is a proper subset of a population.

What is statistics to others?

> Statistics is a theory of information, with inference making as its objective.
>
> [7, p. 2]

> The objective of statistics is to make an inference about a population based on information contained within a sample from that population, and to provide an associated measure of goodness for the inference.
>
> [7, pp. 2–3]

Before being able to make inferences, we need some way to characterise/describe a set of measurements.

1.1 Describing/characterising a set of measurements

A population can be characterised/approximated by taking a sample and creating its relative frequency distribution.[1]

- **Frequency** is the number of occurrences of a given type of event, or the number of members of a population falling into a specified class or category.
  - A **class** or **grouping** is a way of organising data into intervals or categories to summarise a dataset.
- If the frequency is expressed as a proportion of the total number of occurrences/members, it's called the **relative frequency**.

[1] This is an example of a descriptive statistic.

Therefore, a **relative frequency distribution** created from empirical data (in the form of a sample drawn from a population) is a specification of how the frequencies of sample members are distributed according to the values of the variables they exhibit.

If we consider a population to be the outcome of drawing repeatedly from some random process, we can suppose that the population is governed by some underlying theoretical "random distribution" (known as a probability distribution). The subsequent sections will develop a theory of "probability" and a consequence of this will be that we can postulate "random distributions" that model the underlying distribution of a population.

1.1.1 Visualising Data

A **histogram** is a graphical representation of a set of observations.[2] Each class frequency is represented by the area of the rectangle centred on its respective class interval. If all the class intervals are of equal length, the heights of the rectangles are also proportional to the observed frequencies. There are several types of histogram:

- A frequency histogram's bar heights represent the frequency of each bin (interval).
- A relative frequency histogram's bar heights represent the relative frequencies of data points within each bin. (Useful for comparing histograms with different sample sizes)
- Cumulative frequency histograms.

For the purpose of inference, histograms aren't usually adequate. Many histograms can be formed from the same data i.e. histograms have the potential to heavily rely on bin size and location of endpoints.

> The rule of thumb always should be that details robust to variation in bin width and bin origin are likely to be genuine; details fragile to such are likely to be spurious or trivial.
>
> Nick Cox @ **Cross Validated**

This sensitivity to variations cannot be determined a priori (from theoretical deduction) and is instead learned from observation/experience. A safe bet is to use multiple histograms with several bin widths and origins.[3]

[2] i.e. a univariate frequency diagram

[3] I saw somewhere on the internet that an alternative method is to check a **kernel density estimate with a not-too-wide bandwidth**.

CHAPTER 2

Mathematical Framework of Experiments

2.1 Experiments

Populations of interest are almost always simply too large and complex to observe completely. Experiments offer a way to sample data from a population in a controlled way, isolating the effect of specific variables[1] that allow for properties of said population to be calculated.

- An **experiment** is a procedure that has a well-defined set of possible outcomes.
  - When an experiment has more than one possible outcome and the outcome is uncertain, we call it a **random experiment**.
  - If an experiment only has one possible outcome, it's called **deterministic**.
- A **trial** is a single performance of a well-designed experiment.

2.2 Probability

Probability is a numerical measure of how likely an outcome of an experiment is.

Actually assigning probabilities should be based on experience. Ideally, these numbers should be verified by repeating an experiment (if it's even possible to do so):

2.2.1 Frequentist Interpretation

If an experiment is repeated many times, we obtain a sample of observations and from this we can calculate the proportion of outcomes corresponding to some favourable outcome we're interested in. Let $A(n)$ denote the number of trials in which the favourable outcome occurs in n repetitions of the experiment. Then the aforementioned proportion, the relative frequency of the event, is $A(n)/n$. As n grows large ($n \to \infty$), we expect that the long-run relative frequency stabilises at some value $p := \lim_{n\to\infty} A(n)/n$ in $[0, 1]$.

Definition 2.2.1 This value $p := \lim_{n\to\infty} \frac{A(n)}{n}$ is an intuitive measure of one's belief that the favourable outcome A will occur in any given trial of an experiment — the **frequentist probability of the event**.

Long-run frequencies are often computed with simulations but their accuracies depend on how well the simulation captures the random nature of the particular experiment.

Some limitations of the relative frequency interpretation of probability are quite natural consequences of the assumptions we made:

- We rely on repeated trials.
  - Some events of interest are unique or not repeatable.
  - We don't have a way to account for subjective probability.
  - We rely on the trials being identical and independent of each other. How realistic is this to implement in real life?[2]

[1] Is there a better word for this?

[2] Probably not very?

- There's a tacit assumption that we have access to infinitely many trials when calculating the long-run frequencies exactly. However, we can only work with finitely many trials in real life to approximate such frequencies.

A framework of probability that enables one to handle a wider variety of events without having to rely *only* on the observed data from infinitely many repeated trials would certainly be a wonderful thing to have. Indeed, Kolmogorov formulated an axiomatic framework of probability for this expressed level of generality:

2.2.2 Kolmogorov's Axiomatic Framework

Random experiments are formalised using the language of sets and measures. The environment in which a random experiment lives can be modelled by a triple $(\Omega, \mathcal{F}, \mathbb{P})$ known as a **probability space**.

Once we carry out a random experiment we observe some quantity e, corresponding to the theoretical outcome ω that came to fruition as a result of carrying out the experiment, as an element of a(n often numerical) set E. This correspondence is formalised by a function $X : \Omega \to E$ (called a **random variable**), and so we write $e = X(\omega)$.

Loosely speaking:

- The **outcome space** Ω of a random experiment is a non-empty[3] set of all its possible outcomes.
- A family $\mathcal{F} \subseteq 2^{\Omega}$ of subsets of Ω will be the structure that contains sets of outcomes and supports operations that describe how such sets relate to each other.

  e.g. If A and B are subsets of Ω, then we'll desire sets like their union $A \cup B$, intersection $A \cap B$, and (absolute) complement A^{c} also be in $\mathcal{F}$. Everything we'd ideally want to talk about from a probabilistic point of view when it comes to composing events will be in $\mathcal{F}$ — this is colloquially referred to as $\mathcal{F}$ being stable.
- Finally, we'll define an abstract set function $\mathbb{P} : \mathcal{F} \to [0, 1]$, assigning to each subset of outcomes $A \subseteq \Omega$, a number $\mathbb{P}(A)$ — called the probability of A — representing how likely A is to occur in any given trial.
  - The abstract definition of $\mathbb{P}$ is removed from any concept of limits of relative frequencies — it's simply a map that satisfies some basic axioms on the family $\mathcal{F}$.
  - Assigning probabilities to events A that can't be repeated is as simple as including A in $\mathcal{F}$, and assigning a number in $[0, 1]$ to A under $\mathbb{P}$. There is no need to rely on repeated trials.
  - From a terminological perspective, if $\mathbb{P}$ is well-defined on $\mathcal{F}$, in the sense that it that meaningfully assigns a single number to each event $A \in \mathcal{F}$, then A is called an **event**. If no such $\mathbb{P}$ is prescribed, we call A an element of $\mathcal{F}$.

Now for some definitions to make precise the above objects:

2.3 Collections of Events, $\mathcal{F}$

Definition 2.3.1 Suppose that Ω is a non-empty set and $\mathcal{F} \subseteq 2^{\Omega}$ satisfies the following:

(1) $\Omega \in \mathcal{F}$

[3] If it's empty, then pack it in.

(2) $\mathcal{F}$ is closed under (absolute) complementation i.e. $A \in \mathcal{F} \implies A^{\mathsf{c}} \in \mathcal{F}$

(3) $\mathcal{F}$ is closed under (at most) countably infinite unions i.e.

$$\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F} \implies \bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}.$$

Such a collection $\mathcal{F}$ is called a **σ-algebra over Ω**.

Example 2.3.2

- The membership of $\emptyset$ in every σ-algebra follows from $\Omega \in \mathcal{F}$ and closure under complementation i.e. $\mathcal{F} \ni \Omega^{\mathsf{c}} = \emptyset$. Thus, $\mathcal{F} = \{\emptyset, \Omega\}$ is a σ-algebra over Ω called the **trivial σ-algebra** over Ω.
- The power set 2^{Ω} is always a σ-algebra known as the **discrete σ-algebra** over Ω.

Practically speaking, it's often difficult to outright specify all the sets in a σ-algebra. Instead, we can build up to such a specification by considering simpler collections of subsets of Ω — arbitrary collections (as seen in the next subsection), or more structured simple collections called algebras. For the latter, we take the definition of a σ-algebra and relax the closure under countable unions (3) to closure under finite unions (3′).

Definition 2.3.3 Suppose that Ω is a non-empty set and $\mathcal{A} \subseteq 2^{\Omega}$ satisfies the following:

(1) $\Omega \in \mathcal{A}$

(2) $\mathcal{A}$ is closed under (absolute) complementation i.e. $A \in \mathcal{A} \implies A^{\mathsf{c}} \in \mathcal{A}$

(3′) $\mathcal{A}$ is closed under finite unions i.e. if $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$.

Such a collection $\mathcal{A}$ is called an **algebra over Ω**.

Remarks (Conceptual)

- De Morgan's law states that for any collection of subsets $\{A_i\}_{i\in I} \subseteq 2^{\Omega}$:

$$\Big(\bigcup_{i\in I} A_i\Big)^{\mathsf{c}} = \bigcap_{i\in I} A_i^{\mathsf{c}}.$$

  Consequently,

  - an algebra $\mathcal{A}$ is closed under finite intersections (as seen in the footnote of this page), and
  - a σ-algebra $\mathcal{F}$ is closed under countably infinite intersections.

  So we could equivalently have replaced (3′) in the definition of an algebra with

  (3″) $\mathcal{A}$ is closed under finite intersections i.e. if $A, B \in \mathcal{A}$ then $A \cap B \in \mathcal{A}$.

- Sets in an algebra have a simple representation which makes it easy for us to define other objects (like set functions) over them — this will be important in **Section 3.5.1**.

Remarks (Terminological)

- The **(relative) complement of B in A** is $B \setminus A$.
- $A, B \in \mathcal{F}$ are **mutually exclusive** (or **disjoint**) if they have empty intersection i.e. $A \cap B = \emptyset$.
- A collection of sets $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ is **pairwise disjoint**[4] if for all $i \neq j$: $A_i \cap A_j = \emptyset$.

[4] If I slip and say 'disjoint', I mean 'pairwise disjoint'. I'll try my best to say 'empty intersection' to mean $\emptyset = \cap_i A_i$.

- A **partition** of a set B is a family of non-empty subsets of B such that each $\omega \in B$ belongs to a unique subset. The subsets in a partition are called **cells**. We say that B is partitioned into $B_1, \ldots, B_n$ if the B_i are pairwise disjoint and

$$B = \bigsqcup_{i=1}^{n} B_i.$$

2.3.1 Generating (σ-)Algebras

Is there a smallest σ-algebra containing a given collection $\mathcal{C} \subseteq 2^\Omega$?

In many cases, it isn't possible to explicitly describe all sets of a σ-algebra. However, given an arbitrary collection $\mathcal{C}$ of subsets of Ω, one can define the smallest σ-algebra that contains $\mathcal{C}$. We do this by noting that:

- The collection of σ-algebras $\{\mathcal{F}_\alpha\}_{\alpha \in I}$ containing $\mathcal{C}$ is non-empty since 2^Ω is a σ-algebra.
- The intersection of an arbitrary family of σ-algebras $\{\mathcal{F}_\alpha\}_{\alpha \in I}$ on Ω is also a σ-algebra on Ω.

This gives us the following definition:

Definition 2.3.4 The **σ-algebra generated by $\mathcal{C} \subseteq 2^\Omega$**, denoted $\sigma(\mathcal{C})$, is the smallest σ-algebra containing $\mathcal{C}$ and is defined by:

$$\sigma(\mathcal{C}) := \bigcap_{\substack{\mathcal{F}_\alpha\text{-}\sigma\text{-algebra} \\ \text{s.t. } \mathcal{C} \subseteq \mathcal{F}_\alpha}} \mathcal{F}_\alpha.$$

Definition 2.3.5 Analogous statements hold for algebras and so the **algebra generated by $\mathcal{C} \subseteq 2^\Omega$**, denoted $\mathrm{Alg}(\mathcal{C})$, is the smallest algebra containing $\mathcal{C}$ and is defined by:

$$\mathrm{Alg}(\mathcal{C}) := \bigcap_{\substack{\mathcal{A}_\alpha\text{-algebra} \\ \text{s.t. } \mathcal{C} \subseteq \mathcal{A}_\alpha}} \mathcal{A}_\alpha.$$

Now we take a small reprieve from probability, and consider a more general context.

2.4 The Extended Real Line, $\overline{\mathbb{R}}$

The extended real line $\overline{\mathbb{R}}$ will be the setting in which a lot of the subsequent theory will be discussed. Definitions first, importance afterwards.

Definition 2.4.1

- As a set, we define $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.
- We extend addition and multiplication from $\mathbb{R}$ to $\overline{\mathbb{R}}$ as follows. For $a \in \mathbb{R}$, define:
  - $a + \infty = \infty$, $a + (-\infty) = -\infty$
  - $0 \cdot (\pm\infty) = 0$
  - $1/(\pm\infty) = 0$
  - $a \cdot (\pm\infty) = \begin{cases} \pm\infty & \text{if } a > 0, \\ \mp\infty & \text{if } a < 0. \end{cases}$

The expression $(+\infty)+(-\infty)$ is left undefined, for how would one assign a unique value to such an expression?

- Equipping $\overline{\mathbb{R}}$ with its natural ordering i.e. augmenting the total ordering $\leqslant$ on $\mathbb{R}$ to include $-\infty < a < +\infty$ for $a \in \mathbb{R}$, induces a topology $\mathcal{T}_o$ on $\overline{\mathbb{R}}$ called the **order topology**.
  - A basis for this topology is the collection
$$\mathcal{C} = \{(a,b)\colon a,b \in \mathbb{R},\ a<b\} \cup \{(a,+\infty]\colon a \in \mathbb{R}\} \cup \{[-\infty,b)\colon b \in \mathbb{R}\}.$$
- One can also make $\overline{\mathbb{R}}$ into a metric space by defining a metric $d(x,y) = |f(x)-f(y)|$ on it by pulling back the standard metric on $[-1,1]$ via the homeomorphism $f\colon \overline{\mathbb{R}} \to [-1,1]$ defined by
$$f(x) = \begin{cases} x/(1+|x|), & \text{where } x \in \mathbb{R} \\ +1, & \text{where } x = +\infty \\ -1, & \text{where } x = -\infty. \end{cases}$$
Note that for $x \geqslant 0$, $d(+\infty,x) = 1/(1+|x|)$, and for $x \leqslant 0$, $d(-\infty,x) = 1/(1+|x|)$.
- Depending on the context, I reserve the use of the name **extended real line** for any of the following (depending on the context):
  - The set itself $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty,+\infty\}$,
  - the topological space $(\overline{\mathbb{R}}, \mathcal{T}_o)$,
  - or the metric space $(\overline{\mathbb{R}}, d)$.

Remarks 2.4.2 It'll be of immediate interest in the next section that we're able to make statements like "$\mu(A) = +\infty$" to describe a set A that is infinitely large with respect to a way μ of measuring the size of sets.

Furthermore, the extended real line comes with some technical conveniences:

- $\overline{\mathbb{R}}$ comes with the benefit that the infimum and supremum of every subset $A \subseteq \overline{\mathbb{R}}$ exist *in* $\overline{\mathbb{R}}$.
  - As an immediate consequence, both the lim sup and lim inf of a real sequence exist $\overline{\mathbb{R}}$ since we define them as[5]
$$\limsup_{n\to\infty} a_n := \inf_n \left(\sup_{k\geqslant n} a_k\right) \quad \& \quad \liminf_{n\to\infty} a_n := \sup_n \left(\inf_{k\geqslant n} a_k\right).$$
  - As a nice application, that we shall see next chapter, one can define a way to "measure" a set "from the outside" by taking the infimum over all the ways to "cover" that set.
- We may also extend the notion of convergence of sequences to allow for limits[6] in $\{\pm\infty\}$.
  - **Proposition 2.4.3** In particular, every monotone sequence in $\mathbb{R}$ has a limit in $\overline{\mathbb{R}}$.

    Proof. Suppose that $\{a_n\}_{n\in\mathbb{N}}$ is increasing i.e. $\forall n \in \mathbb{N}\colon\ a_n \leqslant a_{n+1}$. Then for all $k \geqslant n$, $a_k \geqslant a_n$ and so $\inf_{k\geqslant n} a_k = a_n$, and $\sup_{k\geqslant n} a_k = \lim_{k\to\infty} a_k$ from which we conclude that
$$\limsup_{n\to\infty} a_n = \inf_n \sup_{k\geqslant n} a_k = \inf_n \lim_{k\to\infty} a_k = \lim_{k\to\infty} a_k$$
$$\liminf_{n\to\infty} a_n = \sup_n \inf_{k\geqslant n} a_k = \sup_n a_n = \lim_{k\to\infty} a_k$$
    An analogous argument holds for $\{a_n\}_{n\in\mathbb{N}}$-decreasing. ∎

[5] The lim sup of a sequence is the notion of the tightest eventual upper bound i.e. how low the sequence's upper bound is in the long run (as $n \to \infty$)
$$\limsup_{n\to\infty} a_n := \inf_n \left(\sup_{k\geqslant n} a_k\right).$$
We do this by taking the supremum of each tail to capture the highest the sequence goes from that starting point. The sequence of suprema tails is a decreasing sequence. Then we take the infimum over all starting points to find the tightest bound that eventually holds.

[6] These are sequences in $\mathbb{R}$ that we would say diverge to infinity.

- As an application, we'll be using monotone sequences of "simple" functions to approximate "measurable" functions.

2.5 Measures on $\mathcal{F}$

With the definition of a σ-algebra in hand, we can define the class of set functions called measures, and afterwards a particular type of measure — the one we desire — called a probability measure $\mathbb{P}$ on $\mathcal{F}$ (over Ω).

It's customary to denote the base space as X for a general measure. We call the pair $(X, \mathcal{F})$ a **measurable space**, and each $A \in \mathcal{F}$ is called a **measurable set**.

Definition 2.5.1 Let $(X, \mathcal{F})$ be a measurable space. A set function $\mu\colon \mathcal{F} \to [0, +\infty]$ is called a **measure** on $\mathcal{F}$ if:

(i) $\mu(\emptyset) = 0$

(ii) For any pairwise disjoint collection $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$:

$$\mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) = \sum_{i\in\mathbb{N}} \mu(A_i).$$

Property (ii) is called **σ-additivity** (or **countable additivity**). The triple $(X, \mathcal{F}, \mu)$ is called a **measure space**.

Proposition 2.5.2 (Properties of μ) A measure $\mu\colon \mathcal{F} \to [0, +\infty]$ is:

- **Finitely additive**:
  If $A_1, \ldots, A_n$ are pairwise disjoint, then $\mu\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mu(A_i)$.
- **Monotone**:
  If $A, B \in \mathcal{F}$ and $A \subseteq B$, then $\mu(A) \leqslant \mu(B)$.
- **Countably sub-additive**:
  For any $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$:
  $$\mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big) \leqslant \sum_{i\in\mathbb{N}} \mu(A_i).$$

Proof.

- Take $A_i = \emptyset$ for $i > n$ and appeal to μ being σ-additive.
- $B = A \sqcup (B \setminus A)$ and finite additivity gives us $\mu(B) = \mu(A) + \mu(B \setminus A) \geqslant \mu(A)$.

The final assertion will use a technique that's repeated so often, I'm dedicating a lemma to it.

Lemma 2.5.3 (**Disjointification**) From any sequence $\{A_n\}_{n\in\mathbb{N}} \subseteq 2^X$, one can construct a pairwise disjoint sequence $\{B_n\}_{n\in\mathbb{N}} \subseteq 2^X$ defined by $B_1 = A_1$ and for $n \geqslant 2$:

$$B_n = A_n \setminus \Big(\bigcup_{i=1}^{n-1} A_i\Big)$$

with the same union

$$\bigcup_{n\in\mathbb{N}} A_n = \bigsqcup_{n\in\mathbb{N}} B_n.$$

Proof. Pairwise disjointness should be clear from the definition: Let $n > m$. Then B_n contains all elements of A_n that haven't already been included in any of the earlier A_i with $i < n$. In particular, B_m is a subset of the union that's been removed from A_n in order to form B_n.

For the equal union statement, one direction of inclusion is clear. Each $B_n \subseteq A_n$ and so $\cup_{n\in\mathbb{N}} B_n \subseteq \cup_{n\in\mathbb{N}} A_n$. For the reverse inclusion, let $x \in \cup_{n\in\mathbb{N}} A_n$. Let k be the smallest index for which $x \in A_k$. This means that x is not an element of all the former A_i for $i < k$ i.e. $x \notin \cup_{n=1}^{k-1} A_n$.

$$\therefore\ x \in A_k \setminus \Big(\bigcup_{n=1}^{k-1} A_n\Big) =: B_k \subseteq \bigcup_{n\in\mathbb{N}} B_n.$$

†

- Let $B_1 = A_1$, and for $n > 1$ define $B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right)$. The B_n are pairwise disjoint and for all n: $\bigcup_1^n B_i = \bigcup_1^n A_i$. By monotonicity,

$$\mu\Big(\bigcup_1^\infty A_i\Big) = \mu\Big(\bigsqcup_1^\infty B_i\Big) = \sum_1^\infty \mu(B_i) \leqslant \sum_1^\infty \mu(A_i).$$

■

Definition 2.5.4

- If $\mu(X) < \infty$, then μ is called **finite**.
- If there exists an at most countably infinite sequence $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$ s.t. each $\mu(A_i) < \infty$ and

$$X = \bigcup_{i\in\mathbb{N}} A_i,$$

  then we call μ a **σ-finite** measure.
- A set $N \in \mathcal{F}$ is called a **μ-null set** if $\mu(N) = 0$.
- A statement is true **almost everywhere** (abbreviated **a.e.**), or true **for μ-almost every x** (abbreviated $\forall_\mu x$), if the set on which it doesn't hold is a μ-null set.
- A set $F \subseteq X$ is called **μ-negligible** if it's a subset of a μ-null set.
- A measure μ is called **complete** if every μ-negligible set $F \subseteq X$ is measurable i.e

$$F \subseteq N,\ \mu(N) = 0 \implies F \in \mathcal{F}.$$

  Note that, in general, a subset of a μ-null set (i.e. a μ-negligible set) need not be measurable — the underlying σ-algebra encodes what is measurable and what isn't. It's a peculiar fact but motivates the following:

2.5.1 Sub-σ-algebras and Subspace Measures

Definition 2.5.5 Let $(X, \mathcal{F})$ be a measurable space. We call $\mathcal{G} \subseteq \mathcal{F}$ a **sub-σ-algebra of $\mathcal{F}$** if $\mathcal{G}$ is a σ-algebra on X in its own right.

Lemma 2.5.6 For any subset $D \subseteq X$, the collection

$$\mathcal{F}|_D := \{A \cap D \colon A \in \mathcal{F}\}$$

defines a σ-algebra of subsets of D called the **trace of $\mathcal{F}$ on D**, or **subspace σ-algebra of subsets of D**.

A subtle distinction is that $\mathcal{F}|_D$ is a σ-algebra on D, not on X. Thus, $\mathcal{F}|_D$ is not in general a sub-σ-algebra of $\mathcal{F}$. However, if D is measurable ($D \in \mathcal{F}$) then $\mathcal{F}|_D$ is a sub-σ-algebra of $\mathcal{F}$.

Proof.

- $\emptyset = \emptyset \cap D \in \mathcal{F}|_D$
- Suppose that $A \in \mathcal{F}|_D$ i.e. $\exists B \in \mathcal{F}$ s.t. $A = B \cap D$. Then

$$\begin{aligned} D \setminus A &= D \setminus (B \cap D) \\ &= \underbrace{(X \setminus B)}_{\in \mathcal{F}} \cap D \in \mathcal{F}|_D \end{aligned}$$

- Let $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}|_D$. Then $\exists \{B_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ s.t. $\forall n\colon A_n = B_n \cap D$. Then, we have that

$$\bigcup_{n\in\mathbb{N}} A_n = \underbrace{\Big(\bigcup_{n\in\mathbb{N}} B_n\Big)}_{\in \mathcal{F}} \cap D \in \mathcal{F}|_D.$$

■

Proposition 2.5.7 If we suppose further that $(X, \mathcal{F}, \mu)$ is a measure space, and that $D \subseteq X$ is a measurable subset ($D \in \mathcal{F}$), then the restriction of μ to D, denoted by $\mu|_D\colon \mathcal{F}|_D \to [0, +\infty]$ and defined for any $A \in \mathcal{F}|_D$ by $\mu|_D(A) = \mu(A)$, is a measure on $\mathcal{F}|_D$.

Proof.

- Since $\mathcal{F}|_D \ni \emptyset$, $\mu|_D(\emptyset) = \mu(\emptyset) = 0$.
- Let $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}|_D$ be a pairwise disjoint collection.

$$\mu|_D\Big(\bigsqcup_{n\in\mathbb{N}} A_n\Big) = \mu\Big(\bigsqcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n) = \sum_{n\in\mathbb{N}} \mu|_D(A_n)$$

■

Definition 2.5.8 We call $\mu|_D$ a **subspace measure on** D.

2.5.2 Completion of Measure

One "completes a measure space" by "adding" all μ-negligible sets to the underlying σ-algebra.

Theorem 2.5.9 Let $(X, \mathcal{F}, \mu)$ be a measure space and $\mathcal{N} = \{F \subseteq X\colon \exists N \in \mathcal{F} \text{ with } \mu(N) = 0, \text{ and } F \subseteq N\}$ denote the collection of all μ-negligible sets. Then

$$\overline{\mathcal{F}} = \{A \cup F\colon A \in \mathcal{F},\ F \in \mathcal{N}\}.$$

is a σ-algebra, and there exists an extension $\overline{\mu}\colon \overline{\mathcal{F}} \to [0, +\infty]$ of μ to $\overline{\mathcal{F}}$ defined by $\overline{\mu}(A \cup F) = \mu(A)$ s.t.

(i) $\overline{\mu}$ is a measure,

(ii) $\overline{\mu}$ is the unique extension to $\overline{\mathcal{F}}$,

(iii) $\overline{\mu}$ is complete.

Moreover, $\overline{\mathcal{F}}$ is the smallest σ-algebra containing $\mathcal{F}$ on which μ extends to a complete measure.

Why is it useful to complete a measure?

It seems to be a construction that avoids the pathology of a measure not respecting one's intuitive notion of 'negligibility' i.e. we certainly expect that if a set N is μ-null, then any subset $F \subseteq N$ should also be measurable (with measure zero) since it's a smaller part of something we can already measure. I believe there are more ramifications later on when discussing functions that differ on μ-null sets.

Proof. [7]

$\overline{\mathcal{F}}$ is a σ-algebra:

- Contains the empty set:
  Every μ-null set $N \in \mathcal{F}$ is μ-negligible (since $N \subseteq N$). In particular, $\mu(\emptyset) = 0$ so $\emptyset \in \mathcal{F} \cap \mathcal{N}$. Therefore, $\emptyset = \emptyset \cup \emptyset \in \overline{\mathcal{F}}$.
- Closure under countable unions:
  Since $\mathcal{F}$ is closed under countable unions, it remains to show that so too is $\mathcal{N}$. Let $\{F_i\}_{i\in\mathbb{N}}$ i.e. $\exists\{N_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$ s.t. $\forall i \in \mathbb{N}$: $F_i \subseteq N_i$ and $\mu(N_i) = 0$. We wish to demonstrate that $F \coloneqq \bigcup_{i\in\mathbb{N}} F_i \in \mathcal{N}$. Certainly, $N \coloneqq \bigcup_{i\in\mathbb{N}} N_i \in \mathcal{F}$,
  $$\mu(N) = \mu\Big(\bigcup_{i\in\mathbb{N}} N_i\Big) \leqslant \sum_{i\in\mathbb{N}} \mu(N_i) = 0,$$
  and $\forall i \in \mathbb{N}$: $F_i \subseteq N_i \implies F \subseteq N$. Therefore, $F \in \mathcal{N}$.
- Closure under complementation:
  Let $(A \cup F) \in \overline{\mathcal{F}}$ (i.e. $A \in \mathcal{F}$, $F \in \mathcal{N}$). We wish to show that $(A \cup F)^{\mathsf{c}} \in \overline{\mathcal{F}}$.
  $$(A \cup F)^{\mathsf{c}} = A^{\mathsf{c}} \cap F^{\mathsf{c}}$$
  Since F is μ-negligible, there exists some $N \in \mathcal{F}$ with $\mu(N) = 0$ and $F \subseteq N$. Using the inclusion, we can write F^{c} as $N^{\mathsf{c}} \cup (N \setminus F)$. Therefore,
  $$\begin{aligned}(A \cup F)^{\mathsf{c}} &= A^{\mathsf{c}} \cap F^{\mathsf{c}} \\ &= A^{\mathsf{c}} \cap (N^{\mathsf{c}} \cup (N \setminus F)) \\ &= \underbrace{(A^{\mathsf{c}} \cap N^{\mathsf{c}})}_{\in \mathcal{F}} \cup \underbrace{(A^{\mathsf{c}} \cap (N \setminus F))}_{\subseteq N}.\end{aligned}$$
  Thus, $(A \cup F)^{\mathsf{c}} \in \overline{\mathcal{F}}$.

Is $\overline{\mu}$ well-defined?

We wish to show that the value of $\overline{\mu}$ doesn't depend on the representative of the equivalence class of sets that are equal modulo μ-negligibles i.e. that if $A_1 \cup F_1 = A_2 \cup F_2$ (with $A_1, A_2 \in \mathcal{F}$ and $F_1, F_2 \in \mathcal{N}$) then $\overline{\mu}(A_1 \cup F_1) = \overline{\mu}(A_2 \cup F_2)$.
Note that since $F_1, F_2 \in \mathcal{N}$, $\exists N_1, N_2 \in \mathcal{F}$ s.t. $\mu(N_1) = \mu(N_2) = 0$ and $F_1 \subseteq N_1$, $F_2 \subseteq N_2$. Thus,

$$A_1 \subseteq A_1 \cup F_1 = A_2 \cup F_2 \subseteq A_2 \cup N_2$$

from which we conclude that

$$\begin{aligned}\overline{\mu}(A_1 \cup F_1) &\coloneqq \mu(A_1) \\ &\leqslant \mu(A_1 \cup N_2) \quad \text{by monotonicity} \\ &\leqslant \mu(A_1) + \mu(N_2) \quad \text{by sub-additivity of } \mu \\ &= \mu(A_2) =: \overline{\mu}(A_2 \cup F_2).\end{aligned}$$

The reverse inequality follows similarly from $A_2 \subseteq A_2 \cup F_2 = A_1 \cup F_1 \subseteq A_1 \cup N_1$.

[7] Folland makes a note that we can make a simplifying assumption that $A \cup N = \emptyset$. I imagine this is to split up the measurable and null parts to obtain a canonical representation for elements of $\overline{\mathcal{F}}$.

Is $\overline{\mu}$ an honest to goodness extension of μ to $\mathcal{F}$?

For any $A \in \mathcal{F}$, $A = A \cup \emptyset$ and $\overline{\mu}$ is well-defined so

$$\overline{\mu}(A) = \overline{\mu}(A \cup \emptyset) \coloneqq \mu(A).$$

(i) Is $\overline{\mu}$ a measure?

- Since $\emptyset \in \mathcal{F}$, $\overline{\mu}(\emptyset) = \mu(\emptyset) = 0$.
- Let $\{A_i \cup F_i\}_{i\in\mathbb{N}} \subseteq \overline{\mathcal{F}}$ be a collection of pairwise disjoint sets. Then

$$\begin{aligned}\overline{\mu}\Big(\bigcup_{i\in\mathbb{N}}(A_i \cup F_i)\Big) &= \overline{\mu}\Big(\underbrace{\Big(\bigcup_{i\in\mathbb{N}} A_i\Big)}_{\in\mathcal{F}} \cup \underbrace{\Big(\bigcup_{i\in\mathbb{N}} F_i\Big)}_{\in\mathcal{N}}\Big)\\ &\coloneqq \mu\Big(\bigcup_{i\in\mathbb{N}} A_i\Big)\\ &= \sum_{i\in\mathbb{N}} \mu(A_i)\\ &=: \sum_{i\in\mathbb{N}} \overline{\mu}(A_i \cup F_i)\end{aligned}$$

(ii) Is $\overline{\mu}$ the unique extension of μ to $\overline{\mathcal{F}}$?
Suppose that ν is another extension of μ to $\mathcal{F}$. Then, for any $A \cup F \in \overline{\mathcal{F}}$:

$$\nu(A \cup F) = \mu(A) =: \overline{\mu}(A \cup F).$$

(iii) Is $\overline{\mu}$ complete? We wish to show that every $\overline{\mu}$-negligible set L is an element of $\overline{\mathcal{F}}$.

- Since L is $\overline{\mu}$-negligible i.e. $L \in \mathcal{N}_{\overline{\mu}}$, there exists some $\overline{\mu}$-null set $M \in \overline{\mathcal{F}}$ s.t. $L \subseteq M$.
- M is of the form $M = A \cup F$ for some $A \in \mathcal{F}$ and $F \in \mathcal{N}_\mu$.
- Since F is μ-negligible, $\exists N \in \mathcal{N}_\mu$ s.t. $\mu(N) = 0$ and $F \subseteq N$.
- Thus, $M = (A \cup F) \subseteq (A \cup N)$.
- Note that $0 = \overline{\mu}(M) = \overline{\mu}(A \cup F) \coloneqq \mu(A)$ so A is a μ-null set, and in particular an element of $\mathcal{N}_\mu \cap \mathcal{F}$. We've already shown that $\mathcal{N}_\mu$ (which we denoted by $\mathcal{N}$) is closed under countable unions so $M \subseteq (A \cup N) \in \mathcal{N}_\mu$.
- Thus, $L \subseteq M \subseteq (A \cup N) \in \mathcal{N}_\mu$ and so $L \in \mathcal{N}_\mu$.
- Then we can write $L = \emptyset \cup L$, where $\emptyset \in \mathcal{F}$ and $L \in \mathcal{N}_\mu$, so $L \in \overline{\mathcal{F}}$.

■

Definition 2.5.10

- $\overline{\mathcal{F}}$ is called the **completion of $\mathcal{F}$ with respect to μ**.
- $\overline{\mathcal{F}}$ is the **join** $\mathcal{F} \vee \mathcal{N} = \sigma(\mathcal{F} \cup \mathcal{N})$ of $\mathcal{F}$ and the collection of negligible sets $\mathcal{N}$.

2.6 Why Not Always $\mathcal{F} = 2^X$?

So far we've considered measures μ on an arbitrary σ-algebra $\mathcal{F}$. Why have we not simply taken the power set 2^X as our σ-algebra each time? Let's get as many measurable sets in there as possible, right?

If our space X is at most countably infinite, then there are no issues with taking $\mathcal{F} = 2^X$. However, it's very often the case that one is interested in uncountably infinite spaces like $X = \mathbb{R}$ (or $\overline{\mathbb{R}}$). The prototypical counterexample is the attempt to define a measure that generalises the notion of length of an interval $I \subseteq \mathbb{R}$ on the entirety of $2^{\mathbb{R}}$:

Claim It is impossible to demand that a set function μ satisfies all 4 of the following conditions:

(μ.1) μ is defined on all of $2^{\mathbb{R}}$ i.e. $\mu\colon 2^{\mathbb{R}} \to [0, +\infty]$.

(μ.2) $\mu((a, b]) = b - a$

(μ.3) μ is translationally invariant i.e. if $A \subseteq \mathbb{R}$, then for every $x \in \mathbb{R}$:

$$\mu(A) = \mu(A + x),$$

where $A + x \coloneqq \{a + x\colon a \in A\}$,

(μ.4) λ is σ-additive i.e. for any pairwise disjoint collection $\{A_i\}_{i\in\mathbb{N}} \subseteq 2^{\mathbb{R}}$:

$$\mu\Big(\bigsqcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

Conditions (μ.2–4) are non-negotiable for a function that's supposed to capture length — the length of an interval is the difference between their endpoints, a set's length doesn't change if you translate it, and the length of a disjoint sum of intervals is simply the sum of their individual lengths.

This leaves (μ.1), the existence of a set to which one cannot assign a notion of "size" with such a measure μ. One such example is called a Vitali set. We construct it as follows:

2.6.1 A Non-Measurable Set (Vitali Set)

Let's define an equivalence relation $\sim$ on $\mathbb{R}$ by $x \sim y \equiv x - y \in \mathbb{Q}$. Denote by $[x]$ the equivalence class of x. It's clear that $\mathbb{R}/\sim$ is uncountable. Now we form the set V by selecting[8] a representative from each element of Λ s.t. each representative is in $(0, 1)$. Therefore, $V \subseteq (0, 1)$. Consider two rational translates $V + p$ and $V + q$ of V. We claim that any two rational translates of V are either equal or disjoint.

Proof. Suppose that they aren't disjoint. If this implies they are equal then the claim has been proven. Let p and q be rational, and $x \in (V + p) \cap (V + q)$. Then

$$\begin{cases} x = \alpha + p, & \text{where } \alpha \in V \\ x = \beta + q, & \text{where } \beta \in V \end{cases}$$

This implies that $\alpha - \beta = q - p \in \mathbb{Q}$ i.e. $\alpha \sim \beta$. Because we constructed V in such a way that there is only one representative (in $(0, 1)$) for each equivalence class, this tells us that $\alpha = \beta$ so $q - p = 0$ i.e. $p = q$. Thus, $V + p = V + q$. ◧

The above claim is equivalent to its contrapositive i.e. $V + p \neq V + q \implies (V + p) \cap (V + q) = \emptyset$. Now consider the disjoint collection of translates $\{V + q\}_{q\in\mathbb{Q}\cap(-1,1)}$. Their union is clearly contained

[8] This supposedly relies on the Axiom of Choice.

in $(-1, 2)$ and it follows that:

$$3 = \mu((-1,2)) \geqslant \mu\Big(\bigsqcup_{\substack{q\in\mathbb{Q}\\ q\in(-1,1)}} (V+q)\Big) \quad \text{by monotonicity}$$
$$\overset{(4)}{=} \sum_{q\in\mathbb{Q}\cap(-1,1)} \mu(V+q)$$
$$\overset{(3)}{=} \sum_{q\in\mathbb{Q}\cap(-1,1)} \mu(V).$$

Since our infinite sum of a constant is finite, $0 = \mu(V) = \mu(V+q)$ for every $q \in \mathbb{Q}\cap(-1,1)$.

$$\therefore\ \mu\Big(\bigsqcup_{q\in\mathbb{Q}\cap(-1,1)} (V+q)\Big) = 0.$$

Now we make the following

Claim

$$(0,1) \subseteq \bigsqcup_{\substack{q\in\mathbb{Q}\\ q\in(-1,1)}} (V+q).$$

From this, it will follow that

$$1 = \mu((0,1)) \leqslant \mu\Big(\bigsqcup_{\substack{q\in\mathbb{Q}\\ q\in(-1,1)}} (V+q)\Big) = 0$$

which is a contradiction.

Proof of claim. Let $x \in (0,1)$. We wish to show that x is a rational translate of some element of V i.e. that $\exists \alpha \in V$ and $\exists q \in \mathbb{Q}\cap(-1,1)$ s.t. $x = \alpha + q$.

Since $x \in \mathbb{R}$ it is certainly a member of an equivalence class $[\alpha] \in \mathbb{R}/\sim$ for some $\alpha \in (0,1)$ i.e. $x - \alpha = q$ for some $\mathbb{Q}$. In particular, since $x \in (-1,1)$, then $x - \alpha \in (-1,1)$ and so $q \in \mathbb{Q}\cap(-1,1)$ i.e. we've written $x = \alpha + q$ in the desired form. ■

Therefore, there does not exist
such a set function on **all of** $2^{\mathbb{R}}$.

2.6.2 Damage Control

We'll continue using the example of μ defined on $2^{\mathbb{R}}$, and modify the setup (i.e. conditions (μ.1–4)) in order to avoid pathological sets that cannot be measured in the general case.

Since there exist subsets that aren't measurable by our most natural notion of length that generalises $\mu((a,b]) = b - a$, we should **either**:

1. modify only (μ.1) by **restricting** the domain of μ to a **proper** subset of $2^{\mathbb{R}}$ — a σ-algebra of sets to which we can meaningfully assign size, **or**

2. relax only (μ.4) in the sense that μ is still defined on $2^{\mathbb{R}}$, is finitely additive, but we only demand that $\mu(\sqcup_{i\in\mathbb{N}} E_i) = \sum_{i\in\mathbb{N}} \mu(E_i)$ holds for **certain collections of sets**.

1. The first approach will be to consider a natural σ-algebra that respects[9] the underlying structure of the space. In our case, $\mathbb{R}$ is equipped with a (standard) topology comprising of sets (that we call open) which give us a way to discuss locality (neighbourhoods), continuity, and other concepts in analysis. These are the "nice" sets in $\mathbb{R}$ and we would like for them (and the countable set-theoretic compositions thereof) to be measurable.

[9] By containing it, and hence retaining any desirable properties of the underlying structure.

Definition 2.6.1 The **Borel σ-algebra over $\mathbb{R}$**, denoted $\mathcal{B}_{\mathbb{R}} \subseteq 2^{\mathbb{R}}$, is generated by[10] the standard topology $\mathcal{T}_{\text{std.}}$ on $\mathbb{R}$ i.e.

$$\mathcal{B}_{\mathbb{R}} = \sigma(\mathcal{T}_{\text{std.}})$$

More generally, for any topological space:

Definition 2.6.2 The **Borel σ-algebra $\mathcal{B}_X$ of a (topological) space $(X, \mathcal{T})$** is the σ-algebra generated by the collection of all open subsets $\mathcal{T}$ of X

$$\mathcal{B}_X = \sigma(\mathcal{T}).$$

Most of the measures of interest in these notes will be defined on $\mathcal{B}_X$. Such measures are called **Borel measures**.

2. The second approach will be the topic of **Chapter 3**. As opposed to the first point which is quite descriptive, this approach will be more constructive:

   - One starts by defining a set function, with some ideal properties like σ-additivity, on a **simple sub-collection**[11] $\mathcal{S} \subseteq 2^X$.
   - Then we extend this set function as much as we can to a larger collection — a σ-algebra generated by the original collection $\mathcal{S}$.
   - Hopefully this extension is unique and those nice properties like σ-additivity carry over.

   Example 2.6.3 As a spoiler for what's to come, we'll see that for the example of constructing the Lebesgue measure λ a by-product of the theory is that λ is a complete, unique measure on $\mathcal{L}$ (the so-called collection of **Lebesgue-measurable sets**) s.t. $\mathcal{L} \supseteq \mathcal{B}_{\mathbb{R}}$. In fact, $\mathcal{L}$ is the completion of $\mathcal{B}_{\mathbb{R}}$ with respect to λ i.e. every set $A \in \mathcal{L}$ is of the form $A = B \cup F$ where $B \in \mathcal{B}_{\mathbb{R}}$ and[12] $F \in \mathcal{N}_\lambda$.

Now we take a brief intermission to discuss probability measures before returning to the above at the start of the next chapter.

2.7 Probability Measures

Definition 2.7.1 A measure $\mu \colon \mathcal{F} \to [0, +\infty]$ with the additional property that $\mu(X) = 1$ is called a **probability measure**.

- We often denote such a measure by $\mathbb{P}$ instead of μ, and the underlying space X by Ω.
- By monotonicity, every $A \subseteq X$ satisfies $0 \leqslant \mathbb{P}(A) \leqslant \mathbb{P}(X) = 1$ so $\mathbb{P}$ may be written as $\mathbb{P} \colon \mathcal{F} \to [0, 1]$.
- The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a **probability space**.

[10] Equivalently, $\mathcal{B}_{\mathbb{R}}$ is generated by the following families: $\{(a, b) : a, b \in \mathbb{R}\}$, $\{[a, b] : a, b \in \mathbb{R}\}$, $\{(a, \infty) : a \in \mathbb{R}\}$, and $\{[a, \infty) : a \in \mathbb{R}\}$.

[11] This simple collection is what was referred to earlier when we said the σ-additivity 'holds for certain collections of sets' — formalised later in **Definition 3.1.1**.

[12] Recall that $\mathcal{N}_\lambda$ is the collection of λ-negligible sets i.e. $\mathcal{N}_\lambda = \{F \subseteq \mathbb{R} \colon \exists N \in \mathcal{F} \text{ with } \lambda(N) = 0 \text{ and } F \subseteq N\}$.

2.7.1 Defining Ω and Counting Subsets

Correctly defining the outcome space of an experiment is imperative — all subsequent calculations depend on it. **If the outcome space is finite**, one can follow the **sample point method** to define a discrete probability measure and compute the probabilities of events:

1. List all finitely many elementary events (or atoms) $\{\omega_i\}_{i\in I} \subseteq \Omega$ where $|I|$ is finite.
2. Assign "reasonable" non-negative probabilities p_i to each $\{\omega_i\}$ such that all the p_i sum to 1.
3. Events $A \subseteq \Omega$ are then defined as disjoint unions of elementary events i.e.

$$A = \bigsqcup_{i\in I} \{\omega_i\}$$

4. Calculate the probability of A as the sum

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(\omega_i).$$

Example For the random experiment of rolling a fair die, $\Omega = \{1,2,3,4,5,6\}$, $\mathcal{F} = 2^\Omega$ and each outcome is equally likely for a single roll i.e. $\forall \omega \in \Omega\colon \mathbb{P}(\{\omega\}) = 1/6$. Since any $A \in \mathcal{F}$ may be written as a disjoint union of elementary events, we have that

$$\mathbb{P}(A) = \mathbb{P}\Big(\bigsqcup_{\omega\in A} \{\omega\}\Big) = \sum_{\omega\in A} \mathbb{P}(\{\omega\}) \quad \text{by additivity.}$$

When it's inconvenient to list all the elements of the outcome space, we can rely on ~~a computer~~ the techniques of combinatorial analysis to determine the number of elements in a particular subset.

Counting

Theorem 2.7.2 (**Fundamental Principle of Counting**) If some procedure can be performed in n_1 ways, and if, following this procedure, a second procedure can be performed in n_2 different ways, ..., and finally the k^{th} procedure can be performed in n_k different ways; then the number of ways the procedures can be performed in the order indicated is the product $n_1 \cdot \ldots \cdot n_k$.

Permutations

An arrangement of a set of n objects in a given order is called a **permutation**[13] of the objects (taken all at a time).

An arrangement of any $r \leqslant n$ of these objects in a given order is called an ***r*-permutation** (or a permutation of n objects taken r at a time). The number of permutations of n objects taken r at a time is denoted $P(n,r)$. The first element in an r-permutation of n objects may be chosen in n different ways. The second can be chosen in $n-1$ different ways. Proceeding inductively, the r^{th} element may be chosen in $n-(r-1)$ ways. Thus, by the fundamental theorem of counting

$$P(n,r) = n \cdot (n-1) \cdot \ldots \cdot (n-(r-1)) = \frac{n!}{(n-r)!}$$

[13] Equivalently, it's a one-to-one function from a finite set into itself.

Partitions

Suppose that we have n objects and we wish to partition them into k distinct groups containing $n_1, \ldots, n_k$ objects, respectively, where each object appears in exactly one group and $n = \sum_{j=1}^{k} n_j$. The number N of ways to partition the objects as such is

$$N = \binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdot \ldots \cdot n_k!}$$

Proof. N is the number of distinct arrangements of n objects in a row for a case in which rearrangement of the objects within a group does not count.

The number of distinct arrangements of the n objects, assuming all are distinct, is $P(n; n) = n!$. This is equal to the product of the number of ways of partitioning the n objects into k groups (ignoring order within groups) and the number of ways of ordering the $n_1, \ldots, n_k$ elements within each group:

$$n! = P(n; n) = N \cdot (n_1! \cdot \ldots \cdot n_k!)$$

■

$\binom{n}{n_1, \ldots, n_k}$ is known as a **multinomial coefficient** because such terms appear in the expansion of the multinomial term $y_1 + \ldots + y_k$ raised to the power of n.

Ordered Samples

When we, for example, choose one ball after another r times from an urn of n balls, we call the choice an **ordered sample** of size r.

- When sampling with replacement, there are n possible choices of ball each time so there are n^r different ordered samples of size r.
- When sampling without replacement, each ordered sample of size r is simply an r-permutation from a set of size n. Thus, the number of ordered samples without replacement is $P(n, r)$.

Combinations

A **combination** is a selection of one or more of the elements of a given set without regard to order. A combination of n objects taken r at a time is called an ***r*-combination**. The number of combinations of n objects taken r at a time is the number of subsets of size r. We denote this number by $C(n, r)$ or ${}^nC_r$. The selection of r objects from a total of n objects is equivalent to partitioning the n objects into $k = 2$ groups, the r selected and the $n - r$ remaining:

$$C(n, r) = \binom{n}{r, n-r} = \frac{n!}{r!(n-r)!} := \binom{n}{r}.$$

Permutations are for lists.
Combinations are for groups.

2.8 Independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Independence in probability is the idea that there can be separate, non-interacting sources of randomness.

Definition 2.8.1

- Two events $A, B \in \mathcal{F}$ are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

- A collection of events $\{A_i\}_{i\in I}$ are **pairwise independent** if every pair of events e.g. A_i and A_j are independent.
- A collection of events $\{A_i\}_{i\in I}$ are **mutually independent** if for every finite set of distinct indices $i_1, \ldots, i_n$ from I:

$$\mathbb{P}\Big(\bigcap_{j=1}^{n} A_{i_j}\Big) = \prod_{j=1}^{n} \mathbb{P}(A_{i_j}).$$

The concept of mutual independence generalises readily from sub-collections of finitely many events to sub-σ-algebras:

Definition 2.8.2 Let $\{\mathcal{F}_i\}_{i\in I}$ be a collection of sub-σ-algebras of $\mathcal{F}$. Then the σ-algebras $\{\mathcal{F}_i\}_{i\in I}$ are **mutually $\mathbb{P}$-independent** if for every finite subset $J \subseteq I$, and every choice of $A_j \in \mathcal{F}_j$ for $j \in J$,

$$\mathbb{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J} \mathbb{P}(A_j).$$

The mutual independence of events is a special case of the mutual independence of σ-algebras. Note that for each $i \in I$, $\mathcal{F}_i = \{\emptyset, A_i, A_i^{\mathsf{c}}, \Omega\}$ is the σ-algebra generated by A_i. The condition of mutual independence of the $\mathcal{F}_i$ then reduces to that of mutual independence of the A_i.

2.9 (Naïve) Conditional Probability

> I wouldn't say 'naïve' is a fitting word for the concept, but neither is 'elementary' for I consider everything in these notes to be at least somewhat sophisticated. I went for the lesser of two evils.

Dependence is the complementary idea to independence. We learn new things every day and update our beliefs when confronted with new evidence. *How* one should update one's beliefs that some event will occur given new evidence pertaining to said event is a central feature in the study of probability. This is formalised by the concept of conditional probability.

Let $A, B \in \mathcal{F}$. Suppose that we observe B with non-zero probability $\mathbb{P}(B) > 0$ and say we're interested in how likely A is to occur given that B has already occurred. We denote this by $\mathbb{P}(A \,|\, B)$.

We can appeal to the relative frequency interpretation of probability as a guiding light for the kind of expression to expect for $\mathbb{P}(A \,|\, B)$. Suppose that an experiment is repeated n times and on each trial, the occurrences of A and B are recorded — the numbers of which are $A(n)$ and $B(n)$ respectively. *Suppose further that we concern ourselves with only the outcomes for which B occurs with positive probability $\mathbb{P}(B) > 0$ unless stated otherwise.* The proportion of times that A occurs in these outcomes where B occurs can be written as

$$\frac{(A\cap B)(n)}{B(n)} = \frac{\frac{(A\cap B)(n)}{n}}{\frac{B(n)}{n}} \xrightarrow{n\to\infty} \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)},$$

where the numerator and denominator are thought of as the long-running relative frequencies of the events $A \cap B$ and B respectively. Thus, the relationship

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

defines the conditional probability of any A given the event B.

In the case that B is $\mathbb{P}$-null, so is $A \cap B$ and the relationship $0 * \mathbb{P}(A \,|\, B) = 0$ tells us nothing about how to determine $\mathbb{P}(A \,|\, B)$.

Since we're considering the fraction of outcomes where A occurs within the repeated trials where B has already occurred, one would naturally expect that $\mathbb{P}(B \,|\, B) = 1$. Since $\mathbb{P}$ is a probability measure on $\mathcal{F}$, it's an immediate consequence that so too is $\mathbb{P}(\cdot \,|\, B) \colon \mathcal{F} \to [0, 1]$, defined for all $A \in \mathcal{F}$ by

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

Definition 2.9.1 Given an event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the map $\mathbb{P}(\cdot \mid B) \colon \mathcal{F} \to [0,1]$ defined for all $A \in \mathcal{F}$ by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

is called the **naïve conditional probability on $\mathcal{F}$ given B**.

Corollaries

Corollary 2.9.2 (**Multiplication Law**) For any events A and B:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B).$$

Alternative Characterisation of Independence

Given $\mathbb{P}(\cdot \mid B) \colon \mathcal{F} \to [0,1]$ with $\mathbb{P}(B) > 0$, if A, B are independent, then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A).$$

This gives us an alternate characterisation of independence. Two events A and B are independent if the knowledge that one occurs gives us no information about whether the other does (or doesn't) i.e.

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \text{and} \quad \mathbb{P}(B \mid A) = \mathbb{P}(B).$$

It's possible to calculate an unconditional probability by conditioning on some events that make the calculation simpler:

Corollary 2.9.3 (**Law of Total Probability**) Let $B_1, \ldots, B_n$ be a partition of Ω. Then for any event $A \in \mathcal{F}$:

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

The choice of partition is important and can turn a complicated problem into smaller, simpler sub-problems.

Proof.

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(A \cap \Omega) \\
&= \mathbb{P}\left(A \cap \left(\bigsqcup_{i=1}^{n} B_i\right)\right) \\
&= \mathbb{P}\left(\bigsqcup_{i=1}^{n} (A \cap B_i)\right) \\
&= \sum_{i=1}^{n} \mathbb{P}(A \cap B_i) \qquad \text{since the } A \cap B_i \text{ are disjoint} \\
&= \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)
\end{aligned}$$

■

Bayes' Rule

Suppose that instead of finding $\mathbb{P}(A \mid B_i)$, we seek the probability of a "cause" B_j given an "effect" A. Bayes' rule is very useful in this case.

Theorem 2.9.4 (**Bayes' Rule**) For a partition $B_1, \ldots, B_n$ of Ω and some event A,

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Proof. From the definition of conditional probability

$$\begin{aligned}\mathbb{P}(B_j \mid A) &= \frac{\mathbb{P}(B_j \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\mathbb{P}(A)} \quad \text{by the multiplication law} \\ &= \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)} \quad \text{by the law of total probability.}\end{aligned}$$

■

Bayes' rule is the fundamental ingredient for a subjective approach to theories of evidence and learning. Its derivation is very simple but there's massive ongoing debate as to how it's used in practise.

According to this point of view, an individual's belief of some world event H can be coded into probabilities $\mathbb{P}(H)$ and given some evidence E, our beliefs are modified $\mathbb{P}(H \mid E)$.

Since $\mathbb{P}(E \mid H)$ is usually easier to calculate, Bayes rule comes in handy:

$$\mathbb{P}(H \mid E) = \frac{\mathbb{P}(E \mid H)\mathbb{P}(H)}{\mathbb{P}(E \mid H)\underbrace{\mathbb{P}(H)}_{\text{a prior}} + \mathbb{P}(E \mid H^{\mathsf{c}})\mathbb{P}(H^{\mathsf{c}})}$$

(Using Bayes' rule and the law of total probability can help update our beliefs based on observed evidence.)

CHAPTER 3

Constructing Measures

This chapter will explore a[1] systematic way to construct (probability) measures.

3.1 Terminology

Definition 3.1.1 Let $\mathcal{C}$, $\mathcal{C}_1$ and $\mathcal{C}_2$ be any collections of subsets of X s.t. $\mathcal{C}_1 \subseteq \mathcal{C}_2$.

- A set function $\nu\colon \mathcal{C} \to [0, +\infty]$ is:
  - **additive within $\mathcal{C}$** if for any $n \geqslant 1$ and any pairwise disjoint collection $\{A_i\}_{i=1}^n$ **s.t.**[2] $\bigsqcup_{i=1}^n A_i \in \mathcal{C}$, we have
  $$\nu\Big(\bigsqcup_{i=1}^n A_i\Big) = \sum_{i=1}^n \nu(A_i).$$
  - **σ-additive within $\mathcal{C}$** if we can replace n with ∞.
- Given any two set functions $\nu_i\colon \mathcal{C}_i \to [0, +\infty]$ for $i = 1, 2$, we say that **ν_2 is an extension of ν_1** if
  $$\forall A \in \mathcal{C}_1\colon \nu_2(A) = \nu_1(A).$$
  This is denoted by $\nu_2\big|_{\mathcal{C}_1} = \nu_1$.

3.2 Chapter Roadmap

- Recall the idea of covering a Euclidean subset and taking the limit to get a notion of area.

0. Introduce the notion of an outer measure $\mu^*\colon 2^X \to [0, +\infty]$.
   - Let $\mathcal{K} \supseteq \{\emptyset, X\}$ be a cover of X. Define a particular outer measure $\rho^*\colon 2^X \to [0, +\infty]$, from a set function $\rho\colon \mathcal{K} \to [0, +\infty]$, for any $A \subseteq X$ by:
   $$\rho^*(A) \coloneqq \inf\left\{\sum_{i=1}^\infty \rho(K_i)\colon \{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K},\ A \subseteq \bigcup_{i\in\mathbb{N}} K_n\right\}$$
   where $\rho(\emptyset) = 0$.

- Introduce the collection Σ of Carathéodory measurable sets with respect to any outer measure μ^*.
  - Show that Σ is a σ-algebra.
  - Show that the restriction of **any** outer measure $\mu^*\colon 2^X \to [0, +\infty]$ to Σ i.e. $\mu^*\big|_\Sigma$ is a measure.

1. Define a simple collection $\mathcal{S}$ (a semi-algebra) of subsets of X *that covers* X.
   - This will be the collection on which we understand how to prescribe some notion of size with a function — our goal being to extend such a function to a measure.

[1] The Riesz representation theorem seems to offer another way to define measures that I may include at a later date.

[2] $\mathcal{C}$ is any collection so it need not be closed under finite unions (nor countably infinite unions in the definition of σ-additivity) — that's why we demand that the unions are also in $\mathcal{C}$.

2. Introduce the algebra $\mathrm{Alg}(\mathcal{S})$ generated by $\mathcal{S}$, and provide an explicit representation for any $A \in \mathrm{Alg}(\mathcal{S})$:
$$A \in \mathrm{Alg}(\mathcal{S}) \iff \exists \{E_j\}_{j=1}^n \subseteq \mathcal{S} \text{ s.t. } A = \bigsqcup_{j=1}^n E_j.$$
3. Define a pre-pre-measure (additive and σ-additive function) μ_0 on $\mathcal{S}$, and a pre-measure on any algebra $\mathcal{A}$.
4. Show that an additive (resp. σ-additive) function on $\mathcal{S}$ can be uniquely extended to an additive (resp. σ-additive) function on $\mathrm{Alg}(\mathcal{S})$.
    - i.e. a pre-pre-measure on $\mathcal{S}$ can be extended uniquely to a pre-measure on $\mathrm{Alg}(\mathcal{S})$.
    - This depends on a few facts about continuity from below and above of σ-additive functions.
5. Then we prove that a pre-measure μ_0 on an algebra $\mathcal{A}$ can be extended to a measure $\mu^*\big|_\Sigma$, where Σ is a σ-algebra that contains $\mathcal{A}$, and the measure $\mu^*\big|_\Sigma$ is complete.
6. Under the assumption of σ-finiteness, the extended measure is unique.
    - The Monotone Class Theorem (a similar structure theorem to the π-λ Theorem) is useful here.
7. Finally, combine all of the above using the following:
    - Define μ^* as the aforementioned infimum (in 0.) but replace ρ with the pre-measure μ_0 on $\mathrm{Alg}(\mathcal{S})$.
        - μ_0 uniquely extends the pre-pre-measure $\nu \colon \mathcal{S} \to [0, \infty]$ to $\mathrm{Alg}(\mathcal{S})$.
    - Then it follows that $\mu = \mu^*\Big|_\Sigma$ is a complete and unique measure extending μ_0 (and hence ν) to Σ.

3.3 Approximation by Covering

It's helpful to recall the procedure used in calculus to define the area of a bounded region $E \subseteq \mathbb{R}^2$. Subdivide the plane into a collection $\mathcal{K} = \{K_n\}_{n \in \mathbb{N}}$ of (almost)[3] disjoint rectangles.

- Approximate the area of E from below by summing the areas of rectangles in the grid that are subsets of E — this is the inner area approximation.
- Approximate the area of E from above by summing the areas of rectangles in the grid that intersect E — this is the outer area approximation.

[3] The shared edge of adjacent rectangles will intersect but they have no Euclidean area so they don't count.

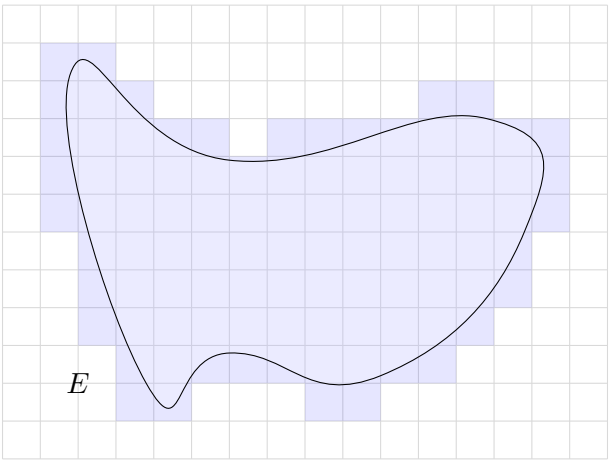


Figure 3.1: The inner and outer area approximations of E for a fixed grid size (each K_n has the same area).

Now we can vary over all possible available coverings in $\mathcal{K}$ to find the *inner area* and *outer area* of E:

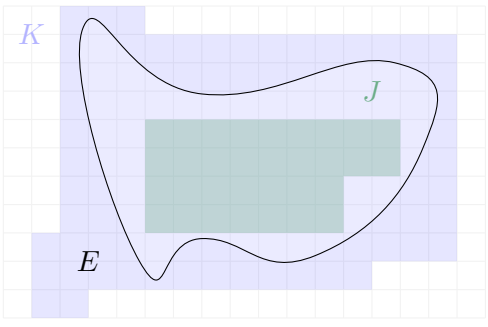
$$\operatorname{inn}(A) = \sup\left\{\operatorname{area}\Big(\bigcup_{i\in\mathbb{N}} K_i\Big) \colon \{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K},\ A \supseteq \bigcup_{i\in\mathbb{N}} K_i\right\}$$

$$\operatorname{out}(A) = \inf\left\{\operatorname{area}\Big(\bigcup_{i\in\mathbb{N}} K_i\Big) \colon \{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K},\ A \subseteq \bigcup_{i\in\mathbb{N}} K_i\right\}$$

If these two values are equal, the common value is the "area" of E. An astute observation is that one can characterise the inner area in terms of the outer area, thus we only need one of these concepts moving forward. I'll explain how we get such a relationship in the general case in **Section 3.4**, but the special case of Euclidean area offers a very simple manipulation:

Recall that the plane has been subdivided into disjoint rectangles and in this case, $\operatorname{area}(\,\cdot\,)$ is additive on $\mathcal{K}$ — this property is not something we'll be able to guarantee in the general case. Because of this, for $B \subseteq A$ we can express $\operatorname{area}(A \setminus B) = \operatorname{area}(A) - \operatorname{area}(B)$.

For notational convenience, let $K \coloneqq \sqcup_{i\in\mathbb{N}} K_i$ be any cover that contains E, and let J be a corresponding union of elements of $\mathcal{K}$ such that $K \setminus J$ covers $K \setminus E$.



The inner area of E can then be expressed as:

$$\begin{aligned}
&\operatorname{inn}(E)\\
&= \sup\{\operatorname{area}(J) \colon J \subseteq E\}\\
&= \sup\{\operatorname{area}(K \setminus (K \setminus J)) \colon K \setminus E \subseteq K \setminus J\}\\
&= \sup\{\operatorname{area}(K) - \operatorname{area}(K \setminus J) \colon K \setminus E \subseteq K \setminus J\}\\
&= \operatorname{area}(K) - \inf\{\operatorname{area}(K \setminus J) \colon K \setminus E \subseteq K \setminus J\}\\
&= \operatorname{out}(K) - \operatorname{out}(K \setminus E)
\end{aligned}$$

This establishes the following relationship:

$$\operatorname{out}(K) - \operatorname{out}(K \setminus E) = \operatorname{inn}(E) = \operatorname{out}(E) \quad \text{i.e.} \quad \operatorname{out}(K) = \operatorname{out}(E) + \operatorname{out}(K \setminus E).$$

3.3.1 Outer Measure

The abstract generalisation of outer area is *outer measure*. The set of rectangles we can use to cover a set are replaced by a more general concept:

Definition 3.3.1 A **covering of X** is a countable collection $\mathcal{K} = \{K_n\}_{n\in\mathbb{N}} \subseteq 2^X$ of sets such that

$$X \subseteq \bigcup_{n\in\mathbb{N}} K_n.$$

Definition 3.3.2 An (abstract) **outer measure** on X is a map $\mu^*\colon 2^X \to [0,+\infty]$ s.t.

(i) $\mu^*(\emptyset) = 0$

(ii) If $E \subseteq F$, then $\mu^*(E) \leqslant \mu^*(F)$.

(iii) If $\{E_i\}_{i\in\mathbb{N}} \subseteq 2^X$, then

$$\mu^*\Big(\bigcup_{i=1}^{\infty} E_i\Big) \leqslant \sum_{i=1}^{\infty} \mu^*(E_i).$$

Proposition 3.3.3 Let $\{\emptyset, X\} \subseteq \mathcal{K} \subseteq 2^X$, and $\rho\colon \mathcal{K} \to [0,+\infty]$ be a set function s.t. $\rho(\emptyset) = 0$. Then, for any $A \in 2^X$:

$$\rho^*(A) \coloneqq \inf\left\{\sum_{i=1}^{\infty} \rho(K_i)\colon \{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K},\ A \subseteq \bigcup_{i\in\mathbb{N}} K_i\right\}$$

is an example of an outer[4] measure on X.

Proof.

(i) We can cover $\emptyset$ by $\{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K}$ where $K_i = \emptyset$ for all $i \in \mathbb{N}$. It's clear that

$$\rho^*(\emptyset) \leqslant \sum_{i=1}^{\infty} \rho(\emptyset) = 0.$$

Any collection $\{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K}$ satisfies $\emptyset \subseteq K_i$ for all $i \in \mathbb{N}$ and so qualifies as a cover of $\emptyset$. Then the non-negativity of ρ implies that

$$\sum_{i=1}^{\infty} \rho(K_i) \geqslant 0.$$

In particular, the infimum $\rho^*(\emptyset)$ over all such coverings of $\emptyset$ is $\geqslant 0$. Therefore, $0 \leqslant \rho^*(\emptyset) \leqslant 0$.

(ii) Let $E \subseteq F \subseteq X$. We wish to show that $\rho^*(E) \leqslant \rho^*(F)$. Take any covering $\{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K}$ of F. Such a covering is also a cover of E. Note that the family of coverings of E is at least as large as the family of coverings of F. Thus,

$$\rho^*(E) \coloneqq \inf_{\substack{\text{covers}\\ \text{of } E}}\left\{\sum_{i\in\mathbb{N}} \rho(K_i)\right\} \leqslant \inf_{\substack{\text{covers}\\ \text{of } F}}\left\{\sum_{i\in\mathbb{N}} \rho(K_i)\right\} =\coloneqq \rho^*(F).$$

(iii) Assume that $\rho^*(E_i) < \infty$ for all i. Otherwise, the inequality we wish to prove

$$\rho^*\Big(\bigcup_{i\in\mathbb{N}} E_i\Big) \leqslant \sum_{i\in\mathbb{N}} \rho^*(E_i)$$

[4] If we chose to pursue inner measure instead, we would define something like

$$\rho_*(A) \coloneqq \sup\left\{\sum_{i=1}^{\infty} \rho(K_i)\colon \{K_i\}_{i\in\mathbb{N}} \subseteq \mathcal{K},\ A \supseteq \bigcup_{i\in\mathbb{N}} K_i\right\}.$$

is trivial (since the right-hand side $= +\infty$). Fix some $\varepsilon > 0$. Since the infimum $\rho^*(E_i)$ is finite by assumption for each i, we can find a covering that almost attains the infimum i.e. $\exists\{K_{ij}\}_{j\in\mathbb{N}} \subseteq \mathcal{K}$ s.t. $E_i \subseteq \bigcup_{j\in\mathbb{N}} K_{ij}$ and

$$\rho^*(E_i) \leqslant \sum_{j\in\mathbb{N}} \rho(K_{ij}) \leqslant \rho^*(E_i) + \frac{\varepsilon}{2^i}.$$

Letting i vary as well, the $\{K_{ij}\}_{i,j} \subseteq \mathcal{K}$ form a covering of the set E (since each $\{K_{ij}\}_j$ covers E_i).

$$\begin{aligned} \therefore\ \rho^*(E) \leqslant \sum_{(i,j)\in\mathbb{N}^2} \rho(K_{ij}) &= \sum_{i\in\mathbb{N}} \Big(\sum_{j\in\mathbb{N}} \rho(K_{ij}) \Big) \\ &\leqslant \sum_{i\in\mathbb{N}} \Big(\rho^*(E_i) + \frac{\varepsilon}{2^i} \Big) \\ &= \sum_{i\in\mathbb{N}} \rho^*(E_i) + \varepsilon. \end{aligned}$$

This inequality holds for all $\varepsilon > 0$ so let $\varepsilon \downarrow 0$ to conclude that ρ^* is σ-sub-additive.

■

The outer measure ρ^* defined above tells us that given a cover of X, and some set function assigning sizes to such sets in the cover, we can cover A in all the different ways possible, assign a number to the "size" of each combination, and then take the infimum of these sums. This lines up with our motivating intuition at the chapter's start — approximating area from outside and inside of a set.

3.4 Outer Measurability and Carathéodory Extension

Consider $E \subseteq X$. Suppose that $\mu^*(X) < \infty$, and let $E \subseteq X$. The *outer measure of* E, given by $\mu^*(E)$, is finite by monotonicity of μ^*.

Looking back at the specific case of area in $\mathbb{R}^2$, we derived the expression $\text{out}(K) = \text{out}(E) + \text{out}(K \setminus E)$ when the inner and outer measures of a bounded region $E \subseteq \mathbb{R}^2$ are equal. This turns out to be particularly important in the general case when characterising measurability:

For any set $E \subseteq X$, sub-additivity of an outer measure μ^* on 2^X gives us the inequality

$$\mu^*(X) \leqslant \mu^*(E) + \mu^*(X \setminus E).$$

We've assumed that $\mu^*(X)$ is finite, and so $\mu^*(X \setminus E)$ is finite by monotonicity (since $X \setminus E \subseteq X$). Thus, we may meaningfully subtract $\mu^*(X \setminus E)$ from both sides of the inequality to get

$$\mu^*(X) - \mu^*(X \setminus E) \leqslant \mu^*(E).$$

This is a lower bound for the outer measure of E. The LHS is analogous to the expression for the inner area in the plane — a quantity that serves as "dual" to the outer measure by virtue of subtracting the outer measure of E's absolute complement in X from the outer measure of X. Let's define it to be the inner measure $\mu_*(E)$ of E. Thus, our inequality is the very intuitive statement that the inner measure (which approximates a set from inside) is at most the outer measure (the approximation from the outside):

$$\mu_*(E) \leqslant \mu^*(E).$$

Once again, if the *inner* and *outer measures of* E are equal i.e.

$$\overbrace{\mu^*(X) - \mu^*(X \setminus E)}^{\mu_*(E)} = \mu^*(E)$$

then rearranging, which is permitted because $\mu^*(X) < \infty$, gives

$$\mu^*(X) = \mu^*(E) + \mu^*(X \setminus E). \tag{Cara$_X$}$$

This equation says that our set E splits X "nicely" without introducing any "extra" outer measure in the split.

This requirement of splitting nicely was generalised by Carathéodory from just splitting X (which required $\mu^*(X) < \infty$) to all subsets $A \subseteq X$:

Definition 3.4.1 A set $E \subseteq X$ is said to be **μ^*-measurable** (or **Carathéodory-measurable**) if it satisfies Carathéodory's splitting criterion i.e. that for every $A \subseteq X$:

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^{\mathsf{c}}).$$

Definition 3.4.2 (Terminological remarks)

- Note that $A \cap E^{\mathsf{c}} = A \setminus E$, and in (Cara_X) we note that $\mu^*(E) = \mu^*(X \cap E)$ so the expression in (Cara_X) is the special case of our Carathéodory criterion for just X.
- The Carathéodory criterion can be interpreted as E being μ^* measurable iff it has a sufficiently nice boundary, amenable to being covered by μ^*. The analogy I have in my head is that coastlines have fractal dimension, so any subset of the surface of the Earth's surface with part of a coastline as its boundary would not satisfy the criterion — since any splitting at a coastline would introduce extra measure. Maybe this is a poor analogy?

(!!) The "$\leqslant$" part of Carathéodory's criterion is just sub-additivity. Verifying the inequality in the reverse direction is sufficient for demonstrating a set is μ^*-measurable.

Theorem 3.4.3 (Carathéodory Extension) Let μ^* be an outer measure on X.

- The set of μ^*-measurable sets Σ is a σ-algebra, and
- the restriction $\mu := \mu^*\big|_\Sigma$ is a complete measure.

Proof.

- Σ is a σ-algebra over X:

  (1) For any $A \subseteq X$, $\mu^*(A \cap \emptyset) + \mu^*(A \cap \emptyset^{\mathsf{c}}) = \mu^*(\emptyset) + \mu^*(A \cap X) = \mu^*(A)$. Thus, $\emptyset \in \Sigma$.

  (2) Carathéodory's criterion is symmetric in A and A^{c} so $A \in \Sigma \implies A^{\mathsf{c}} \in \Sigma$.

  (3′) According to (!!), we need only verify the stated inequality. First, I'll do this for two sets $E, F \in \Sigma$. Let $A \subseteq X$. Then

$$\begin{aligned}
\mu^*(A) &\overset{E\in\Sigma}{=} \mu^*(A \cap E) + \mu^*(A \cap E^{\mathsf{c}}) \\
&\overset{F\in\Sigma}{=} \mu^*((A \cap E) \cap F) + \mu^*((A \cap E) \cap F^{\mathsf{c}}) \\
&\qquad + \mu^*((A \cap E^{\mathsf{c}}) \cap F) + \mu^*((A \cap E^{\mathsf{c}}) \cap F^{\mathsf{c}}) \\
&= \mu^*(A \cap (E \cap F)) + \mu^*(A \cap (E \cap F^{\mathsf{c}})) + \mu^*(A \cap (E^{\mathsf{c}} \cap F)) + \mu^*(A \cap (E \cup F)^{\mathsf{c}}) \\
&\geqslant \mu^*(A \cap ((E \cap F) \cup (E \cap F^{\mathsf{c}}) \cup (E^{\mathsf{c}} \cap F))) + \mu^*(A \cap (E \cup F)^{\mathsf{c}}) \quad \text{by sub-add} \\
&= \mu^*(A \cap (E \cup F)) + \mu^*(A \cap (E \cup F)^{\mathsf{c}})
\end{aligned}$$

  Thus, $E \cup F \in \Sigma$.

  This fact extends to closure under finite unions by proceeding inductively so Σ is an algebra.

  (3) Now consider a countable collection $\{E_i\}_{i\in\mathbb{N}} \subseteq \Sigma$. We wish to show that $\Sigma \ni E := \bigcup_{i\in\mathbb{N}} E_i$ i.e. that for any $A \subseteq X$:

$$\mu^*(A) \geqslant \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} E_i\Big)\Big) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} E_i\Big)^{\mathsf{c}}\Big)$$

  An intermediate factoid will let us construct a pairwise disjoint collection $\{F_i\}_{i\in\mathbb{N}}$ from the $\{E_i\}_{i\in\mathbb{N}}$ which has the same union, but facilitates simpler computation because we won't need to worry about overlapping sets:

Lemma 3.4.4 If $F_1, \ldots, F_n \in \Sigma$ are pairwise disjoint, then for any $A \subseteq X$:

$$\mu^*\Big(A \cap \Big(\bigcup_{i=1}^{n} E_i\Big)\Big) = \mu^*(A \cap E_1) + \ldots + \mu^*(A \cap E_n).$$

Proof. It's sufficient to prove for E_1 and E_2. Let $E_1 \cap E_2 = \emptyset$. Since $E_1 \in \Sigma$, $\forall A \subseteq X$:

$$\begin{aligned}
\mu^*(A \cap (E_1 \cup E_2)) &= \mu^*((A \cap (E_1 \cup E_2)) \cap E_1) + \mu^*((A \cap (E_1 \cup E_2)) \cap E_1^{\mathsf{c}}) \\
&= \mu^*((A \cap E_1 \cap E_1) \cup (A \cap E_1 \cap E_2)) \\
&\quad + \mu^*((A \cap E_1 \cap E_1^{\mathsf{c}}) \cup (A \cap \underbrace{E_1^{\mathsf{c}} \cap E_2}_{E_2 \subseteq E_1^{\mathsf{c}}})) \\
&= \mu^*(A \cap E_1) + \mu^*(A \cap E_2).
\end{aligned}$$

†

So now define $\{F_i\}_{i\in\mathbb{N}}$ by $F_1 = E_1$, and for $i > 1$:

$$F_i = E_i \setminus \Big(\bigcup_{n=1}^{i-1} E_i\Big).$$

Since Σ is an algebra, $\Sigma \ni \bigcup_{i=1}^{n} F_i$. Thus,

$$\begin{aligned}
\mu^*(A) &= \mu^*\Big(A \cap \bigcup_{i=1}^{n} F_i\Big) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{n} F_i\Big)^{\mathsf{c}}\Big) \\
&\overset{3.4.4}{=} \sum_{i=1}^{n} \mu^*(A \cap F_i) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{n} F_i\Big)^{\mathsf{c}}\Big) \\
&\geqslant \sum_{i=1}^{n} \mu^*(A \cap F_i) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} F_i\Big)^{\mathsf{c}}\Big) \quad \text{by monotonicity.}
\end{aligned}$$

This inequality holds for all n. Let $n \to \infty$ to yield

$$\begin{aligned}
\mu^*(A) &\geqslant \sum_{i=1}^{\infty} \mu^*(A \cap F_i) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} F_i\Big)^{\mathsf{c}}\Big) \\
&\geqslant \mu^*\Big(\bigcup_{i=1}^{\infty}(A \cap F_i)\Big) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} F_i\Big)^{\mathsf{c}}\Big) \quad \text{by } \sigma\text{-sub-additivity} \\
&= \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} F_i\Big)\Big) + \mu^*\Big(A \cap \Big(\bigcup_{i=1}^{\infty} F_i\Big)^{\mathsf{c}}\Big).
\end{aligned}$$

Therefore, Σ is a σ-algebra.

- $\mu := \mu^*\big|_{\Sigma}$ is a measure:

  Since we've already shown that $\mu^*(\emptyset) = 0$ and $\emptyset \in \Sigma$, it follows that $\mu(\emptyset) = 0$. All that remains to demonstrate is σ-additivity of $\mu := \mu^*\big|_{\Sigma}$ i.e. for any pairwise disjoint collection $\{E_i\}_{i\in\mathbb{N}} \subseteq \Sigma$:

$$\mu\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mu(E_i).$$

  We already know that μ^* is countably-sub-additive i.e.

$$\mu\Big(\bigcup_{i=1}^{\infty} E_i\Big) := \mu^*\Big(\bigcup_{i=1}^{\infty} E_i\Big) \leqslant \sum_{i=1}^{\infty} \mu^*(E_i) =: \sum_{i=1}^{\infty} \mu(E_i).$$

It remains to demonstrate the reverse inequality. Consider 3.4.4 with $A = X$:

$$\mu^*\Big(\bigcup_{i=1}^{\infty} E_i\Big) \geqslant \mu^*\Big(X \cap \Big(\bigcup_{i=1}^{n} E_i\Big)\Big) \overset{3.4.4}{=} \mu^*(X \cap E_1) + \ldots + \mu^*(X \cap E_n).$$

This holds for all $n \in \mathbb{N}$ and so the inequality follows.

- μ is a complete measure:

  We wish to show that any subset B of a μ-null set $N \in \Sigma$ is also an element of Σ. Let $A \subseteq X$.

  - Since μ^* is monotone, notice that $\mu^*(B) \leqslant \mu^*(N) = 0$.
  - Since $(A \cap B) \subseteq B$, its outer measure is also zero. Thus,

$$\mu^*(A \cap B) + \mu^*(A \cap B^{\mathsf{c}}) = \mu^*(A \cap B^{\mathsf{c}}) \leqslant \mu^*(A).$$

■

The restriction theorem above works for <u>any</u> outer measure μ^*.

It's good to know that the above theorem holds so generally. **However**, if we construct an outer measure μ^* on X from a set function ρ on a cover $\mathcal{K} \supseteq \{\emptyset, X\}$ of X, like we did with ρ^* in **Proposition 3.3.3**, then there's no guarantee that μ extends ρ. We need some extra structure for ρ and $\mathcal{K}$.

3.5 Refined Carathéodory Extension

In practice, one begins with a particular class of set functions ν (which will be referred to as *pre-pre-measures*) whose behaviour we understand on a simple collection of objects $\mathcal{S}$ (a *semi-algebra*) that covers X.

The following example inspires the subsequent definition:

Example Consider $X = \mathbb{R}$ and the collection of half-open intervals

$$\mathcal{S} = \{(a, b] \colon a, b \in \mathbb{R}\} \cup \{(a, \infty) \colon a \in \mathbb{R}\} \cup \{(-\infty, b] \colon b \in \mathbb{R}\} \cup \{\emptyset\}.$$

Note that this collection is closed under finite intersections i.e. all of the following are elements of $\mathcal{S}$:

- $(a, b] \cap (-\infty, c] = \begin{cases} \emptyset, & \text{if } c \leqslant a \\ (a, c], & \text{if } a < c < b \\ (a, b], & \text{if } c \geqslant b \end{cases}$
- $(a, b] \cap (c, \infty) = \begin{cases} (a, b], & \text{if } c \leqslant a \\ (c, b], & \text{if } a < c < b \\ \emptyset, & \text{if } c \geqslant b \end{cases}$
- $(-\infty, b] \cap (a, \infty) = \begin{cases} (a, b], & \text{if } a < b \\ \emptyset, & \text{if } b \leqslant a \end{cases}$
- and, trivially, for any $A \in \mathcal{S}$: $A \cap \emptyset = \emptyset \in \mathcal{S}$.

Both algebras and σ-algebras demand closure under complementation but this motivating example exhibits slightly weaker behaviour. Instead of $A^{\mathsf{c}} \in \mathcal{S}$, we instead have that, for example

$$(a,b]^{\mathsf{c}} = \underbrace{(-\infty, a]}_{\in \mathcal{S}} \sqcup \underbrace{(b, +\infty)}_{\in \mathcal{S}}$$

is a disjoint union of elements of $\mathcal{S}$. The same holds for all other elements of $\mathcal{S}$:

- $(-\infty, b]^{\mathsf{c}} = (b, \infty) \in \mathcal{S}$
- $\emptyset^{\mathsf{c}} = \mathbb{R} = \underbrace{(-\infty, a]}_{\in \mathcal{S}} \sqcup \underbrace{(a, +\infty)}_{\in \mathcal{S}}$ for any $a \in \mathbb{R}$.
- $(a, +\infty)^{\mathsf{c}} = (-\infty, a] \in \mathcal{S}$.

Thus, $\mathcal{S}$ is closed under complements being expressible as a disjoint union of elements of $\mathcal{S}$. This property looks like it'll be compatible with an additive set function μ_0 as we now have a way to express $\mu_0(A^{\mathsf{c}})$ as a sum $\sum_{j=1}^{n} \mu_0(E_j)$ of the values of its disjoint "pieces."

Definition 3.5.1 A collection $\mathcal{S} \subseteq 2^X$ is called a **semi-algebra** if the following hold:

- $\emptyset, X \in \mathcal{S}$
- $\mathcal{S}$ is closed under finite intersections i.e. $A, B \in \mathcal{S} \implies A \cap B \in \mathcal{S}$.
- If $A \in \mathcal{S}$, then $A^{\mathsf{c}} = \bigsqcup_{j=1}^{n} E_j$ where $\{E_j\}_1^n \subseteq \mathcal{S}$.

Our goal is to define a measure μ (on a suitable σ-algebra) that extends ν. We'll do this by first extending ν to a set function on the algebra $\mathrm{Alg}(\mathcal{S})$ generated by $\mathcal{S}$. Then we will define an outer measure μ^* (using ν) in the same spirit as the last section, and finally restrict to get μ. Since our end goal is a measure, we must necessarily have σ-additivity being carried through both extensions. Since neither domain of ν and μ_0 are σ-algebras, we must take care to specify the sense in which both maps are σ-additive.

Definition 3.5.2 Let $\mathcal{C} \subseteq 2^X$. A set function $\xi \colon \mathcal{C} \to [0, \infty]$ is a **measure on $\mathcal{C}$** if:

(a) ξ is additive within $\mathcal{C}$,

(b) ξ is σ-additive within $\mathcal{C}$.

Remarks 3.5.3 Since (finite additivity + σ-sub-additivity) is equivalent to σ-additivity, one can relax either:

- **(a)** to $\xi(\emptyset) = 0$, or
- **(b)** to σ-sub-additivity,

but not both at the same time!

Thus, I'll define:

Definition 3.5.4

- ν is called a **pre-pre-measure** on a semi-algebra $\mathcal{S}$ if it's a *measure on* $\mathcal{S}$.
- μ_0 is called a **pre-measure** on an algebra $\mathcal{A}$ if it's a *measure on* $\mathcal{A}$.

3.5.1 Extension From $\mathcal{S}$ to $\mathrm{Alg}(\mathcal{S})$

Important pedagogical point: When it comes to the construction of measures, we'll see that the algebra generated by a semi-algebra is very simple — one can explicitly represent all sets in such an algebra as a disjoint union of finitely many semi-algebra elements. This is **not** the case for a σ-algebra generated by a semi-algebra — we have no idea what the sets look like i.e. there is no explicit representation.

Without an explicit representation, proving that certain properties hold for all sets in $\sigma(\mathcal{S})$ must rely on techniques that don't leverage such an explicit representation. This will be one of the major difficulties of extending the set function from a semi-algebra to a σ-algebra generated by said semi-algebra.

Lemma 3.5.5 Let $\mathcal{S} \subseteq 2^X$ be a semi-algebra. Denote by $\mathrm{Alg}(\mathcal{S})$ the algebra generated by $\mathcal{S}$. Then

$$\mathrm{Alg}(\mathcal{S}) = \left\{ \bigsqcup_{i\in I} S_i \colon I\text{-finite}, \{S_i\}_{i\in I} \subseteq \mathcal{S} \text{ disjoint} \right\}.$$

Proof. Denote the collection on the right side by Δ. It's clear that $\mathcal{S} \subseteq \Delta$ because one can take any $S_i \in \mathcal{S}$, and let I be a singleton. If it can be shown further that Δ is an algebra, then $\mathrm{Alg}(\mathcal{S}) \subseteq \Delta$.

(i) $X \in \mathcal{S} \implies X \in \Delta$

(ii) Let $\bigsqcup_{i\in I} S_i$ and $\bigsqcup_{j\in J} T_j$ be two elements of Δ. Then

$$\Big(\bigsqcup_{i\in I} S_i\Big) \cap \Big(\bigsqcup_{j\in J} T_j\Big) = \bigsqcup_{(i,j)\in I\times J} S_i \cap T_j$$

which is an element of Δ since $\mathcal{S}$ is closed under finite intersections, and $\{S_i \cap T_j\}_{(i,j)\in I\times J} \subseteq \mathcal{S}$ is a finite, disjoint collection.

(iii) Is Δ closed under complementation? Let $\bigsqcup_{i\in I} S_i \in \Delta$. Since each $S_i \in \mathcal{S}$-semi-algebra, we may write the complement of S_i as a disjoint union of finitely many sets $\{S_{ij}\}_{j\in J_i} \subseteq \mathcal{S}$ i.e.

$$S_i^{\mathsf{c}} = \bigsqcup_{j\in J_i} S_{ij} \in \Delta$$

By De Morgan's law,

$$\Big(\bigsqcup_{i\in I} S_i\Big)^{\mathsf{c}} = \bigcap_{i\in I} S_i^{\mathsf{c}} \overset{\text{(ii)}}{\in} \Delta.$$

Thus, Δ is an algebra containing $\mathcal{S}$ and so $\mathrm{Alg}(\mathcal{S}) \subseteq \Delta$.

For the reverse inclusion, note that any element of Δ is a disjoint union of elements of $\mathcal{S}$, and because $\mathcal{S} \subseteq \mathrm{Alg}(\mathcal{S})$-algebra the aforementioned disjoint union is also an element of $\mathrm{Alg}(\mathcal{S})$. ■

This representation makes our first extension theorem easier to prove:

Theorem 3.5.6 Let $\mathcal{S}$ be a semi-algebra on X, and $\nu \colon \mathcal{S} \to [0, +\infty]$ a pre-pre-measure on $\mathcal{S}$. Then, there exists a pre-measure μ_0 uniquely extending ν to $\mathrm{Alg}(\mathcal{S})$ which is defined by:

$$\mu_0\Big(\bigsqcup_{i\in I} S_i\Big) = \sum_{i\in I} \nu(S_i).$$

Proof. First off, is μ_0 well-defined?

Suppose that $A \in \text{Alg}(\mathcal{S})$ has two distinct representations

$$A = \bigsqcup_{i \in I} S_i = \bigsqcup_{j \in J} T_j.$$

We wish to show that

$$\mu_0(A) = \sum_{i \in I} \nu(S_i) = \sum_{j \in J} \nu(T_j).$$

Since $S_i \subseteq A$ and $S_i \in \mathcal{S}$, we can see that

$$\mathcal{S} \ni S_i = S_i \cap A = S_i \cap \bigsqcup_{j \in J} T_j = \bigsqcup_{j \in J} \underbrace{S_i \cap T_j}_{\in \mathcal{S}}.$$

By the additivity of ν on $\mathcal{S}$, we have that

$$\nu(S_i) = \sum_{j \in J} \nu(S_i \cap T_j)$$

which implies

$$\begin{aligned} \mu_0(A) &= \sum_{i \in I} \nu(S_i) \\ &= \sum_{i \in I} \sum_{j \in J} \nu(S_i \cap T_j). \end{aligned}$$

This argument was symmetric in S_i and T_j so we simply repeat the same argument to get that:

$$\begin{aligned} \mu_0(A) &= \sum_{j \in J} \nu(T_j) \\ &= \sum_{j \in J} \sum_{i \in I} \nu(T_j \cap S_i). \end{aligned}$$

An analogous statement holds for when A has both a finite and countably infinite representation, the proof of which leverages σ-additivity in the obvious place. Thus, μ_0 is well-defined.

Is μ_0 a pre-measure on $\text{Alg}(\mathcal{S})$?

- Is $\mu_0(\emptyset) = 0$?

  Since ν is a (pre-pre-)measure, $\nu(\emptyset) = 0$ and so $\nu(\emptyset) = 0$.

- Is μ_0 countably additive?

  Let $\{A_i\}_{i \in \mathbb{N}} \subseteq \text{Alg}(\mathcal{S})$ be a pairwise disjoint collection whose union $A = \bigsqcup_{i \in \mathbb{N}} A_i \in \text{Alg}(\mathcal{S})$. We wish to show that

  $$\mu_0(A) = \sum_{i \in \mathbb{N}} \mu_0(A_i).$$

  Each $A_i \in \text{Alg}(\mathcal{S})$ has a disjoint representation

  $$A_i = \bigsqcup_{j \in J_i} S_{i,j}.$$

  Thus, we can write A as

  $$A = \bigsqcup_{i \in \mathbb{N}} \bigsqcup_{j \in J_i} S_{i,j}.$$

  Furthermore, we demanded that $A \in \text{Alg}(\mathcal{S})$ so it has its own disjoint union representation as

  $$A = \bigsqcup_{k \in K} T_k.$$

  Leveraging the first part of this proof, we make two observations:

$$(*)\ \ \mathcal{S} \ni T_k = T_k \cap A = T_k \cap \bigsqcup_{i\in\mathbb{N}} A_i = \bigsqcup_{i\in\mathbb{N}} \big(T_k \cap \bigsqcup_{j\in J_i} S_{ij}\big) = \bigsqcup_{i\in\mathbb{N}} \bigsqcup_{j\in J_i} T_k \cap S_{ij}$$

$$(**)\ \ \mathcal{S} \ni S_{ij} = S_{ij} \cap A = \bigsqcup_{k\in K} S_{ij} \cap T_k$$

$$\begin{aligned}
\therefore \quad \mu_0(A) \coloneqq \sum_{k\in K} \nu(T_k) &\overset{(*)}{=} \sum_{k\in K} \nu\Big(\bigsqcup_{i\in\mathbb{N}} \bigsqcup_{j\in J_i} T_k \cap S_{ij}\Big) \\
&= \sum_{k\in K} \sum_{i\in\mathbb{N}} \sum_{j\in J_i} \nu(T_k \cap S_{ij}) \quad \text{by the } \sigma\text{-additivity of } \nu \\
&= \sum_{i\in\mathbb{N}} \sum_{j\in J_i} \sum_{k\in K} \nu(T_k \cap S_{ij}) \\
&= \sum_{i\in\mathbb{N}} \sum_{j\in J_i} \nu\Big(\bigsqcup_{k\in K} S_{ij} \cap T_k\Big) \quad \text{by the additivity of } \nu \\
&\overset{(**)}{=} \sum_{i\in\mathbb{N}} \sum_{j\in J_i} \nu(S_{ij}) \\
&=: \sum_{i\in\mathbb{N}} \mu_0(A_i)
\end{aligned}$$

Is μ_0 an extension of ν? Any $A \in \mathcal{S}$ can be written as a trivial[5] union of itself and so $\mu_0(A) = \nu(A)$ so μ_0 and ν coincide on $\mathcal{S}$.

Is μ_0 unique? Suppose that μ_0 and $\widetilde{\mu}_0$ are two such extensions. They agree on the semi-algebra $\mathcal{S}$. We wish to show they coincide on $\mathrm{Alg}(\mathcal{S})$. Let $A \in \mathrm{Alg}(\mathcal{S})$ so it has a representation $A = \bigsqcup_{i\in I} S_i$ where $\{S_i\}_{i\in I} \subseteq \mathcal{S}$ is a pairwise disjoint collection. Then

$$\begin{aligned}
\mu_0(A) &\overset{\sigma\text{-add}}{=} \sum_{i\in I} \mu_0(S_i) \\
&= \sum_{i\in I} \widetilde{\mu}_0(S_i) \quad \text{since they agree on } \mathcal{S} \\
&\overset{\sigma\text{-add}}{=} \widetilde{\mu}_0(A).
\end{aligned}$$

■

Under certain conditions, we can go even further beyond and uniquely extend this pre-measure μ_0 on $\mathrm{Alg}(\mathcal{S})$ to a measure $\mu\colon \sigma(\mathcal{S}) \to [0, +\infty]$.

3.5.2 Extension From $\mathrm{Alg}(\mathcal{S})$ to $\sigma(\mathcal{S})$

We've done most of the hard work already.

Using **Proposition 3.3.3**, we can replace ρ with our pre-measure μ_0 and $\mathcal{K}$ with $\mathrm{Alg}(\mathcal{S})$ (which is also a cover). Then $\mu^*\colon 2^X \to [0, +\infty]$ defined for all $A \subseteq X$ by:

$$\mu^*(A) \coloneqq \inf\left\{\sum_{i=1}^{\infty} \mu_0(E_i)\colon \{E_i\}_{i\in\mathbb{N}} \subseteq \mathrm{Alg}(\mathcal{S}),\ A \subseteq \bigcup_{i\in\mathbb{N}} E_i\right\}$$

is an outer measure.

By **Theorem 3.4.3**, we've also demonstrated that the collection Σ of Carathéodory measurable sets (with respect to μ^*) is a σ-algebra, and that $\mu^*\big|_\Sigma$ is a complete measure. With the extra structure of the pre-measure μ_0 on the algebra $\mathrm{Alg}(\mathcal{S})$, we can show that $\mathcal{A} \subseteq \Sigma$:

[5] Is this the right word?

Let $A \in \mathcal{A}$ and $E \subseteq X$. We wish to show that $A \in \Sigma$. The inequality holds trivially if $\mu^*(E) = \infty$. WLOG, assume $\mu^*(E) < \infty$. Thus, given $\varepsilon > 0$, we can find a covering $\{E_i\}_{i\in\mathbb{N}} \subseteq \mathcal{A}$ of E that almost achieves the infimum i.e. s.t.

$$\mu^*(E) \leqslant \sum_{i\in\mathbb{N}} \mu_0(E_i) \leqslant \mu^*(E) + \varepsilon.$$

In particular, $E_i \cap A \in \mathcal{A}$ and $(E \cap A) \subseteq \bigcup_{i\in\mathbb{N}} E_i \cap A$ i.e. $\{E_i \cap A\}_{i\in\mathbb{N}}$ is a covering of $E \cap A$. Therefore,

$$\mu^*(E \cap A) \leqslant \sum_{i\in\mathbb{N}} \mu_0(E_i \cap A).$$

By the same logic,

$$\mu^*(E \cap A^{\mathsf{c}}) \leqslant \sum_{i\in\mathbb{N}} \mu_0(E_i \cap A^{\mathsf{c}}).$$

Then

$$\begin{aligned}
\mu^*(E \cap A) + \mu^*(E \cap A^{\mathsf{c}}) &\leqslant \sum_{i\in\mathbb{N}} \mu_0(E_i \cap A) + \sum_{i\in\mathbb{N}} \mu_0(E_i \cap A^{\mathsf{c}}) \\
&= \sum_{i\in\mathbb{N}} \mu_0((E_i \cap A) \sqcup (E_i \cap A^{\mathsf{c}})) \quad \text{by additivity} \\
&= \sum_{i\in\mathbb{N}} \mu_0(E_i) \\
&\leqslant \mu^*(E) + \varepsilon \quad \text{by construction.}
\end{aligned}$$

Let $\varepsilon \downarrow 0$ to conclude that $A \in \Sigma$.

Since $\mathcal{A} \subseteq \Sigma$ and Σ is a σ-algebra, it follows that $\sigma(\mathcal{A}) \subseteq \Sigma$.

What remains is to show that $\mu^*\big|_{\Sigma}$ is a unique extension of μ_0.

Is it an extension?

Let $A \in \mathrm{Alg}(\mathcal{S})$. If we let $E_1 = A \in \mathrm{Alg}(\mathcal{S})$ and $E_i = \emptyset \in \mathrm{Alg}(\mathcal{S})$ for $i > 1$, then $\{E_i\}_{i\in\mathbb{N}}$ is a cover of A. Therefore,

$$\mu^*(A) \leqslant \sum_{i\in\mathbb{N}} \mu_0(E_i) = \mu_0(A).$$

For the reverse inequality, let $\{E_i\}_{i\in\mathbb{N}} \subseteq \mathrm{Alg}(\mathcal{S})$ be a covering of A. Then we can "disjointify" this cover by defining $F_1 = E_1$ and for $i > 1$:

$$F_i = E_i \setminus \big(\bigcup_{j=1}^{i-1} E_j\big).$$

Note that the union of the E_i and F_i is the same and they both cover A. Thus, $\{F_i \cap A\}_{i\in\mathbb{N}} \subseteq \mathrm{Alg}(\mathcal{S})$ is a pairwise disjoint collection whose union is A.

$$\therefore\ \mu_0(A) \overset{\sigma\text{-add}}{=} \sum_{i\in\mathbb{N}} \mu_0(F_i \cap A) \leqslant \sum_{i\in\mathbb{N}} \mu_0(F_i) \leqslant \sum_{i\in\mathbb{N}} \mu_0(E_i).$$

The final two inequalities follow from monotonicity of μ_0. Taking the infimum over all covers $\{E_i\}_{i\in\mathbb{N}}$ yields the reverse inequality $\mu_0(A) \leqslant \mu^*(A)$.

$$\therefore \mu^*\big|_{\mathrm{Alg}(\mathcal{S})} = \mu_0.$$

3.5.3 Uniqueness Of Our Extension

We'll prove uniqueness on $\sigma(\text{Alg}(\mathcal{S}))$. This will require the further assumption of σ-finiteness. The reason for this is to break up each $A \in \sigma(\text{Alg}(\mathcal{S}))$ into "finite pieces" so a statement about the equality can be made by passing through a limit.

Proposition 3.5.7 Suppose that $\mu_1, \mu_2 \colon \sigma(\text{Alg}(\mathcal{S})) \to [0, +\infty]$ are two measures extending a measure $\mu_0 \colon \text{Alg}(\mathcal{S}) \to [0, +\infty]$ s.t.

$$\mu_1\big|_{\text{Alg}(\mathcal{S})} = \mu_2\big|_{\text{Alg}(\mathcal{S})},$$

and that both measures are σ-finite on X. Then $\mu_1 = \mu_2$.

Proof. Since μ_1 is σ-finite on X, and because $\text{Alg}(\mathcal{S})$ is a cover of X, there exists a sequence $\{E_i\}_{i\in\mathbb{N}} \subseteq \text{Alg}(\mathcal{S})$ s.t. $X = \bigcup_{i\in\mathbb{N}} E_i$ and $\mu_1(E_i) < \infty$ for every $i \in \mathbb{N}$. Since μ_1 and μ_2 coincide on $\text{Alg}(\mathcal{S})$, they coincide on every E_i in particular, and so μ_1 is σ-finite on X iff μ_2 is σ-finite on X. ◧

The rest of the proof relies on the concepts of a monotone class, and the characterisation of σ-additive functions by continuity from above/below of σ-additive functions.

Monotone Classes

Definition 3.5.8 A collection $\mathcal{M} \subseteq 2^X$ is called a **monotone class** if:

- $\mathcal{M}$ is closed under increasing limits i.e.

$$\text{If } \{A_j\}_{j\in\mathbb{N}} \subseteq \mathcal{M} \text{ is s.t. } A_j \subseteq A_{j+1} \text{ for every } j, \text{ then } A := \bigcup_{j\in\mathbb{N}} A_j \in \mathcal{M}.$$

- $\mathcal{M}$ is closed under decreasing limits i.e.

$$\text{If } \{B_j\}_{j\in\mathbb{N}} \subseteq \mathcal{M} \text{ is s.t. } B_j \supseteq B_{j+1} \text{ for every } j, \text{ then } B := \bigcup_{j\in\mathbb{N}} B_j \in \mathcal{M}.$$

The intersection of any collection of monotone classes is also a monotone class. Thus, we introduce the notion of the monotone class generated by a collection $\mathcal{C} \subseteq 2^X$, denoted by $\mathcal{M}(\mathcal{C})$, as the intersection of all monotone classes containing $\mathcal{C}$.

Theorem 3.5.9 (Monotone Class Theorem) Let $\mathcal{A} \subseteq 2^X$ be an algebra. Then $\mathcal{M}(\mathcal{A}) = \sigma(\mathcal{A})$.

Continuity From Above/Below

Let $\mathcal{C} \subseteq 2^X$.

Definition 3.5.10 A set function $\xi \colon \mathcal{C} \to [0, +\infty]$ is called

- **continuous from below at** $E \in \mathcal{C}$ if for any increasing sequence $\mathcal{C} \supseteq \{E_i\}_{i\in\mathbb{N}} \uparrow E := \bigcup_{i\in\mathbb{N}} E_i$, then we have that
$$\xi(E_n) \overset{n\to\infty}{\longrightarrow} \xi(E).$$

- **continuous from above at** $E \in \mathcal{C}$ if for any decreasing sequence $\mathcal{C} \supseteq \{E_i\}_{i\in\mathbb{N}} \downarrow E := \bigcap_{i\in\mathbb{N}} E_i$ s.t.
$$\exists N_0 \in \mathbb{N} \text{ s.t. } \xi(E_{N_0}) < \infty,$$
then we have that
$$\xi(E_n) \overset{n\to\infty}{\longrightarrow} \xi(E).$$

- **continuous at** E if ξ is both continuous from below and above at E.

Remarks 3.5.11

- Whenever there is no explicit reference to a set, any mention of **continuity from below/above** is a blanket reference applying to all sets in the domain of ξ.
- The stipulation of the finiteness condition after some point in the definition of continuity from below at E is to avoid a scenario like the following:

  Example Let $X = \mathbb{R}$ and consider λ (the measure extending the notion of length on $\mathbb{R}$). Let $E_n = [n, +\infty)$. It's clear that $\lambda(E_n) = \infty$ for every $n \in \mathbb{N}$. We have that

  $$\bigcap_{n\in\mathbb{N}} E_n = \emptyset$$

  and we note that $\lambda(\emptyset) = 0$. However,

  $$\lambda(E_n) \overset{n\to\infty}{\not\longrightarrow} \lambda(\emptyset).$$

Lemma 3.5.12 Let $\mathcal{A} \subseteq 2^X$ be an algebra, and $\xi\colon \mathcal{A} \to [0, \infty]$ be additive on $\mathcal{A}$. Then,

1. If ξ is σ-additive, then ξ is continuous at E for all $E \in \mathcal{A}$.
2. If ξ is continuous from below, then ξ is σ-additive.
3. If ξ is continuous from above at $\emptyset$ and ξ is finite on X, then ξ is σ-additive.

Proof.

1. Suppose that ξ is σ-additive.

   - We wish to show that ξ is continuous from below at any $E \in \mathcal{A}$.
     Consider any increasing sequence $\mathcal{A} \supseteq \{E_n\}_{n\in\mathbb{N}} \uparrow E$. By **disjointification**, we construct a pairwise disjoint sequence $\{F_n\}_{n\in\mathbb{N}}$ with the same union as the E_n. This allows us to leverage the σ-additivity of ξ. Since the E_n are increasing, our F_n take the simplified form $F_1 = E_1$ and $F_n = E_n \setminus E_{n-1}$. Note that

     $$\begin{aligned}\bigsqcup_{j=1}^{n} F_j &= F_1 \sqcup \bigsqcup_{j\geqslant 2} F_j \\ &= F_1 \sqcup \bigsqcup_{j\geqslant 2} (E_j \setminus E_{j-1}) \\ &= E_1 \sqcup (E_2 \setminus E_1) \sqcup \ldots \sqcup (E_n \setminus E_{n-1}) \\ &= E_n\end{aligned}$$

     Then, we conclude that

     $$\begin{aligned}\xi(E) &= \xi\Big(\bigcup_{n\in\mathbb{N}} E_n\Big) \\ &= \xi\Big(\bigsqcup_{n\in\mathbb{N}} F_n\Big) \\ &\overset{\sigma\text{-add}}{=} \sum_{n\in\mathbb{N}} \xi(F_n) \\ &= \lim_{n\to\infty} \sum_{j=1}^{n} \xi(F_j) \\ &= \lim_{n\to\infty} \xi\Big(\bigsqcup_{j=1}^{n} F_j\Big) \quad \text{by additivity} \\ &= \lim_{n\to\infty} \xi(E_n)\end{aligned}$$

- We wish to show that ξ is continuous from below at any $E \in \mathcal{A}$. Consider any decreasing sequence $\mathcal{A} \supseteq \{E_n\}_{n\in\mathbb{N}} \downarrow E$, that after some point N_0 is finite i.e.

$$\exists N_0 \in \mathbb{N} \text{ s.t. } n \geqslant N_0 \implies \xi(E_n) < \infty.$$

The idea is to transform the decreasing sequence into an increasing one.

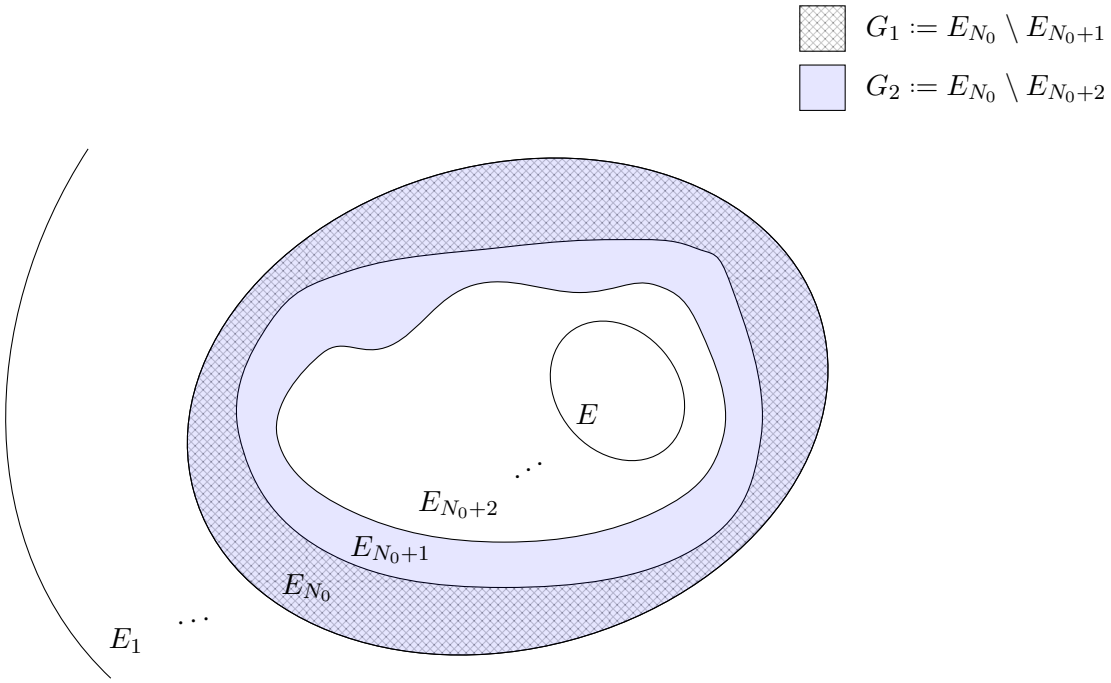


Figure 3.2: A visualisation of the construction of our increasing sequence.

Note that $G_1, G_2 \in \mathcal{A}$. Proceeding inductively, define

$$G_k := E_{N_0} \setminus E_{N_0+k} \in \mathcal{A}.$$

This sequence $\mathcal{A} \supseteq \{G_k\}_{k\in\mathbb{N}} \uparrow (E_{N_0} \setminus E)$. Thus, by the first part

$$\xi(G_k) \uparrow \xi(E_{N_0} \setminus E).$$

Writing it out more explicitly,

$$\lim_{k\to\infty} (\xi(E_{N_0}) - \xi(E_{N_0+k})) = \lim_{k\to\infty} \xi(E_{N_0} \setminus E_{N_0+k}) = \xi(E_{N_0} \setminus E) = \xi(E_{N_0}) - \xi(E),$$

the first and final equality of which are possible thanks to $\mu(E_{N_0}) < \infty$ and the monotonicity of ξ. Therefore, ξ is continuous from above.

2. Let ξ be continuous from below. Take any set $E = \bigsqcup_{k\in\mathbb{N}} \in \mathcal{A}$ where $\mathcal{A} \supseteq \{E_k\}_{k\in\mathbb{N}} \uparrow E$ is a pairwise disjoint increasing sequence. By monotonicity, for any n:

$$\xi\Big(\bigsqcup_{k=1}^{n} E_k\Big) \leqslant \xi(E).$$

The left-hand side is equal to $\sum_{k=1}^{n} \xi(E_k)$ by additivity and taking the limit as $n \to \infty$ gives

$$\sum_{k\in\mathbb{N}} \xi(E_k) \leqslant \xi(E).$$

This holds generally. We haven't used continuity from below yet — we'll do so for the reverse inequality:

Consider the sequence $\{F_n\}_{n\in\mathbb{N}}$ defined by

$$F_n = \bigsqcup_{k=1}^{n} E_k.$$

$F_n \in \mathcal{A}$ for every $n \in \mathbb{N}$ and F_n increases to E. Since ξ is continuous from below, $\xi(F_n) \uparrow \xi(E)$. By the additivity of ξ, the reverse inequality follows. Thus, ξ is σ-additive.

3. Suppose that ξ is continuous from above at $\emptyset$ and $\xi(X) < \infty$. Since ξ is finite, it is therefore σ-finite on X. Consider a sequence of pairwise disjoint sets $\mathcal{A} \supseteq \{E_k\}_{k\in\mathbb{N}} \uparrow E \coloneqq \bigsqcup_{k\in\mathbb{N}} E_k$. We want to construct a sequence that decreases to the empty set. Define the sequence $\{F_n\}_{n\in\mathbb{N}}$ by

$$F_n = \bigsqcup_{k\geqslant n} E_k.$$

Despite F_n being a countable union (and algebras being closed only under finite unions), we can write it as a finite difference of elements of $\mathcal{A}$:

$$F_n = E \setminus \Big(\bigsqcup_{j=1}^{n-1} E_j \Big).$$

Thus, $\{F_n\}_{n\in\mathbb{N}} \subseteq \mathcal{A}$. Then $F_n \downarrow \emptyset$ and $\xi(F_1) < \infty$ because $F_1 \subseteq X$ and $\xi(X) < \infty$. Thus,

$$\xi(F_n) \to \xi(\emptyset) = 0.$$

Now we write:

$$\begin{aligned}
\xi(E) &= \xi\Big(\Big(\bigsqcup_{k=1}^{n} E_k \Big) \sqcup \Big(\bigsqcup_{k>n}^{n} E_k \Big) \Big) \\
&= \xi\Big(\bigsqcup_{k=1}^{n} E_k \Big) + \xi(F_{n+1}) \\
&= \sum_{k=1}^{n} \xi(E_k) + \xi(F_{n+1}) \\
&\overset{j\to\infty}{\longrightarrow} \sum_{k=1}^{\infty} \xi(E_k) + 0
\end{aligned}$$

■

Proof of Proposition 3 (Continued). Fix $n \in \mathbb{N}$. Define

$$\mathcal{B}_n = \{E \in \sigma(\mathrm{Alg}(\mathcal{S})) \colon \mu_1(E \cap E_n) = \mu_2(E \cap E_n)\}.$$

Note that both quantities $\mu_i(E \cap E_n)$ are finite. $\mathcal{B}_n \subseteq \sigma(\mathrm{Alg}(\mathcal{S}))$ by definition. We wish to prove that this is in fact an equality.

If $E \in \mathrm{Alg}(\mathcal{S})$, then $E \cap E_n \in \mathrm{Alg}(\mathcal{S})$ and so $\mu_1(E \cap E_n) = \mu_2(E \cap E_n)$. Therefore, the inclusion $\mathrm{Alg}(\mathcal{S}) \subseteq \mathcal{B}_n$ holds. It's very difficult to show that $\mathcal{B}_n$ is a σ-algebra but it's easier to show that it's a monotone class.

- Consider an increasing sequence $\mathcal{B}_n \supseteq \{A_j\}_{j\in\mathbb{N}} \uparrow A \coloneqq \bigcup_{j\in\mathbb{N}} A_j$. We must show that $A \in \mathcal{B}_n$. Since $A_j \in \mathcal{B}_n$, we know that

$$\mu_1(A_j \cap E_n) = \mu_2(A_j \cap E_n).$$

Note that $A_j \cap E_n$ is increasing towards $A \cap E_n$. Since, μ_1 and μ_2 are σ-additive, they are continuous from below. Thus,

$$\mu_1(A \cap E_n) \overset{\infty\leftarrow j}{\longleftarrow} \mu_1(A_j \cap E_n) = \mu_2(A_j \cap E_n) \overset{j\to\infty}{\longrightarrow} \mu_2(A \cap E_n)$$

and because the j^{th} terms of both sequences are equal, they have the same limit. Thus, $\mu_1(A \cap E_n) = \mu_2(A \cap E_n)$ and so $A \in \mathcal{B}_n$.

- Consider a decreasing sequence $\mathcal{B}_n \supseteq \{B_j\}_{j\in\mathbb{N}} \downarrow B \coloneqq \bigcap_{j\in\mathbb{N}} B_j$. Since $B_j \in \mathcal{B}_n$,

$$\mu_1(B_j \cap E_n) = \mu_2(B_j \cap E_n).$$

Since μ_1 and μ_2 are σ-additive, they are both therefore continuous from above. In particular, μ_1 is continuous from above at $B \cap E$ i.e. for any decreasing sequence $(B_j \cap E_n) \downarrow (B \cap E_n)$, if we have that $\exists N_0 \in \mathbb{N}$ s.t. $\mu_1(B_{N_0} \cap E_n) < \infty$, then

$$\mu_1(B_j \cap E_n) \overset{j\to\infty}{\longrightarrow} \mu_1(B \cap E_n).$$

The σ-finiteness of μ_1 guarantees our finiteness condition above because for every $j \in \mathbb{N}$:

$$\begin{aligned}(B_j \cap E_n) \subseteq E_n \overset{\mu_1\text{-monotone}}{\Longrightarrow} \mu_1(B_j \cap E_n) &\leqslant \mu_1(E_n) \\ &< \infty.\end{aligned}$$

Finally,

$$\mu_1(B \cap E_n) \overset{\infty\leftarrow j}{\longleftarrow} \mu_1(B_j \cap E_n) = \mu_2(B_j \cap E_n) \overset{j\to\infty}{\longrightarrow} \mu_2(B \cap E_n)$$

and we conclude that $B \in \mathcal{B}_n$.

Since $\mathcal{B}_n$ is a monotone class that contains $\mathrm{Alg}(\mathcal{S})$, it also contains $\mathcal{M}(\mathrm{Alg}(\mathcal{S}))$. Combining all of the above, we've shown that

$$\mathrm{Alg}(\mathcal{S}) \subseteq \mathcal{M}(\mathrm{Alg}(\mathcal{S})) \subseteq \mathcal{B}_n \subseteq \sigma(\mathrm{Alg}(\mathcal{S})).$$

By the Monotone Class Theorem, $\mathcal{M}(\mathrm{Alg}(\mathcal{S}))$ and we conclude that $\mathcal{B}_n = \sigma(\mathrm{Alg}(\mathcal{S}))$.

Now to show that both measures agree on all of $\sigma(\mathrm{Alg}(\mathcal{S}))$. Let $A \in \sigma(\mathrm{Alg}(\mathcal{S}))$. Equivalently, $A \in \mathcal{B}_n$ i.e.

$$\mu_1(A \cap E_n) = \mu_2(A \cap E_n).$$

Since $E_n \uparrow X$, we can use that μ_1, μ_2 being σ-additive implies continuity from below to conclude that

$$\mu_1(A) = \mu_1(A \cap X) \overset{\infty\leftarrow n}{\longleftarrow} \mu_1(A \cap E_n) = \mu_2(A \cap E_n) \overset{n\to\infty}{\longrightarrow} \mu_2(A \cap X) = \mu_2(A).$$

Thus concludes the proof of uniqueness. ■

All our work is summarised in the following result:

Theorem 3.5.13 (**Carathéodory's Extension Theorem For Semi-Algebras**) Let ν be a pre-pre-measure on a semi-algebra $\mathcal{S}$. Then there exists a unique pre-measure $\mu_0 \colon \mathrm{Alg}(\mathcal{S}) \to [0, +\infty]$ on X extending ν. Define μ^* by

$$\mu^*(A) \coloneqq \inf\left\{\sum_{i=1}^{\infty} \mu_0(E_i) \colon \{E_i\}_{i\in\mathbb{N}} \subseteq \mathrm{Alg}(\mathcal{S}),\ A \subseteq \bigcup_{i\in\mathbb{N}} E_i\right\}$$

for every $A \subseteq X$, and let Σ denote the Carathéodory measurable sets with respect to μ^*. Then there is a unique, complete measure $\mu \colon \Sigma \to [0, +\infty]$ defined by $\mu = \mu^*\big|_{\Sigma}$ on a σ-algebra Σ containing $\mathcal{S}$. This measure extends μ_0 from $\mathrm{Alg}(\mathcal{S}) \subseteq \Sigma$.

3.6 Defining the Lebesgue Measure

Recall the semi-algebra

$$\mathcal{S} = \{(a, b] \colon a, b \in \mathbb{R}\} \cup \{(a, \infty) \colon a \in \mathbb{R}\} \cup \{(-\infty, b] \colon b \in \mathbb{R}\} \cup \{\emptyset\}.$$

If we can demonstrate that the set function $\nu \colon \mathcal{S} \to [0, +\infty]$ defined for any $I \in \mathcal{S}$ by $\nu(I) = \text{len}(I)$, is a pre-pre-measure, then we may apply Carathéodory's extension theorem for semi-algebras — this will extend our notion of interval length to a full measure on Σ.

Claim ν is a pre-pre-measure.

Proof. Let $\{I_i\}_{i\in\mathbb{N}} \subseteq \mathcal{S}$ be a pairwise disjoint collection whose union $I = \bigsqcup_{i\in\mathbb{N}} I_n$ is in $\mathcal{S}$. We wish to show that

$$\nu(I) = \sum_{i\in\mathbb{N}} \nu(I_i).$$

For any finite $N \in \mathbb{N}$, it's clear that

$$\nu(I) \geqslant \sum_{i=1}^{N} \nu(I_i).$$

Since this holds for every N, it follows that

$$\nu(I) \geqslant \sum_{i=1}^{\infty} \nu(I_i).$$

For the reverse inequality, there are two cases:

- If $I \in \mathcal{S}$ is a finite interval i.e. $I = (a, b]$ for $a, b \in \mathbb{R}$, if we order the constituent $I_i = (a_i, b_i]$ from left to right (i.e. $a_i < b_i \leqslant a_{i+1} < b_{i+1}$ for every $i \in \mathbb{N}$) then for each $\varepsilon > 0$ we can cover each $I_i = (a_i, b_i]$ by an open interval $\widetilde{I_i} = (a_i, b_i + \varepsilon 2^{-i})$. This collection $\{\widetilde{I_i}\}_{i\in\mathbb{N}}$
  - satisfies $\nu(\widetilde{I_i}) = \nu(I_i) + \varepsilon 2^{-i}$
  - and is an open cover of $[a + \varepsilon, b]$.

  By compactness, we may extract a finite sub-cover and extract a finite sub-cover $\{\widetilde{I_{i_j}}\}_{j=1}^{K}$ of $[a + \varepsilon, b]$. Since it's a cover, we certainly have that

  $$b - a - \varepsilon = \text{len}([a + \varepsilon, b]) \leqslant \sum_{j=1}^{K} \nu(\widetilde{I_{i_j}}) = \sum_{j=1}^{K} \left(\nu(I_{i_j}) + \varepsilon 2^{-i_j}\right) \leqslant \varepsilon + \sum_{j=1}^{K} \nu(I_{i_j})$$

  Rearranging gives

  $$\sum_{j=1}^{K} \nu(I_{i_j}) \geqslant \underbrace{b - a}_{=\nu(I)} - 2\varepsilon$$

  from which it follows that

  $$\sum_{i=1}^{\infty} \nu(I_i) \geqslant \sum_{j=1}^{K} \nu(I_{i_j}) \geqslant \nu(I) - 2\varepsilon.$$

  Let $\varepsilon \downarrow 0$ and the result follows.

- If I is an interval of infinite length, then for each $n \in \mathbb{N}$, $\{I_i \cap (-n, n]\}_{i\in\mathbb{N}}$ is a pairwise disjoint collection with union $I \cap (-n, n]$. From the argument for finite I,

  $$\nu(I \cap (-n, n]) = \sum_{i=1}^{\infty} \nu(I_i \cap (-n, n]) \leqslant \sum_{i=1}^{\infty} \nu(I_i).$$

  The above inequality is true for every n so let $n \to \infty$, and note that $\nu(I \cap (-n, n]) = \infty$, concluding that

  $$\sum_{i=1}^{\infty} \nu(I_i) = \infty = \nu(I).$$

■

Thus, ν is a pre-pre-measure on $\mathcal{S}$, and by the Carathéodory extension theorem for semi-algebras there exists a unique and complete measure λ on Σ (which we shall call $\mathcal{L}$, the collection of Lebesgue-measurable sets) s.t. for every interval I: $\lambda(I) = \text{len}(I)$. This is the **Lebesgue measure on $\mathbb{R}$**.

3.7 Product Measures

So far we've constructed measures, and covered the important example of the Lebesgue measure on $\mathbb{R}$. Problems in probability aren't limited to a single trial of an experiment — we often repeat experiments to obtain a sample of values on which we perform numerical computations. Sometimes we're even interested in considering what happens in two different experiments at the same time.

The natural object to support this is, of course, the Cartesian product of two outcome spaces. However, now we must construct an appropriate σ-algebra on this product to describe such events, and on top of that an appropriate measure. The construction we follow naturally builds independence into a model of such experiments.

Since there's no point in re-writing a perfectly good resource, I link here **a set of lecture notes** by John K. Hunter, Professor Emeritus at UC Davis. Chapter 5 discusses product measures. I enclose here a very brief summary of the main points.

> Let $(X_1, \mathcal{F}_1, \mu_1)$ and $(X_2, \mathcal{F}_2, \mu_2)$ be measure spaces. We want to construct a product measure $\mu_1 \times \mu_2$ on an appropriate σ-algebra that satisfies the product rule
>
> $$(\mu_1 \times \mu_2)(A \times B) = \mu_1(A)\mu_2(B)$$
>
> for all $A \in \mathcal{F}_1$ and for all $B \in \mathcal{F}_2$.
>
> - Define a **rectangle** to be a subset of the form $A_1 \times A_2 \subseteq X_1 \times X_2$ where $A_i \subseteq X_i$ for $i = 1, 2$. The *sides* of the rectangle are A_1 and A_2.
> - Define a **measurable rectangle** to be a set of the form $R = A \times B$ where $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$.
> - Prove that the collection $\mathcal{R}$ of measurable rectangles is a semi-algebra.
>   - Denote by $\mathcal{F}_1 \otimes F_2$, the σ-algebra on σ-algebra on $X_1 \times X_2$ generated by $\mathcal{R}$. We call this the **product σ-algebra on $X_1 \times X_2$**.
> - Define $\nu \colon \mathbb{R} \to [0, +\infty]$ by $\nu(A \times B) = \mu_1(A)\mu_2(B)$. Prove that ν is a pre-pre-measure on $\mathbb{R}$.
> - By Carathéodory's extension theorem for semi-algebras, there exists a measure $\mu_1 \times \mu_2$ on the σ-algebra[6] generated by $\mathcal{R}$, denoted by $\mathcal{F}_1 \otimes \mathcal{F}_2$, such that $\forall A \in \mathcal{F}_1$, $\forall B \in \mathcal{F}_2$:
>
> $$(\mu_1 \times \mu_2)(A \times B) = \mu_1(A)\mu_2(B).$$
>
> - If μ_1 and μ_2 are σ-finite measures, then $\mu_1 \times \mu_2$ is **unique** and called **the product measure of μ_1 and μ_2**.

One of the most important examples of a product measure is the Lebesgue measure on $\mathbb{R}^n$.

Example The **Lebesgue measure on $\mathbb{R}^n$** is the completion of the n-fold product of the Lebesgue measure on $\mathbb{R}$ i.e.

$$\lambda_{\mathbb{R}^n} = \overline{\bigotimes_{1}^{n} \lambda_{\mathbb{R}}}.$$

[6] This σ-algebra is called the **product σ-algebra** and will be seen again in the chapter on random vectors. It is defined differently but is the same object.

CHAPTER 4

Measurable Functions

4.1 Random Variables

Now that we have measures on σ-algebras in our toolbox, it's natural to consider functions between measurable spaces that preserve/respect the σ-algebra structure. Such functions are called *measurable*. If the domain of such a function is a probability space, then we call the function a *random variable*.

From a practical perspective, one carries out a random experiment and observes some quantity exhibited by its outcome — not the outcome itself.

From a probabilistic perspective, the outcome space Ω is often too granular to work with directly. To remedy this we can shift our focus to some (observable) numerical quantity, typically in $\mathbb{R}$, that depends on the outcomes of a random experiment. One can think of such numerical quantities as summarising agents e.g. there are several outcomes that correspond to the sum of two dice rolls being 9.

In either case, the correspondence between outcomes $\omega \in \Omega$ and such numbers $x \in E$ defines a function $X : \Omega \to E$ which models the outcomes of a random experiment.

Example Gamblers are more concerned with their losses over many trials (of a game) than with the details of the games that give rise to them.

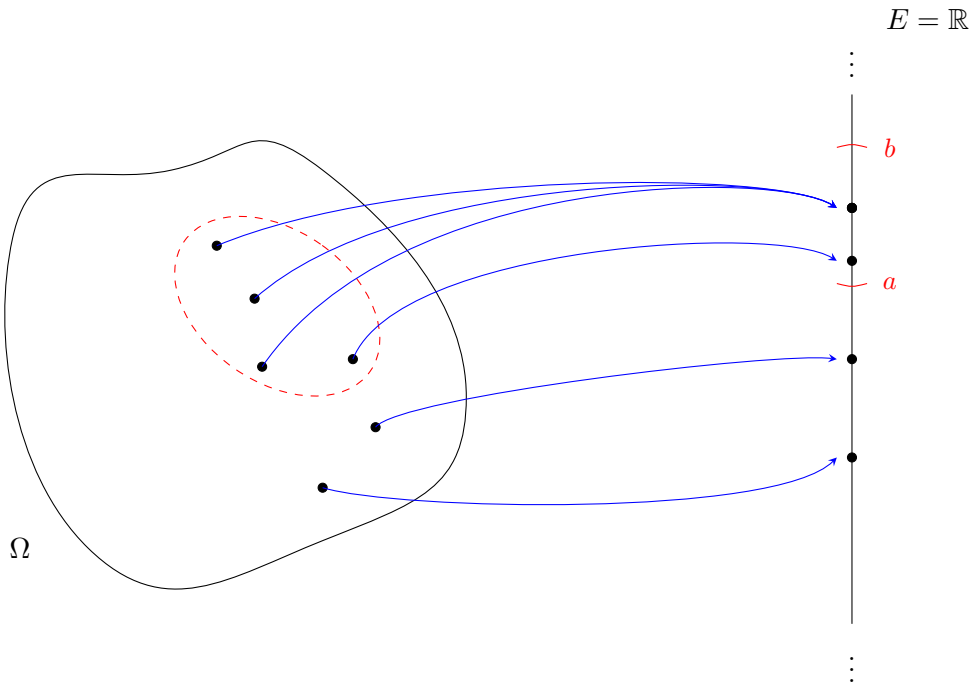


Figure 4.1: A visualisation of a finite space Ω, a mapping X into $E = \mathbb{R}$, and an interval summarising a collection of outcomes.

Take $E = \mathbb{R}$. Since the outcomes in Ω are governed by randomness, so too are these output numbers. The numbers serve to summarise the outcomes in Ω, so in a sense we'd expect the numbers in E to respect the event structure. We wish to find the probabilities that X assumes values in some regions of E as proxy for the direct probabilities which may be harder to work with. Thus, we're concerned with quantities like $\mathbb{P}(X \in B) := \mathbb{P}\big(X^{-1}(B)\big)$ where $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$.

Since $\mathbb{P}$ is defined on $\mathcal{F}$, a statement like $\mathbb{P}(X \in B)$ is only well-defined if $X^{-1}(B) \in \mathcal{F}$. This condition is what it means for X to respect the σ-algebra structure of both domain and codomain:

Definition 4.1.1

- Let $(X, \mathcal{F})$ and $(E, \mathcal{E})$ be measurable spaces. A function $f\colon X \to E$ is called **$(\mathcal{F}, \mathcal{E})$-measurable** if for every $B \in \mathcal{E}$,
$$f^{-1}(B) \in \mathcal{F}.$$
- The class of $(\mathcal{F}, \mathcal{E})$-measurable functions shall henceforth be denoted by $\mathrm{Meas}_{\mathcal{F},\mathcal{E}}(X\,;E)$.
- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A measurable, real-valued function on Ω i.e.
$$X \in \mathrm{Meas}_{\mathcal{F},\mathcal{B}_{\mathbb{R}}}(\Omega\,;\mathbb{R})$$
is called a **(real-valued) random variable**.

Remarks 4.1.2

- When it's understood that $f\colon (X, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ (resp. $(\mathbb{C}, \mathcal{B}_{\mathbb{C}})$), then the following are equivalent shorthands for $(\mathcal{F}, \mathcal{B}_{\mathbb{R}})$ (resp. $(\mathcal{F}, \mathcal{B}_{\mathbb{C}})$)-measurability:
  - f is **$\mathcal{F}$-measurable**,
  - f is **measurable**.
- The Borel σ-algebra $\mathcal{B}_X$ on X is the smallest/minimal σ-algebra $\mathcal{F}$ on X s.t. all continuous functions $f\colon X \to \mathbb{R}$ are $(\mathcal{F}, \mathcal{B}_{\mathbb{R}})$-measurable.
- Let $f\colon \mathbb{R} \to \mathbb{C}$. [8]
  - f is **Borel measurable** if it's $(\mathcal{B}_{\mathbb{R}}, \mathcal{B}_{\mathbb{C}})$-measurable.
  - f is **Lebesgue measurable** if it's $(\mathcal{L}, \mathcal{B}_{\mathbb{C}})$-measurable.

  Likewise for $f\colon \mathbb{R} \to \mathbb{R}$.
- We don't (typically) consider Lebesgue or even $(\mathcal{L}, \mathcal{L})$-measurable functions, choosing instead to work with Borel measurable functions.

  Reasons for this include the asymmetry of $(\mathcal{L}, \mathcal{B}_{\mathbb{R}})$-measurability i.e. $f, g \in \mathrm{Meas}_{\mathcal{L},\mathcal{B}_{\mathbb{R}}}(\mathbb{R}\,;\mathbb{R}) \not\Longrightarrow f \circ g \in \mathrm{Meas}_{\mathcal{L},\mathcal{B}_{\mathbb{R}}}(\mathbb{R}\,;\mathbb{R})$, and when it comes to $(\mathcal{L}, \mathcal{L})$-measurable functions, continuous functions need not be $(\mathcal{L}, \mathcal{L})$-measurable — this is a big drawback compared to continuous functions being automatically Borel measurable.
  - A **fantastic write-up** by `Nate Eldredge` on Math Stack Exchange encapsulates the different types of measurability, their scopes and limitations. I won't make any attempt to paraphrase it, and readily encourage you to read the full thing.

In practice, one checks the measurability of $f\colon X \to E$ on a sub-collection $\mathcal{C} \subseteq \mathcal{E}$ that generates $\mathcal{E}$ i.e. $\sigma(\mathcal{C}) = \mathcal{E}$.

Lemma 4.1.3 Let $\mathcal{C} \subseteq \mathcal{E}$ be a generating set i.e. $\sigma(\mathcal{C}) = \mathcal{E}$. Then
$$f \in \mathrm{Meas}_{\mathcal{F},\mathcal{E}}(X\,;E) \iff \forall A \in \mathcal{C}\colon f^{-1}(A) \in \mathcal{F}.$$

Proof. The forward implication is clear. For the reverse implication, define
$$\mathcal{I} \coloneqq \big\{A \subseteq E : f^{-1}(A) \in \mathcal{F}\big\}.$$
Note that this is a σ-algebra that contains $\mathcal{C}$, and therefore $\sigma(\mathcal{I}) \supseteq \sigma(\mathcal{C})$. However, $\sigma(\mathcal{I}) = \mathcal{I}$ and $\sigma(\mathcal{C}) = \mathcal{E}$ so $\mathcal{I} \supseteq \mathcal{E}$, and therefore f is $(\mathcal{F}, \mathcal{E})$-measurable. ∎

Here are some examples of random variables:

Example Let $A \in \mathcal{F}$. The **indicator of A** is the function $\mathbb{1}_A \colon \Omega \to \{0,1\}$ that *indicates* whether an event A occurs or not. It's defined for any $\omega \in \Omega$ by

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

We consider $\{0,1\}$ to be equipped with $2^{\{0,1\}} = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}$. Let $B \in 2^{\{0,1\}}$. Then:

$$\mathbb{1}_A^{-1}(B) = \begin{cases} \Omega & \text{if } 0 \in B \ni 1 \\ A^{\mathsf{c}} & \text{if } 0 \in B \not\ni 1 \\ A & \text{if } 0 \notin B \ni 1 \\ \emptyset & \text{if } 0 \notin B \not\ni 1 \end{cases}$$

In every case, the pre-image is in $\mathcal{F}$. Thus, $\mathbb{1}_A \in \mathrm{Meas}_{\mathcal{F}, 2^{\{0,1\}}}(\Omega; \{0,1\})$.

Example Suppose that we roll two independent and identical fair die. The outcome space is the collection of pairs (i,j). The random variable $X \colon \Omega \to \mathbb{R}$ defined by $X((i,j)) = i + j$ returns the sum of the two rolls. This summarises the outcome space from 36 elements to the natural numbers from 2 to 12 (inclusive).

4.2 Properties of Measurable Functions

From the perspective of modelling nature, I don't think we typically define random variables that can attain the values $\pm\infty$. However, when it comes to measure-theoretical work, especially to do with convergence, we work with random variables $X \colon \Omega \to \overline{\mathbb{R}} := [-\infty, +\infty]$. The natural topology associated with $\overline{\mathbb{R}}$ is the order topology. We take $\{x \in \overline{\mathbb{R}} \colon x > a\} = (a, +\infty]$ with $a \in \mathbb{R}$ to be a neighbourhood of $+\infty$. The Borel σ-algebra of $(\overline{\mathbb{R}}, \mathcal{T}_{\mathrm{ord}})$ is generated by the intervals $\{(a, +\infty]\}_{a \in \mathbb{R}}$.

Definition 4.2.1 A function $f \colon X \to \overline{\mathbb{R}}$ is $\mathcal{F}$-measurable if for every $a \in \mathbb{R}$, $f^{-1}((a, +\infty]) \in \mathcal{F}$.

I'll also denote $f^{-1}((a, +\infty])$ by $\{f > a\}$.

Lemma 4.2.2 Let $f, g \colon X \to \overline{\mathbb{R}}$ be $\mathcal{F}$-measurable. The following sets are measurable:

- $\{f = +\infty\} = \bigcap_{n \in \mathbb{N}} \{f > n\}$ and an analogous statement can be made for $\{f = -\infty\}$.
- $\{f > g\} \equiv \{x \colon f(x) > g(x)\} = \bigcup_{q \in \mathbb{Q}} (\{f > q\} \cap \{g < q\})$
  and equality holds by the density of the rationals in the reals.
- $\{f \geqslant g\} = \{f < g\}^{\mathsf{c}}$
- $\{f = g\} = \{f \geqslant g\} \cap \{f \leqslant g\}$
- $\{f + g \text{ is undefined}\} = (\{f = +\infty\} \cap \{g = -\infty\}) \cup (\{f = -\infty\} \cap \{g = +\infty\})$
- $\{f + g \text{ is well-defined}\} = \{f + g \text{ is undefined}\}^{\mathsf{c}}$

■

Conventionally, we set $f(x) + g(x) = 0$ and $f(x)/g(x) = 0$ wherever the sums or ratios are not well-defined.

Proposition 4.2.3 Let $f, g \colon X \to \overline{\mathbb{R}}$ be $\mathcal{F}$-measurable, and $c \in \mathbb{R}$. Then $c \cdot f$, $f + g$, f/g, $|f|$, $\max(f, g)$, $\min(f, g)$, and fg are $\mathcal{F}$-measurable functions.

Proof.

- If $c = 0$, then
$$\{c \cdot f > a\} = \begin{cases} \emptyset & \text{if } a \geqslant 0, \\ X & \text{if } a < 0. \end{cases}$$
For $c \neq 0$,
$$\{c \cdot f > a\} = \begin{cases} \{f > a/c\} & \text{if } c > 0, \\ \{f < a/c\} & \text{if } c < 0. \end{cases}$$
- Let $A \coloneqq \{f + g \text{ is well-defined}\}$. Then
$$\{f + g > a\} = A \cap \{f + g > a\} = A \cap \{f > g - a\} = A \cap \bigcup_{q \in \mathbb{Q}} (\{f > q\} \cap \underbrace{\{a - g < q\}}_{= \{g > a - q\})}$$
- $1/g$ is defined on $A = X \setminus \{g = 0\}$. All sets henceforth are intersected with A.
  - If $a = 0$, then $\{1/g > a\} = \{1/g > 0\} = \{g > 0\} \setminus \{g = +\infty\}$.
  - If $a > 0$, then $\{1/g > a\} = \{g > 0\} \cap \{g < 1/a\}$.
  - If $a < 0$, then $\{1/g > a\} = (\{1/g > a\} \cap \{g > 0\}) \cup (\{1/g > a\} \cap \{g < 0\})$
$$= \{g > 0\} \cup \{g < 1/a\}.$$
- $\{|f| > a\} = \{f > a\} \cup \{f < -a\}$
- $\{\max(f, g) > a\} = \{f > a\} \cup \{g > a\}$
- $\{\min(f, g) > a\} = \{f > a\} \cap \{g > a\}$
- Prove that f^2 is $\mathcal{F}$-measurable, and it follows that $fg = \dfrac{(f+g)^2 - (f-g)^2}{4}$ is measurable?
  Indeed,
$$(f^2)^{-1}((a, \infty]) = \{x \in X \colon (f(x))^2 > a\} = \begin{cases} X & \text{if } a < 0, \\ \{f < -\sqrt{a}\} \cup \{f > \sqrt{a}\} & \text{if } a \geqslant 0 \end{cases}$$
  and both of these are in $\mathcal{F}$. The claim follows.

■

Proposition 4.2.4 Let $\{f_n\}_{n \in \mathbb{N}} \subseteq \mathrm{Meas}_{\mathcal{F}, \mathcal{B}_{\overline{\mathbb{R}}}}(X ; \overline{\mathbb{R}})$. Then both $\sup_{n \in \mathbb{N}} f_n$ and $\inf_{n \in \mathbb{N}} f_n$ are $\mathcal{F}$-measurable.

Proof.
$$\left(\sup_{n \in \mathbb{N}} f_n\right)^{-1}((a, +\infty]) = \bigcup_{n \in \mathbb{N}} f_n^{-1}((a, +\infty])$$

■

Corollary 4.2.5 Let $\{f_n\}_{n \in \mathbb{N}} \subseteq \mathrm{Meas}_{\mathcal{F}, \mathcal{B}_{\overline{\mathbb{R}}}}(X ; \overline{\mathbb{R}})$. Then:

- $\limsup_{n \to \infty} f_n$ and $\liminf_{n \to \infty} f_n$ are measurable.
- $A \coloneqq \{x \text{ s.t. } \lim_{n \to \infty} f_n(x) \text{ exists}\}$ is measurable, and defining $\lim_{n \to \infty} f_n(x) = 0$ where it doesn't exist, $\lim_{n \to \infty} f_n(x)$ is measurable.

Proof.

- Since both $\sup_n f_n$ and $\inf_n f_n$ are $\mathcal{F}$-measurable, it follows that

$$\limsup_{n\to\infty} f_n \coloneqq \inf_{n\geqslant 1}\left(\sup_{k\geqslant n} f_k\right)$$

  is measurable. The $\liminf_n f_n$ is analogously measurable.

- $A = \{\limsup_{n\to\infty} f_n = \liminf_{n\to\infty} f_n\}$ which is measurable. For every $x \in A$,

$$f(x) = \lim_{n\to\infty} f_n(x) = \limsup_{n\to\infty} f_n(x)$$

  so f is measurable. Otherwise, on $X \setminus A$ we have that f is constant and hence $f^{-1}(\{0\}) = X \setminus A \in \mathcal{F}$. Thus, f is measurable.

■

Corollary 4.2.6 The limit of a monotone sequence $\{f_n\}_{n\in\mathbb{N}} \subseteq \mathrm{Meas}_{\mathcal{F},\mathcal{B}_{\overline{\mathbb{R}}}}(X;\overline{\mathbb{R}})$ is $\mathcal{F}$-measurable.

Proof. By 2.4.3, every monotone sequence has a limit in $\overline{\mathbb{R}}$. For any $x \in X$, the sequence $\{f_n(x)\}_{n\in\mathbb{N}}$ is a monotone sequence and thus has a limit

$$\liminf_{n\to\infty} f_n(x) = \lim_{n\to\infty} f_n(x) = \limsup_{n\to\infty} f_n(x).$$

Thus, $\lim_{n\to\infty} f_n(x)$ is measurable. ■

Here is the main result of this section that will be used repeatedly[1] throughout these notes.

Theorem 4.2.7 Every *non-negative* measurable function $f\colon X \to \overline{\mathbb{R}}$ is the pointwise limit of a non-decreasing sequence of non-negative, simple, measurable functions $\{s_n\}_{n\in\mathbb{N}}$ i.e.

$$\forall x \in X\colon \quad \lim_{n\to\infty} s_n(x) = f(x).$$

Proof. For every $n \in \mathbb{N}$, if f is unbounded above then we may cut the height of f off at n, and then consider sub-dividing $[0, n]$ into intervals of equal height 2^{-n} and then approximating f by its floor over this partition of the range. Pictorially:

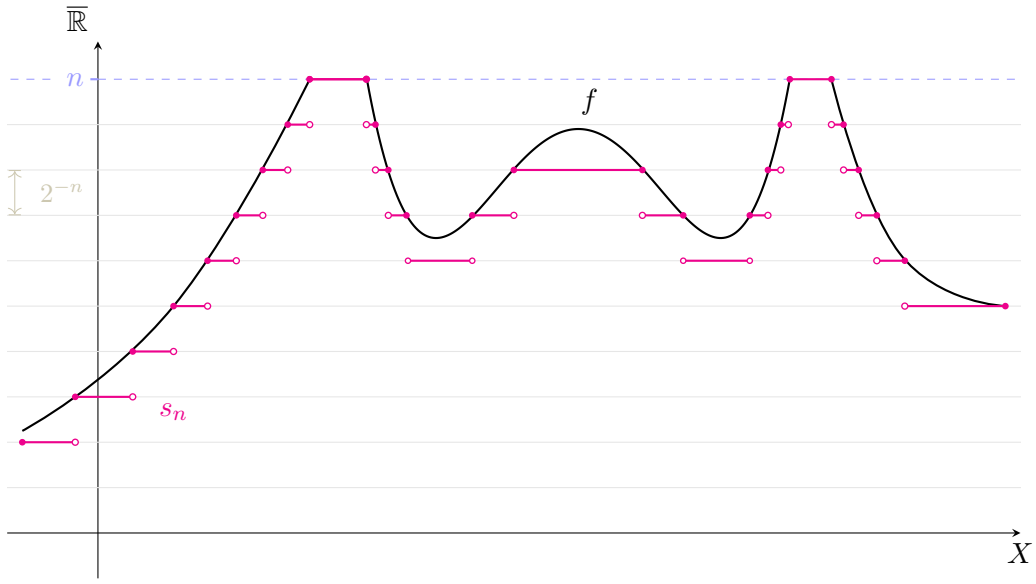


Figure 4.2: A visualisation of $s_n(x)$ approximating f for some $n \in \mathbb{N}$.

[1] I'm typing this in post.

i.e. for $k = 1, 2, \ldots, n2^n$:

$$\begin{aligned} s_n(x) &= \sum_{k=1}^{n2^n} \left(\frac{k-1}{2^n}\right) \mathbb{1}_{\left\{\frac{k-1}{2^n} \leqslant f < \frac{k}{2^n}\right\}}(x) + n\mathbb{1}_{\{f \geqslant n\}}(x) \\ &= \sum_{k=1}^{n2^n} \left(\frac{k-1}{2^n}\right) \mathbb{1}_{f^{-1}\Big(\underbrace{\left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)}_{\in \mathcal{B}_{[0,+\infty]}}\Big)}(x) + n\mathbb{1}_{f^{-1}\Big(\underbrace{[n, +\infty]}_{\in \mathcal{B}_{[0,+\infty]}}\Big)}(x) \end{aligned}$$

Since $f \in \mathrm{Meas}_{\mathcal{F}, \mathcal{B}_{[0,+\infty]}}(X\,; [0, +\infty])$, each s_n is a sum of finitely many indicator functions on sets in $\mathcal{F}$, and is therefore simple and measurable.

- Do the $\{s_n\}_{n\in\mathbb{N}}$ constitute a non-decreasing sequence?

  I'll just perform the inductive step. Say we begin with $s_n(x)$ and consider $s_{n+1}(x)$. At $n+1$, each interval in the range is further sub-divided in half e.g. for all x that satisfy $f(x) \in \left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)$, $s_n(x) = \frac{k-1}{2^n}$. Now note that:

$$\begin{aligned} f(x) \in \left[\tfrac{k-1}{2^n}, \tfrac{k}{2^n}\right) &= \left[\tfrac{k-1}{2^n}, \left(\tfrac{k-1}{2^n} + \tfrac{k}{2^n}\right)/2\right) \sqcup \left(\left(\tfrac{k-1}{2^n} + \tfrac{k}{2^n}\right)/2, \tfrac{k}{2^n}\right] \\ &= \left[\tfrac{2(k-1)}{2^{n+1}}, \tfrac{2k-1}{2^{n+1}}\right) \sqcup \left(\tfrac{2k-1}{2^{n+1}}, \tfrac{2k}{2^{n+1}}\right] \\ &= \left[\tfrac{2k-2}{2^{n+1}}, \tfrac{2k-1}{2^{n+1}}\right) \sqcup \left(\tfrac{2k-1}{2^{n+1}}, \tfrac{2k}{2^{n+1}}\right] \end{aligned}$$

  On the first sub-interval, we have that

$$s_{n+1}(x) = \frac{2k-2}{2^{n+1}} = \frac{k-1}{2^n} = s_n(x)$$

  and on the second sub-interval,

$$s_{n+1}(x) = \frac{2k-1}{2^{n+1}} > \frac{k}{2^n} = s_n(x),$$

  and for x s.t. $f(x) \in [n, +\infty]$, we may again sub-divide $[n, +\infty] = [n, n+1) \sqcup [n+1, +\infty]$. Over the first sub-interval, $s_{n+1}(x) \geqslant s_n(x)$ since we may run the same argument as above, and for the second we have that $s_{n+1}(x) = n + 1 \geqslant n = s_n(x)$.

- Is the pointwise limit equal to f?
  - If $f(x) < \infty$, then $\exists n \in \mathbb{N}$ s.t. $n > f(x)$ so that $|s_{n+1}(x) - s_n(x)| < 2^{-n}$.
  - If $f(x) = +\infty$, then $s_n(x) = n \overset{n\to\infty}{\longrightarrow} +\infty$.

Therefore, for every $x \in X$:

$$\lim_{n\to\infty} s_n(x) = f(x)$$

and by **Corollary 4.2.6**, f is measurable because it's the monotone limit of a sequence

$$\{s_n\}_{n\in\mathbb{N}} \subseteq \mathrm{Meas}_{\mathcal{F}, \mathcal{B}_{\overline{\mathbb{R}}}}(X\,; \overline{\mathbb{R}}).$$

■

Corollary 4.2.8 Any non-negative, **bounded** function $f\colon X \to \overline{\mathbb{R}}$ is measurable iff it is the **uniform** limit of a sequence of measurable functions.

Proof. Since f is bounded, the construction above works for any x to bound f by

$$s_n(x) := \frac{k-1}{2^n} \leqslant f(x) \leqslant \frac{k}{2^n} = s_n(x) + \frac{1}{2^n}$$

where k is suitably chosen. This implies that

$$0 \leqslant f(x) - s_n(x) \leqslant \frac{1}{2^n}$$
$$\text{i.e.} \quad 0 \leqslant |f(x) - s_n(x)| \leqslant \frac{1}{2^n}$$

The upper bound doesn't depend on x and so we have the uniform bound

$$\sup_{x \in X} |f(x) - s_n(x)| \leqslant \frac{1}{2^n}.$$

Thus, the convergence of $\{s_n\}_{n \in \mathbb{N}} \longrightarrow f$ as $n \to \infty$ is uniform. ■

The general case where f is not necessarily non-negative follows from writing it as the difference of two non-negative functions $f = f^+ - f^-$ where $f^+ = \max(f, 0)$ and $f^- = \max(0, -f)$. This proof follows because f is measurable iff f^+ and f^- are measurable:

$\Longrightarrow$ If f is measurable, then the constant function 0 is certainly measurable and it follows from **Proposition 4.2.3** that $\max(f, 0)$ and $\max(-f, 0)$ are measurable.

$\Longleftarrow$ f is the difference of two measurable functions and is therefore measurable by the same proposition.

4.3 Probability Distribution of X

Continuing on from **Section 4.1**, the discussion about $\mathbb{P}(X \in B)$ was good motivation for the concept of a map "pushing forward" a measure from one space onto another i.e. to describe the distribution of the outputs of a measurable function on a measurable space.

Definition 4.3.1 If Φ is a measurable map from one measurable space $(X, \mathcal{F})$ into a second $(E, \mathcal{E})$, then the **distribution of Φ under a measure μ** (or the **law of Φ with respect to μ**) on $(X, \mathcal{F})$ is the *unique* **pushforward measure** $\Phi_\sharp \mu$ defined on $(E, \mathcal{E})$ by

$$(\Phi_\# \mu)(B) = \mu(\Phi^{-1}(B)) \quad \text{for } B \in \mathcal{E}.$$

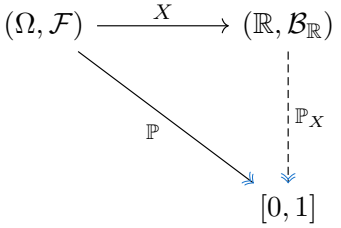
Proof. We must prove that $(\Phi_\# \mu)$ is indeed a measure.

- $(\Phi_\# \mu)(\emptyset) := \mu(\Phi^{-1}(\emptyset)) = \mu(\emptyset) = 0$
- Let $\{B_i\}_{i \in \mathbb{N}} \subseteq \mathcal{E}$ be a pairwise disjoint collection. Then,

$$\begin{aligned}
(\Phi_\# \mu)\Big(\bigsqcup_{i \in \mathbb{N}} B_i\Big) &:= \mu\Big(\Phi^{-1}\Big(\bigsqcup_{i \in \mathbb{N}} B_i\Big)\Big) \\
&= \mu\Big(\bigsqcup_{i \in \mathbb{N}} \Phi^{-1}(B_i)\Big) \\
&= \sum_{i \in \mathbb{N}} \mu(\Phi^{-1}(B_i)) \quad \text{since } \Phi \text{ is measurable, each } \Phi^{-1}(B_i) \in \mathcal{F} \\
&=: \sum_{i \in \mathbb{N}} (\Phi_\# \mu)(B_i)
\end{aligned}$$

■

Example A real-valued random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ induces a unique probability distribution $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ as the push-forward measure $X_\# \mathbb{P}$ of $\mathbb{P}$ via X according to the following diagram:



The conventions of my diagrams are:

- A single-headed arrow represents an "ordinary" map e.g. a random variable $X : \Omega \to \mathbb{R}$ between measurable spaces $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$
- A double-headed arrow represents a (probability) measure (e.g. $\mathbb{P} : \mathcal{F} \to [0, 1]$)
  - There's an extra meaning for the **double head** — the probability measures will be from a measurable space's **second** entry e.g. the σ-algebra $\mathcal{F}$ in $(\Omega, \mathcal{F})$.
- A dashed line represents an induced map
- A dotted line represents a map induced by an induced map (we'll see this shortly)

Thus, the **probability distribution** $\mathbb{P}_X : \mathcal{B}_{\mathbb{R}} \to [0, 1]$ *of the real-valued random variable* X is defined for $B \in \mathcal{B}_{\mathbb{R}}$ by:

$$\mathbb{P}_X(B) := X_{\#}\mathbb{P}(B) = \mathbb{P}\big(X^{-1}(B)\big) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

In order to specify the probability measure associated with a random variable, sometimes it's more convenient to specify alternative representations (CDFs, PMFs and PDFs) of the measure from which the corresponding experiment is clear.

One alternative specification that always exists is the distribution function of X. In the same way that one can reduce the problem of checking measurability of X from all sets in $\mathcal{B}_{\mathbb{R}}$ to a generating set $\{(-\infty, x]\}_{x \in \mathbb{R}}$, we can characterise $\mathbb{P}_X$ by its behaviour on $\{(-\infty, x]\}_{x \in \mathbb{R}}$:

All random variables admit a **cumulative distribution function** — a function $F_X : \mathbb{R} \to [0, 1]$ that specifies a probability measure by returning the probability that the random variable X assumes a value less than or equal to x:

$$\begin{aligned} F_X(x) &:= \mathbb{P}_X((-\infty, x]) \\ &:= \mathbb{P}(\{\omega \in \Omega : X(\omega) \leqslant x\}) \\ &= \mathbb{P}(X \leqslant x) \end{aligned}$$

Note that F_X is a non-decreasing function that satisfies:

- $F_X(x) \to 0$ as $x \to -\infty$
- $F_X(x) \to 1$ as $x \to +\infty$
- F_X is right-continuous[2] i.e. $F_X(x) = \lim_{y \to x+} F_X(y)$ for all $x \in \mathbb{R}$.

Conversely, for any distribution function F that satisfies the above properties, we can always construct a probability space and random variable such that F is the distribution function of X.

4.4 Support of Probability Distribution

Loosely speaking, the support of a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ can be thought of as where $\mathbb{P}$ lives in Ω. The traditional definition of the support of a function speaks of the closure of a subset

[2] Note: A cumulative distribution function is not necessarily continuous — it's very possible to have $F_X(x-) \neq F_X(x+)$. For example, in the next section we will see that the CDF of a discrete probability measure is a step function.

of the function's domain. However, we don't have a topology to speak of on Ω so we can't speak of support.

Practically speaking, the random variables of interest will map into a space like $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, $(\overline{\mathbb{R}}, \mathcal{B}_{\overline{\mathbb{R}}})$, $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ etc.[3] These are examples of Borel spaces.

4.4.1 Borel Spaces

Definition 4.4.1 First, some terminology:

- Let $(E, \mathcal{E})$ and $(S, \mathcal{S})$ be two measurable spaces. A map $\phi\colon E \to S$ is called **bimeasurable** if ϕ is measurable, invertible, and ϕ^{-1} is also measurable.
- Two spaces $(E, \mathcal{E})$ and $(S, \mathcal{S})$ are **measurably isomorphic** (or **Borel isomorphic**) if there exists a bimeasurable map between them. We denote this by
$$(E, \mathcal{E}) \cong_{\text{Meas}} (S, \mathcal{S}).$$
- A **Polish space** is a topological space $(Y, \mathcal{T}_Y)$ that is metrisable and separable.

Now for the definition of such spaces:

Definition 4.4.2 Let $(E, \mathcal{E})$ be a measurable space. If there exists a bimeasurable map $\varphi\colon E \to B$ **where B is a Borel subset of $\mathbb{R}$**, then $(E, \mathcal{E})$ is a **Borel space**.

> As a small spoiler for what's to come and why this is important, it's a fact that a regular conditional distribution of $\mathcal{F} = \sigma(X)$ given a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ exists if X maps into a Borel space $(E, \mathcal{E})$.

Definition 4.4.3 A measurable space $(E, \mathcal{E})$ is called a **standard Borel space** if it satisfies the following equivalent conditions:

- $(E, \mathcal{E})$ is measurably isomorphic to a Polish space $(Y, \mathcal{T}_Y)$ equipped with its Borel σ-algebra $\mathcal{B}_Y = \sigma(\mathcal{T}_Y)$ i.e.
$$(E, \mathcal{E}) \cong_{\text{Meas}} (Y, \mathcal{B}_Y).$$
- $(E, \mathcal{E})$ is measurably isomorphic to **some Borel subset B of some Polish space** $(Y, \mathcal{T}_Y)$, equipped with its subspace[4] Borel σ-algebra $\mathcal{B}_Y|_B = \{F \cap B\colon F \in \mathcal{B}_Y\} = \{F \in \mathcal{B}_Y\colon F \subseteq B\}$. i.e.
$$(E, \mathcal{E}) \cong_{\text{Meas}} (B, \mathcal{B}_Y|_B).$$

Remarks

- **These parts** of the above two definitions highlight their equivalence[5] because $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a Polish space.
- Henceforth, I will simply refer to such spaces as Borel spaces.
- The topology on E is given by the pullback of $\mathcal{T}_Y$ under the bimeasurable map ϕ i.e.
$$\mathcal{T}_E = \left\{\phi^{-1}(U) : U \in \mathcal{T}_Y\right\}.$$

Theorem 4.4.4 Let $\{(E_i, \mathcal{E}_i)\}_{i\in\mathbb{N}}$ be a collection of Borel spaces. Then
$$\left(E = \prod_{i\in\mathbb{N}} E_i,\ \mathcal{E} = \bigotimes_{n\in\mathbb{N}} \mathcal{E}_i\right)$$
is a Borel space.

[3] Or $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ in the case of a random *vector* which shall be seen later.

[4] The second equality follows because $B \in \mathcal{B}_Y$ implies that $F \cap B \in \mathcal{B}_Y$.

[5] From the bottom of my heart, this has been an unholy pain in the ass. Everyone uses different names, nobody bothers to explain the equivalence of definitions. Everyone uses a different canonical uncountable Borel space. I'm doing the opposite of authors and I state them as definitions, the theorems that link them, and refuse to prove any of them.

4.4.2 Measurable Classification of Borel Spaces

The following theorem in [8, p. 83], crediting Kuratowski, classifies Borel spaces in the sense of measurable isomorphism:

Theorem 4.4.5 Suppose that $(E, \mathcal{E})$ is Borel isomorphic to a Borel subset B of a complete, separable metric space Y (equipped with the σ-algebra $\{F \in \mathcal{B}_Y : F \subseteq B\}$). Then either:

- E is at most countably infinite and $\mathcal{E} = 2^E$, or
- $(E, \mathcal{E})$ is Borel isomorphic to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

So the isomorphism theorem above says that:

- All finite and countable Borel spaces are trivial in the sense that all subsets are measurable, and
- up to measurable isomorphism, there is "the" uncountable Borel space.

4.4.3 Topological Support of Probability Measure

Now we're in a position to define the support of a (probability) measure:

Definition 4.4.6 The **topological support** of a measure μ on a Borel space $(E, \mathcal{E})$ is the set of all points $x \in E$ for which every neighbourhood $N_x \in \mathcal{T}$ s.t. $x \in N_x$ has positive measure:

$$\operatorname{supp}(\mu) := \{x \in E \mid \forall N_x \in \mathcal{T} : (x \in N_x \implies \mu(N_x) > 0)\}.$$

Henceforth, any mentions of random variables will map into Borel spaces. Nothing really changes from the measurability point of view — we simply also have the ability to discuss topological concepts such as closure, support, separability, metrisability etc.

Remarks 4.4.7

- An equivalent definition is the largest $C \in \mathcal{B}_E$ (with respect to inclusion) such that every open set which has non-empty intersection with C has positive measure i.e. the largest C s.t. $\forall U \in \mathcal{T}$:
$$U \cap C \neq \emptyset \implies \mu(U \cap C) > 0.$$
- Under some **regularity**[6] **assumptions**, the support of μ is the smallest closed set of full measure

  i.e. the smallest closed set with μ-"almost empty" complement

  i.e. it is a closed set C s.t.

  - $\mu(E \setminus C) = 0$
  - If C_1 is closed and $\mu(E \setminus C_1) = 0$, then $C \subseteq C_1$.

  If the support exists, it's equal to the intersection of all closed subsets of E whose complements have measure 0.

[6] I should fill the blank in on these **assumptions**. I read somewhere (Math Stack Exchange perhaps?) that such assumptions will generally hold for the measures in these notes.

CHAPTER 5

Measuring Functions (Integral)

So far we've defined measures — set functions which associate a general notion of "size" to measurable sets. Last chapter was about measurable functions. We can use measures to define a "way" to compute the "size" of measurable functions — the *Lebesgue integral with respect to a given non-negative measure*. Historically speaking, the Riemann integral (and other integrals at the time) were inadequate — most acutely in the sense that we can quite easily formulate a function of interest for which its Riemann integral does not exist. To remedy such woes, Henri Lebesgue and Emile Borel formulated a more satisfactory theory of integration. Let's explore an example where the Riemann-Darboux integral falls short.

5.1 Historical Shortcomings

Suppose that one poses the question "What is the probability that a number between 0 and 1 (inclusive) chosen uniformly at random is a rational number?" If you've already seen the uniform distribution, great. If not, it doesn't matter — just keep in mind that the integral of an indicator function over a subset of $\mathbb{R}$ gives the length (Lebesgue measure) of that set i.e.

$$b - a = \lambda([a,b]) = \int_a^b \mathrm{d}x = \int_{\mathbb{R}} \mathbb{1}_{[a,b]}(x)\,\mathrm{d}x,$$

and the problem of finding that probability is closely linked (for reasons that will become apparent later on) to the integral

$$\text{“}\int_0^1 \mathbb{1}_{\mathbb{Q}}(x)\,\mathrm{d}x\text{.”}$$

Just because we can write such a collection of symbols down, that doesn't mean we've written something meaningful down. At first glance, one would expect that because both $\mathbb{Q}$ and $\mathbb{R}\setminus\mathbb{Q}$ are dense in $\mathbb{R}$, but $\mathbb{Q}$ is countable (and therefore its complement is uncountable), that a satisfactory theory of integration would assign the value 0 to the integral of $\mathbb{1}_{\mathbb{Q}\cap[0,1]}(\cdot)$. This function vanishes "almost everywhere" on $[0,1]$. Recall the definition of the Riemann-Darboux[1] integral:

Definition 5.1.1

- A **partition P of $[a,b]$** is a set of $n \geqslant 1$ points $\{t_i\}_{i=0}^n$ s.t. $a := t_0 < t_1 < \ldots < t_{n-1} < t_n =: b$.
- The lower sum $L(f,P)$ of f with respect to a partition $P = \{t_i\}_{i=0}^n$ of $[a,b]$ is defined by

$$L(f,P) = \sum_{i=0}^{n-1}(t_{i+1} - t_i) \inf_{t\in[t_i,t_{i+1}]} f(t).$$

  The upper sum $U(f,P)$ is defined analogously — replace inf with sup.
- A function $f\colon [a,b] \to \mathbb{R}$ is **Riemann-Darboux integrable** if for every partition $P = \{t_i\}_{i=0}^n$ of $[a,b]$, we have that

$$\sup_P L(f,P) = \inf_P U(f,P). \tag{Riem}$$

In the case that (Riem) holds, the common value is denoted by $\int_a^b f(x)\,\mathrm{d}x$ and called the Riemann integral of f over $[a,b]$.

[1] This formulation is equivalent to the original Riemann integral but is easier to use.

For any partition P, any interval $[t_{i-1}, t_i]$ will contain rational and irrational points (by the completeness of $\mathbb{R}$) so the infimum and supremum of $\mathbb{1}_{\mathbb{Q}}$ over $[t_{i-1}, t_i]$ will be 0 and 1 respectively. Thus, $L(f, P) = 0$ and $U(f, P) = 1$. These two values don't agree and so the Riemann integral of $\mathbb{1}_{\mathbb{Q}}$ over $[0, 1]$ does not exist.

To extend the woes by 1, we can construct a sequence to demonstrate that the class of Riemann integrable functions aren't closed under pointwise limits! Let $q_1, q_2, \ldots$ denote an enumeration of $\mathbb{Q} \cap [0, 1]$, and consider the sequence of step functions f_n defined by

$$f_n(x) = \begin{cases} 1, & \text{if } x \in \{q_1, \ldots, q_n\} \\ 0, & \text{otherwise.} \end{cases}$$

Each $f_n(x)$ is Riemann integrable over $[0, 1]$ with Riemann integral 0, the $f_n(x)$ converge pointwise to $\mathbb{1}_{\mathbb{Q}\cap[0,1]}$ pointwise, but we know the latter isn't Riemann integrable.

We can somewhat see the issue here — the Riemann integral is insufficient insofar as it depends on the limit of Riemann sums which requires us to partition the *domain* of functions we wish to integrate. Instead of partitioning the domain, Henri Lebesgue and Emile Borel developed a theory of integration that relies on partitioning the range of a function, and fitting a sequence of simple functions to serve as the analogue to the 'upper' and 'lower sums' in the Riemann case. The benefit of this approach is that we may consider functions with more exotic domains (e.g. $\mathbb{Q}$) as long as we have some notion of their "size." This is the theory of Lebesgue integration.

I wrote the following section after watching Nicolas Lanchier's **video lectures** which are based on his book *Stochastic Modeling* [2]. I've tried to fill in some details so all errors henceforth are especially my own.

5.2 Definition of the Lebesgue Integral

We're going to define the Lebesgue integral of a **measurable** function $X : \Omega \to \mathbb{R}$ with respect to any non-negative measure μ in four steps:

Step 1 - (Measurable) Indicator Functions

Let $A \subseteq \Omega$ and $X = \mathbb{1}_A$ be measurable. We define the Lebesgue integral of $X = \mathbb{1}_A$ with respect to μ by

$$\int_\Omega \mathbb{1}_A \, \mathrm{d}\mu := \mu(A).$$

This definition is inspired by the earlier comment on the Riemann integral of an indicator function. Is our definition well-defined? Since $X = \mathbb{1}_A \in \mathrm{Meas}_{\mathcal{F}, \mathcal{B}_{\mathbb{R}}}(\Omega; \mathbb{R})$, for any $B \in \mathcal{B}_{\mathbb{R}}$, $\mathbb{1}_A^{-1}(B) \in \mathcal{F}$. In particular, $A = \mathbb{1}_A^{-1}(\{1\}) \in \mathcal{F}$. Thus, $\mu(A)$ is well-defined.

Step 2 - Simple (Measurable) Functions

A function $X : \Omega \to [0, +\infty)$ is called **simple** iff X is measurable and the image of X is a finite subset of $[0, +\infty)$ i.e.

$$\begin{aligned} \mathrm{Im}(X) &:= \{x \in \mathbb{R} : \exists \omega \in \Omega \text{ s.t. } X(\omega) = x\} \\ &= \{a_1, \ldots, a_m\}. \end{aligned}$$

Equivalently, note that if $A_i := X^{-1}(\{a_i\})$, then $\Omega = \bigsqcup_{i=1}^m A_i$ and so X may be written as a finite linear combination of indicator functions

$$X = \sum_{i=1}^m a_i \mathbb{1}_{A_i}.$$

This is called the **standard representation of X**.

Then we define

$$\int_\Omega X \,\mathrm{d}\mu = \int_\Omega \left(\sum_{i=1}^m a_i \mathbb{1}_{A_i}\right) \mathrm{d}\mu \coloneqq \sum_{i=1}^m a_i \int_\Omega \mathbb{1}_{A_i} \,\mathrm{d}\mu \overset{\text{Step 1}}{=} \sum_{i=1}^m a_i \mu(A_i).$$

We must check to see if this definition is well-defined i.e. that the value of the integral is the same whichever *representation*[2] one chooses for X.

Proof. Let X be simple, measurable and have two representations i.e. $\{A_i\}_{i=1}^n \subseteq \mathcal{F} \supseteq \{B_j\}_{j=1}^m$ are two partitions of Ω and $a_i, b_j > 0$ are s.t.

$$X = \sum_{i=1}^n a_i \mathbb{1}_{A_i} = \sum_{j=1}^m b_i \mathbb{1}_{B_i}.$$

Now we construct a refinement $\{C_{ij}\}_{i,j=1}^n$ of both partitions by $C_{ij} = A_i \cap B_j \in \mathcal{F}$. Then we may compute the integral of X. For the first representation:

$$\begin{aligned}
\int_\Omega X \,\mathrm{d}\mu &= \sum_{i=1}^n a_i \mu(A_i) \\
&= \sum_{i=1}^n a_i \mu(A_i \cap \Omega) \\
&= \sum_{i=1}^n a_i \mu\Big(A_i \cap \bigsqcup_{j=1}^m B_j\Big) \\
&= \sum_{i=1}^n a_i \mu\Big(\bigsqcup_{j=1}^m (A_i \cap B_j)\Big) \\
&= \sum_{i=1}^n a_i \mu\Big(\bigsqcup_{j=1}^m C_{ij}\Big) \\
&= \sum_{i=1}^n a_i \sum_{j=1}^m \mu(C_{ij})
\end{aligned}$$

An analogous computation shows that

$$\int_\Omega X \,\mathrm{d}\mu = \sum_{j=1}^m b_i \mu(B_i) = \sum_{j=1}^m b_j \sum_{i=1}^n \mu(C_{ij}).$$

If $\omega \in C_{ij}$, then $X(\omega) = a_i = b_j$. Thus, we may define $c_{ij} = a_i = b_j$ and note that both integral representations give the same value. ■

Step 3 - Non-Negative (Measurable) Functions

We've already seen that for the Riemann integral, subdividing the domain and trying to approximate a non-negative measurable function like $\mathbb{1}_{\mathbb{Q}}(x)$ with a simple function (corresponding to the Riemann sum) failed when taking the limit. Instead, we sub-divide the range of X, and approximate X from below by a non-negative, simple, and measurable function. By Step 2, the Lebesgue integral of such simple functions are non-negative and well-defined. The Lebesgue integral of X with respect to μ is then defined as

$$\int_\Omega X \,\mathrm{d}\mu \coloneqq \sup_{0 \leqslant s \leqslant X} \int_\Omega s \,\mathrm{d}\mu.$$

This supremum is still well-defined if equal to $+\infty$, and the supremum definition avoids any dependence on how one may go about constructing a sequence of simple functions that converges to X.

[2] By representation of a simple function, one refers to the collection of pairs (A_i, a_i) that define the linear combination. If A_i and A_j are disjoint, for example, then $\mathbb{1}_{A_i \sqcup A_j} = \mathbb{1}_{A_i} + \mathbb{1}_{A_j}$ for example.

Step 4 - Any Measurable Functions

We follow a similar approach in that we wish to use Step 3 to define our integral. We can write X as the difference of two non-negative functions, namely the positive and negative parts of X, denoted by $X^+ := \max(X, 0)$ and $X^- := \max(0, -X)$ respectively.

$$X = X^+ - X^-$$

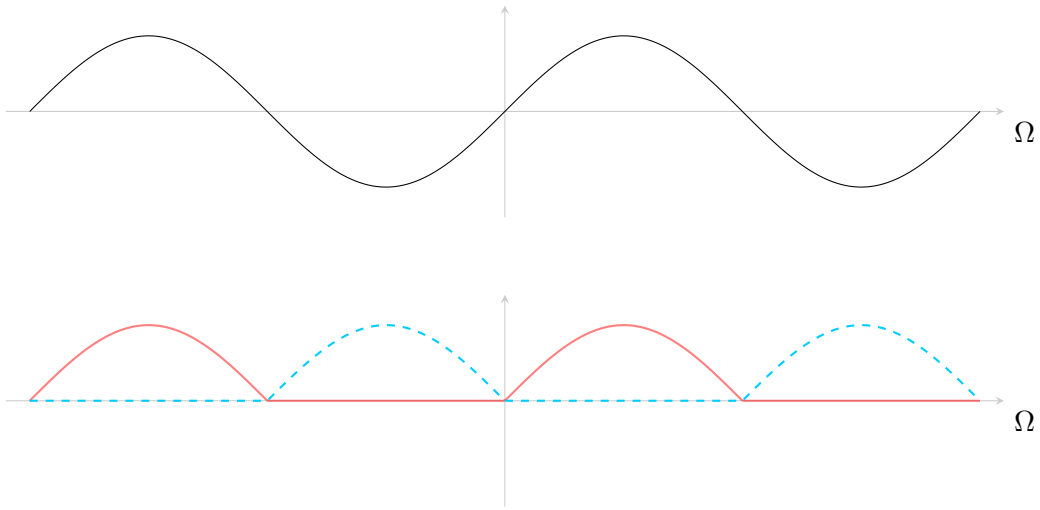


Figure 5.1: Graphs of the measurable function X, and its positive and negative parts.

Then we define the Lebesgue integral of X with respect to the non-negative measure μ by

$$\int_\Omega X \, \mathrm{d}\mu = \int_\Omega X^+ \, \mathrm{d}\mu - \int_\Omega X^- \, \mathrm{d}\mu$$

If the integrals of the positive and negative parts are both equal to $+\infty$, we have the problem of $\infty - \infty$ not being well-defined. Thus, our definition requires that both integrals (which are non-negative) are finite and we impose the condition that the Lebesgue integral of $|X|$ is finite i.e.

$$\int_\Omega X^+ \, \mathrm{d}\mu - \int_\Omega X^- \, \mathrm{d}\mu = \int_\Omega |X| \, \mathrm{d}\mu < \infty.$$

Since the sum of the non-negative integrals is finite, the difference is certainly finite and so $\int_\Omega X \, \mathrm{d}\mu$ makes sense.

Definition 5.2.1 We say that X **is Lebesgue integrable** with respect to a non-negative measure μ (on $\mathcal{F}$), also denoted by $X \in L^1(\Omega, \mathcal{F}, \mu)$, if

$$\int_\Omega |X| \, \mathrm{d}\mu < \infty.$$

Lemma 5.2.2 (Properties of the Lebesgue Integral)

- If $f, g \geqslant 0$ are measurable and equal μ-a.e. then

$$\int_\Omega f \, \mathrm{d}\mu = \int_\Omega g \, \mathrm{d}\mu.$$

- Linearity: If $f, g \in L^1(\Omega, \mathcal{F}, \mu)$, and $a, b \in \mathbb{R}$, then $af + bg$ is Lebesgue integrable and

$$\int_\Omega (af + bg) \, \mathrm{d}\mu = a \int_\Omega f \, \mathrm{d}\mu + b \int_\Omega g \, \mathrm{d}\mu.$$

- Monotonicity: If $f \leqslant g$, then
$$\int_\Omega f \,\mathrm{d}\mu \leqslant \int_\Omega g \,\mathrm{d}\mu.$$

Remarks 5.2.3

- In the context where $X : \Omega \to \mathbb{R}$ is a map from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into the measurable space $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ i.e. X is a real-valued random variable, then the Lebesgue integral of X with respect to $\mathbb{P}$ is defined as the **expectation of X**:
$$\int_\Omega X \,\mathrm{d}\mathbb{P} =: \mathbb{E}(X).$$
Furthermore, the expectation inherits all the properties of the Lebesgue integral.

- If X is Riemann integrable, then X is Lebesgue integrable with respect to the Lebesgue measure λ, and these two integrals coincide.

Calling back to our motivating example at the beginning, we wanted to find the probability that a number chosen uniformly at random in $(0, 1)$ is rational. Recall that we can enumerate $\mathbb{Q} \cap [0, 1]$ with a countable set $\{q_i\}_{i\in\mathbb{N}}$.

$$\begin{aligned}
\mathbb{P}(X \in \mathbb{Q} \cap [0,1]) &= \int_\Omega \mathbb{1}_{\mathbb{Q}\cap[0,1]} \,\mathrm{d}\lambda \\
&= \lambda(\mathbb{Q} \cap [0,1]) \\
&= \lambda\Big(\bigsqcup_{i=1}^{\infty} \{q_i\}\Big) \quad \text{by enumerating } (\mathbb{Q}\cap[0,1])\text{-countable} \\
&= \sum_{i=1}^{\infty} \lambda(\{q_i\}) \quad \text{by } \sigma\text{-additivity of } \lambda \\
&= 0 \quad \text{because } \lambda([q_i, q_i]) = 0.
\end{aligned}$$

5.3 MCT, Fatou, DCT, and Fubini

The results are stated here and their proofs (with some neat visualisations) can be found in **Appendix C**.

The first two theorems give conditions under which one can exchange limit and integral for a sequence of measurable functions.

Theorem 5.3.1 (Monotone Convergence Theorem) Let $\{X_n\}_{n\in\mathbb{N}}$ be a non-decreasing sequence of non-negative measurable functions with pointwise limit X. Then, X is measurable and
$$\lim_{n\to\infty} \int_\Omega X \,\mathrm{d}\mu = \int_\Omega \Big(\lim_{n\to\infty} X_n\Big) \,\mathrm{d}\mu.$$

The statement of the theorem can be rephrased as a probability statement by letting $\mu = \mathbb{P}$. In this case, the statement is that for a non-decreasing sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables with pointwise limit X, then X is also a random variable and $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Theorem 5.3.2 (Dominated Convergence Theorem) Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of measurable functions dominated by some integrable function and with pointwise limit X. Then X is integrable,
$$\lim_{n\to\infty} \int_\Omega |X_n - X| \,\mathrm{d}\mu = 0 \quad \text{and} \quad \lim_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu = \int_\Omega \Big(\lim_{n\to\infty} X_n\Big) \,\mathrm{d}\mu.$$

- Proving the DCT requires the following theorem that doesn't guarantee any convergence but we can still define a lim inf and the result is an inequality:

  Theorem 5.3.3 (Fatou's Lemma) Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of non-negative measurable functions. Then
  $$\int_\Omega \left(\liminf_{n\to\infty} X_n\right) \mathrm{d}\mu \leqslant \liminf_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu.$$

This third theorem gives sufficient conditions under which two integrals, of a measurable function of two variables, can commute:

Theorem 5.3.4 (Fubini-Tonelli Theorem) Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be two σ-finite measure spaces. Define:

- $\Omega = \Omega_1 \times \Omega_2$
- $\mathcal{F} = \sigma(\{A \times B \colon A \in \mathcal{F}_1,\, B \in \mathcal{F}_2\})$
- The unique product measure μ on $\mathcal{F}$ satisfying $\mu(A \times B) = \mu_1(A)\mu_2(B)$

For every measurable function $X \colon \Omega \to \mathbb{R}$ that is either non-negative or integrable on a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$,
$$\int_\Omega X \,\mathrm{d}\mu = \int_{\Omega_1}\int_{\Omega_2} X \,\mathrm{d}\mu_2 \,\mathrm{d}\mu_1 = \int_{\Omega_2}\int_{\Omega_1} X \,\mathrm{d}\mu_1 \,\mathrm{d}\mu_2.$$

The case where f is non-negative is often stated as a separate theorem called Tonelli's theorem. In that case, we don't need to check if f is integrable. One may also Fubini's theorem stated as above but without the non-negativity criterion.

5.4 Absolute Continuity & Radon-Nikodým Derivative

The Radon-Nikodym Theorem justifies the existence and uniqueness of conditional expectation. It also does many other things that we'll see more immediately in this chapter!

The motivation for the theorem is as follows: Let $\phi \colon (\Omega, \mathcal{F}, \mu) \to (\mathbb{R}, \mathcal{B}_\mathbb{R})$ be a non-negative, measurable function. Define
$$\nu(A) \coloneqq \int_A \phi \,\mathrm{d}\mu = \int_\Omega \phi \mathbb{1}_A \,\mathrm{d}\mu$$
for all $A \in \mathcal{F}$. Since $\phi \geqslant 0$ and measurable, the integral is well-defined and non-negative (possibly $+\infty$). Now take a pairwise disjoint collection $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ whose union we denote by $A \in \mathcal{F}$. Then
$$\begin{aligned}
\nu\Big(\bigsqcup_{n=1}^{\infty} A_n\Big) &\coloneqq \int_\Omega \phi \mathbb{1}_{\bigsqcup_{n=1}^\infty A_n} \,\mathrm{d}\mu \\
&= \int_\Omega \sum_{n=1}^{\infty} \phi \mathbb{1}_{A_n} \,\mathrm{d}\mu \\
&= \int_\Omega \lim_{k\to\infty} \sum_{n=1}^{k} \phi \mathbb{1}_{A_n} \,\mathrm{d}\mu
\end{aligned}$$

This is a non-decreasing sequence of measurable functions so we can apply the MCT.

$$\begin{aligned}
&\overset{\text{MCT}}{=} \lim_{k\to\infty} \int_\Omega \sum_{n=1}^{k} \phi \mathbb{1}_{A_n} \,\mathrm{d}\mu \\
&\overset{\text{lin.}}{=} \lim_{k\to\infty} \sum_{n=1}^{k} \int_\Omega \phi \mathbb{1}_{A_n} \,\mathrm{d}\mu \\
&= \sum_{n=1}^{\infty} \int_\Omega \phi \mathbb{1}_{A_n} \,\mathrm{d}\mu \\
&=: \sum_{n=1}^{\infty} \nu(A_n)
\end{aligned}$$

Therefore, ν is σ-additive and so $\nu\colon \mathcal{F} \to [0, +\infty]$ is a measure on $\mathcal{F}$.

The Radon-Nikodym Theorem is related to the above notion but instead it tells us that if we have two measures μ and ν (on the same space), whether or not one can find a ϕ s.t.

$$\forall A \in \mathcal{F}: \quad \nu(A) = \int_A \phi \,\mathrm{d}\mu. \tag{$\star$}$$

Can we always find a ϕ for which ($\star$) is true for every $A \in \mathcal{F}$?

Answer No! If there exists an $A \in \mathcal{F}$ s.t. $\mu(A) = 0$ and $\nu(A) > 0$, then no matter which ϕ is chosen, the Lebesgue integral of ϕ over A w.r.t. μ is an integral over a set of μ-measure zero so the integral must always be zero.

This motivates the definition of absolute continuity as a way to avoid the existence of such an $A \in \mathcal{F}$ with $\mu(A) = 0$ but $\nu(A) > 0$.

Definition 5.4.1 **ν is absolutely continuous with respect to μ**, denoted[3] $\nu \ll \mu$, if for all $A \in \mathcal{F}$:

$$\mu(A) = 0 \implies \nu(A) = 0.$$

This condition of absolutely continuity and the σ-finiteness of μ and ν are precisely the things that guarantee there exists (in a μ-almost everywhere sense) a ϕ s.t. ($\star$) is true:

Theorem 5.4.2 (Radon-Nikodym Theorem) Let μ and ν be two σ-finite measures on the same measurable space s.t. $\nu \ll \mu$. Then there exsts a function $\phi \in \mathrm{Meas}_{\mathcal{F},\mathcal{B}_\mathbb{R}}(\Omega\,; [0, +\infty))$ s.t. for all $A \in \mathcal{F}$:

$$\nu(A) = \int_A \phi \,\mathrm{d}\mu$$

which is unique in the sense that two such versions of ϕ are equal μ-almost everywhere. We denote ϕ by $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ and call it the Radon-Nikodym derivative of ν w.r.t μ.

5.5 Pushforward Measure & Change of Variables

Recall that for a real-valued random variable X on a probability space, one can define a probability measure $\mathbb{P}_X$ on the Borel σ-algebra $\mathcal{B}_\mathbb{R}$ by setting

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B) = \mathbb{P}\big(X^{-1}(B)\big) = \int_\Omega \mathbb{1}_{X^{-1}(B)} \,\mathrm{d}\mathbb{P} \quad \text{for all } B \in \mathcal{B}_\mathbb{R}.$$

(The last expression with the Lebesgue integral is new and follows from $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A)$.) This is called the measure induced by X in measure theory, and the distribution of X in probability theory. To study a random variable in practice, probabilists don't work with the probability measure $\mathbb{P}$ but with its distribution $\mathbb{P}_X$ (a nicer, friendlier measure defined on $\mathcal{B}_\mathbb{R}$) by using the following result:

[3] This notation is inspired by the case $\mu(A) = 0$ forcing $\nu(A) = 0$.

Theorem 5.5.1 (Change of Variables Formula) Let $X \in \mathrm{Meas}_{\mathcal{F},\mathcal{B}_\mathbb{R}}(\Omega;\mathbb{R})$ be a random variable, and $h\colon \mathbb{R} \to \mathbb{R}$ a measurable function. If h is non-negative or integrable, then

$$\int_\Omega h(X)\,\mathrm{d}\mathbb{P} =: \mathbb{E}(h(X)) = \int_\mathbb{R} h\,\mathrm{d}\mathbb{P}_X.$$

Remarks 5.5.2

- The greyed out expression is simply the definition of $\mathbb{E}(h(X))$. The integral is well-defined because $h(X)$ is a measurable function on Ω taking values in $\mathbb{R}$. It's not exactly clear what it means to integrate with respect to a probability measure but we do know how to integrate on $\mathbb{R}$, and so this theorem connects the two notions.
- Why are we using the quantity $\mathbb{E}(h(X))$?

  We can use it to generate many things e.g. $h = \mathrm{id}_\mathbb{R}$ gives us $\mathbb{E}(X)$, and if we then consider $X = \mathbb{1}_A$ for some measurable set A, then we can recover $\mathbb{P}_X(A)$. Higher order moments can also be generated with $\mathbb{E}(h(X))$.

Proof. The steps of the proof will follow the construction of the integral:

1. Assume first that $h = \mathbb{1}_B$ for some $B \in \mathcal{B}_\mathbb{R}$. Note that

$$\begin{aligned}(h \circ X)(\omega) = (\mathbb{1}_B \circ X)(\omega) &= \begin{cases} 1, & \text{if } X(\omega) \in B \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} 1, & \text{if } \omega \in X^{-1}(B) \\ 0, & \text{otherwise} \end{cases} \\ &:= \mathbb{1}_{X^{-1}(B)}(\omega).\end{aligned}$$

   Then

$$\mathbb{E}(h(X)) = \mathbb{E}(\mathbb{1}_B(X)) = \mathbb{E}\big(\mathbb{1}_{X^{-1}(B)}\big) = \mathbb{P}\big(X^{-1}(B)\big) = \mathbb{P}_X(B) = \int_\mathbb{R} \mathbb{1}_B\,\mathrm{d}\mathbb{P}_X$$

2. Let h be a simple, measurable function with standard representation

$$h = \sum_{i=1}^n a_i \mathbb{1}_{A_i}.$$

   Then

$$\begin{aligned}\mathbb{E}(h(X)) &= \mathbb{E}\left(\Big(\sum_{i=1}^n a_i \mathbb{1}_{A_i}\Big) \circ X\right) \\ &= \mathbb{E}\left(\sum_{i=1}^n a_i (\mathbb{1}_{A_i} \circ X)\right) \\ &= \sum_{i=1}^n a_i \mathbb{E}(\mathbb{1}_{A_i} \circ X) \\ &= \sum_{i=1}^n a_i \int_\mathbb{R} \mathbb{1}_{A_i}\,\mathrm{d}\mathbb{P}_X \quad \text{by Step 1} \\ &= \int_\mathbb{R} \underbrace{\sum_{i=1}^n a_i \mathbb{1}_{A_i}}_{=:h}\,\mathrm{d}\mathbb{P}_X \quad \text{by linearity.}\end{aligned}$$

3. Let $h \geqslant 0$ be measurable. Then there exists a non-decreasing sequence of simple, measurable functions with pointwise limit h e.g.

$$s_n(x) = \min(n, 2^{-n}\lfloor 2^n h(x) \rfloor) \quad \forall x \in \mathbb{R}.$$

This sequence was the 2^{-n} sub-division of the range in the construction of the Lebesgue Integral. By the Monotone Convergence Theorem,

$$\begin{aligned}
\mathbb{E}(h(X)) &= \mathbb{E}\Big(\Big(\lim_{n\to\infty} s_n\Big) \circ X\Big) \\
&= \mathbb{E}\Big(\lim_{n\to\infty} (s_n \circ X)\Big) \\
&\overset{\text{MCT}}{=} \lim_{n\to\infty} \mathbb{E}((s_n \circ X)) \\
&= \lim_{n\to\infty} \int_{\mathbb{R}} s_n \, \mathrm{d}\mathbb{P}_X \quad \text{by Step 2 since } s_n\text{-simple} \\
&\overset{\text{MCT}}{=} \int_{\mathbb{R}} \lim_{n\to\infty} s_n \, \mathrm{d}\mathbb{P}_X \\
&= \int_{\mathbb{R}} h \, \mathrm{d}\mathbb{P}_X
\end{aligned}$$

4. Let h be integrable i.e. $\int_{\mathbb{R}} |h| \, \mathrm{d}\mathbb{P}_X < \infty$. Any measurable function that changes sign can be written as its positive part minus its negative part, both non-negative measurable functions.

$$\begin{aligned}
\mathbb{E}(h(X)) &:= \mathbb{E}\big((h^+ - h^-) \circ X\big) \\
&= \mathbb{E}\big(h^+ \circ X\big) - \mathbb{E}\big(h^- \circ X\big) \\
&= \int_{\mathbb{R}} h^+ \, \mathrm{d}\mathbb{P}_X - \int_{\mathbb{R}} h^- \, \mathrm{d}\mathbb{P}_X \quad \text{by Step 3} \\
&= \int_{\mathbb{R}} (h^+ - h^-) \, \mathrm{d}\mathbb{P}_X \quad \text{by linearity} \\
&=: \int_{\mathbb{R}} h \, \mathrm{d}\mathbb{P}_X
\end{aligned}$$

■

5.6 Types of Random Variables

5.6.1 Absolutely Continuous

Now we combine the notion of distribution and the Radon-Nikodym derivative to properly define probability mass and density functions of random variables.

Definition 5.6.1 A (real-valued) random variable X is called **absolutely continuous** if its distribution $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$.

We now only need the σ-finiteness of both measures $\mathbb{P}_X$ and λ in order to apply the Radon-Nikodym theorem. Since $\mathbb{P}_X$ is a probability measure, it's finite and hence σ-finite on $\mathcal{B}_{\mathbb{R}}$. The Lebesgue measure on $\mathbb{R}$ is σ-finite because $\mathbb{R}$ can be covered with a countable sequence of measurable sets each with finite λ-measure e.g.

$$\mathbb{R} = \bigcup_{n\in\mathbb{N}} [-n, n]$$

where each $[-n, n] \in \mathcal{B}_{\mathbb{R}}$ and $\lambda([-n, n]) = 2n < \infty$.

By the Radon-Nikodým Theorem (**5.4.2**), $\exists\lambda$-a.e. unique $\phi_X := \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\lambda}$ which is what we interpret as the **probability density function** of X. Applying **Theorem 5.5.1** and then the **Theorem 5.4.2** gives:

$$\mathbb{E}(h(X)) \overset{\text{CVF}}{=} \int_{\mathbb{R}} h \, \mathrm{d}\mathbb{P}_X \overset{\text{RNT}}{\underset{?}{=}} \int_{\mathbb{R}} h\phi_X \, \mathrm{d}\lambda.$$

It's not immediately clear to me why ? is true. Let's verify it.

Claim (**KEB**) Let $\mathbb{P}_X \ll \lambda$, and $h\colon \mathbb{R} \to \mathbb{R}$ be a non-negative measurable or integrable function. Then:

$$\int_{\mathbb{R}} h \,\mathrm{d}\mathbb{P}_X = \int_{\mathbb{R}} h\phi_X \,\mathrm{d}\lambda$$

Proof.

1. Consider $h = \mathbb{1}_B$ for some $B \in \mathcal{B}_{\mathbb{R}}$.

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{1}_B \,\mathrm{d}\mathbb{P}_X &=: \mathbb{P}_X(B) \\ &\overset{\text{RNT}}{=} \int_B \phi \,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} \phi \mathbb{1}_B \,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} h\phi \,\mathrm{d}\lambda\end{aligned}$$

2. Let h be simple, non-negative and measurable. By linearity, we have that

$$\begin{aligned}\int_{\mathbb{R}} \sum_{i=1}^n a_i \mathbb{1}_{A_i} \,\mathrm{d}\mathbb{P}_X &= \sum_{i=1}^n a_i \int_{\mathbb{R}} \mathbb{1}_{A_i} \,\mathrm{d}\mathbb{P}_X \\ &= \sum_{i=1}^n a_i \int_{\mathbb{R}} \mathbb{1}_{A_i} \phi \,\mathrm{d}\lambda \quad \text{by Step 1} \\ &= \int_{\mathbb{R}} \sum_{i=1}^n a_i \mathbb{1}_{A_i} \phi \,\mathrm{d}\lambda \\ &:= \int_{\mathbb{R}} h\phi \,\mathrm{d}\lambda\end{aligned}$$

3. Let $h \geqslant 0$ be measurable. Then h is the pointwise limit of a non-decreasing sequence of simple functions $\{s_n\}_{n\in\mathbb{N}}$.

$$\begin{aligned}\int_{\mathbb{R}} h \,\mathrm{d}\mathbb{P}_X &= \int_{\mathbb{R}} \lim_{n\to\infty} s_n \,\mathrm{d}\mathbb{P}_X \\ &\overset{\text{MCT}}{=} \lim_{n\to\infty} \int_{\mathbb{R}} s_n \,\mathrm{d}\mathbb{P}_X \\ &= \lim_{n\to\infty} \int_{\mathbb{R}} s_n \phi \,\mathrm{d}\lambda \quad \text{by Step 2} \\ &\overset{\text{MCT}}{=} \int_{\mathbb{R}} \lim_{n\to\infty} s_n \phi \,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} h\phi \,\mathrm{d}\lambda\end{aligned}$$

In the case that h is integrable i.e. $\int_{\mathbb{R}} |h| \,\mathrm{d}\mathbb{P}_X < \infty$, I'm guessing that one can decompose $h = h^+ - h^-$ and then

$$\begin{aligned}\int_{\mathbb{R}} h \,\mathrm{d}\mathbb{P}_X &= \int_{\mathbb{R}} h^+ \,\mathrm{d}\mathbb{P}_X - \int_{\mathbb{R}} h^- \,\mathrm{d}\mathbb{P}_X \\ &= \int_{\mathbb{R}} h^+\phi \,\mathrm{d}\lambda - \int_{\mathbb{R}} h^-\phi \,\mathrm{d}\lambda \quad \text{by Step 3} \\ &= \int_{\mathbb{R}} (h^+ - h^-)\phi \,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} h\phi \,\mathrm{d}\lambda.\end{aligned}$$

■

So overall we went from having no idea how to integrate with respect to $\mathbb{P}$, to still not really knowing how to integrate with respect to $\mathbb{P}_X$ in practice, and finally arrived at an integral with respect to λ which we know how to compute if the function happens to be Riemann integrable:

$$\begin{aligned}\mathbb{E}(h(X)) &:= \int_\Omega h(X)\mathrm{d}\mathbb{P} \\ &\overset{\text{CVF}}{=} \int_{\mathbb{R}} h\mathrm{d}\mathbb{P}_X \\ &\overset{\text{KEB}}{=} \int_{\mathbb{R}} h\phi_X \mathrm{d}\lambda\end{aligned}$$

This final integral is something that we use a lot in practice i.e. to compute the probability or expected value involving a random variable X, we need only look at the integral of some function of X multiplied by the density function. Indeed, when we take $h = \mathrm{id}$ the expression

$$\mathbb{E}(X) = \int_{\mathbb{R}} \mathrm{id}\, \phi_X \,\mathrm{d}\lambda$$

has the natural interpretation as a one-number summary of the centre of its distribution — a measure of its central tendency.

Further to this, we can recover a formula to calculate pushforward probabilities:

$$\begin{aligned}\mathbb{P}_X(B) = \mathbb{P}\big(X^{-1}(B)\big) =: \mathbb{E}\big(\mathbb{1}_{X^{-1}(B)}\big) = \mathbb{E}(\mathbb{1}_B(X)) = \int_\Omega \mathbb{1}_B(X)\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} \mathbb{1}_B\,\mathrm{d}\mathbb{P}_X = \int_{\mathbb{R}} \mathbb{1}_B\phi_X\,\mathrm{d}\lambda \\ = \int_B \phi_X\,\mathrm{d}\lambda.\end{aligned}$$

5.6.2 Discrete

Definition 5.6.2 A random variable $X \in \mathrm{Meas}_{\Omega,E}(\mathcal{F};\mathcal{E})$ is called **discrete** if its distribution has at most countably infinite support.

Remarks (!!)

- In particular, if X has at most countably infinite image, then the support of $\mathbb{P}_X$ is certainly at most countably infinite and so X is discrete.
- If X is discrete, then it follows that $\mathbb{P}_X$ is absolutely continuous with respect to the counting measure $\mu_{\mathrm{supp}(X)}$ on $\mathrm{supp}(X) \subseteq E$ defined by

$$\mu_{\mathrm{supp}(X)}(B) = \mathrm{card}(B \cap \mathrm{supp}(X)).$$

Proof. Let $\mu_{\mathrm{supp}(X)}(B) = 0$. Then $B \cap \mathrm{supp}(X) = \emptyset$ and so $B \subseteq E \setminus \mathrm{supp}(X)$. This means that for every $x \in B$, there exists some open neighbourhood $\mathcal{T}_E \ni N_x \ni x$ s.t. $\mathbb{P}_X(N_x) = 0$. The $\{N_x\}_{x\in B}$ form an open cover of B. Since $(E,\mathcal{E})$ is a Borel space, the underlying topological space is Polish i.e. metrisable and separable. This tells us that it's also second-countable so we can extract a countable subcover $\{N_{x_i}\}_{i\in\mathbb{N}}$ of the $\{N_x\}_{x\in B}$. Therefore,

$$\mathbb{P}_X(B) \leqslant \mathbb{P}_X\Big(\bigcup_{i\in\mathbb{N}} N_{x_i}\Big) = \sum_{i\in\mathbb{N}} \mathbb{P}_X(N_{x_i}) = 0.$$

Thus, $\mathbb{P}_X \ll \mu_{\mathrm{supp}(X)}$. ■

Following the same logic as the absolutely continuous case, if we use the Change of Variables formula **5.5.1**, followed by the **KEB** Claim (which involves the Radon-Nikodym Theorem, noting that $\mathbb{P}_X \ll \mu_{\mathrm{supp}(X)}$) then we arrive at

$$\mathbb{E}(h(X)) := \int_\Omega h\,\mathrm{d}\mathbb{P} \overset{\textbf{CVF}}{=} \int_{\mathbb{R}} h\,\mathrm{d}\mathbb{P}_X \overset{\textbf{KEB}}{=} \int_{\mathbb{R}} h\phi_X\,\mathrm{d}\mu_{\mathrm{supp}(X)} = \sum_{x\in\mathrm{supp}(X)} h(x)\phi_X(x)$$

Example If we let $h = \mathrm{id}$, then

$$\mathbb{E}(h(X)) = \mathbb{E}(X) = \sum_{x \in \mathrm{supp}(X)} x\phi_X(x)$$

which is the usual formula[4] of the expected value of a discrete random variable X. Note that the condition for $\mathbb{E}(X)$ (an integral w.r.t the counting measure) existing is that of the sum being absolutely convergent.[5]

We interpret $\phi_X(x)$ as $\mathbb{P}_X(\{x\})$ and so ϕ_X is the familiar **probability mass function** of a discrete random variable X.

Example We can also recover a familiar expression for the pushforward probability of some $B \in \mathcal{B}_{\mathbb{R}}$. Note that $\mu_{\mathrm{supp}(X)}$ is σ-finite on $\mathrm{supp}(X)$. Thus, the Radon-Nikodým theorem applies and:

$$\begin{aligned}
\mathbb{P}_X(B) = \mathbb{P}\big(X^{-1}(B)\big) =: \mathbb{E}\big(\mathbb{1}_{X^{-1}(B)}\big) = \mathbb{E}(\mathbb{1}_B(X)) = \int_\Omega \mathbb{1}_B(X)\,\mathrm{d}\mathbb{P} &= \int_{\mathbb{R}} \mathbb{1}_B\,\mathrm{d}\mathbb{P}_X \\
&= \int_{\mathbb{R}} \mathbb{1}_B\phi_X\,\mathrm{d}\mu_{\mathrm{supp}(X)} \\
&= \int_B \phi_X\,\mathrm{d}\mu_{\mathrm{supp}(X)} \\
&= \sum_{x \in B \cap \mathrm{supp}(X)} \phi_X(x) \\
&= \sum_{x \in B} \phi_X(x).
\end{aligned}$$

So in summary:

- An absolutely continuous random variable is characterised by a density function.
- A discrete random variable is characterised by a probability mass function.

But both density and mass functions can be viewed under a unifying framework as the Radon-Nikodym derivatives of $\mathbb{P}_X$ with respect to a reference measure — a measure for which we can better understand $\int \text{function}\,\mathrm{d}(\text{ref. measure})$ e.g. λ and $\mu_{\mathbb{N}}$ respectively.

[4] If the numbers $\phi_X(x) = \mathbb{P}_X(\{x\})$ are regarded as masses at the points x, then $\mathbb{E}[X]$ represents the position of the centre of gravity of their sum. Sometimes it's said that X has an atom or point mass of size $p_X(x)$ at x.

[5] Absolute convergence guarantees that we avoid the situation where the value of the infinite sum $\mathbb{E}(X)$ can change upon reordering the x by Riemann's rearrangement theorem.

CHAPTER 6

Random Vectors

The natural extension to a single random variable X, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and representing one observation from a single trial of an experiment, is a collection of random variables $X_1, \ldots, X_n$ formalising the act of carrying out n successive trials of an experiment.

Say we perform the same experiment n times. Each trial (represented by X_i) results in observing x_i. Once we complete the successive trials, we have a collection of observations $x_1, \ldots, x_n$. The event corresponding to this collection is of course equal to the intersection of all possible outcomes that result in each X_i outputting x_i:

$$\{X_1 = x_1, \ldots, X_n = x_n\} \coloneqq \bigcap_{i=1}^{n} \{X_i = x_i\} \coloneqq \bigcap_{i=1}^{n} \{\omega \in \tilde{\Omega} \colon X_i(\omega) = x_i\}.$$

Note that the underlying probability space $\tilde{\Omega}$ (and its corresponding σ-algebra) is abstract, and changes its structure to support the random variables defined on it. This is a subtle point that will be expanded on later (in **Chapter 13**) but in this case, we think of $\tilde{\Omega}$ as a collection of tuples $(\omega_1, \ldots, \omega_n)$ representing the n successive outcomes of the experiment, and each X_i as the composition of $\mathbf{X}$ with the natural i^{th} coordinate projection. Each ω_i certainly lives in Ω so $\tilde{\Omega} = \Omega^n$. Henceforth, I will commit the heinous crime against precision of notation and refer to the underlying space as Ω.

We represent the collection of outcomes by an n-tuple $(x_1, \ldots, x_n)$. The reason for this is because the mathematical object that formalises the idea of modelling multiple random variables at a time is called a random vector:

Definition 6.0.1 A **real random vector** on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a map $\mathbf{X} \colon \Omega \to \mathbb{R}^n$ that is $(\mathcal{F}, \mathcal{B}_{\mathbb{R}^n})$-measurable i.e. $\forall B \in \mathcal{B}_{\mathbb{R}^n}$, $\mathbf{X}^{-1}(B) \in \mathcal{F}$.

Proposition 6.0.2 Any mapping $\mathbf{X} \colon \Omega \to \mathbb{R}^n$ must be of the form $\omega \longmapsto (X_1(\omega), \ldots, X_n(\omega))$, where each component $X_i(\omega) \in \mathbb{R}$ for all $i = 1, \ldots, n$. We shall see that $\mathbf{X}$ is $(\mathcal{F}, \mathcal{B}_{\mathbb{R}^n})$-measurable iff for each $i = 1, \ldots, n$, X_i is $(\mathcal{F}, \mathcal{B})$-measurable.

The reason for this can be found in Section 1.2 from Folland [8, pp. 22–24]. The cliffnotes version is that for a collection of non-empty measurable spaces $\{(X_\alpha, \mathcal{F}_\alpha)\}_{\alpha \in A}$, if we denote by

$$\pi_\alpha \colon \underbrace{\prod_{\alpha \in A} X_\alpha}_{=: X} \to X_\alpha$$

the canonical projections (sending every element of X to its X_α component), then we may define a σ-algebra on X called the **product σ-algebra on X**, denoted $\otimes_{\alpha \in A} \mathcal{F}_\alpha$. This σ-algebra is generated by the collection of pre-images under π_α of all measurable sets $E_\alpha \in \mathcal{F}_\alpha$:

$$\{\pi_\alpha^{-1}(E_\alpha) \colon E_\alpha \in \mathcal{F}_\alpha, \alpha \in A\}.$$

The sets $\pi_\alpha^{-1}(E_\alpha)$ or the finite intersection of such sets are called cylindrical sets. (The product σ-algebra is the coarsest one that makes the coordinate maps f_α measurable.) If A is countable, then $\otimes_{\alpha \in A} \mathcal{F}_\alpha$ is generated by

$$\left\{ \prod_{\alpha \in A} E_\alpha \colon E_\alpha \in \mathcal{F}_\alpha \right\}$$

because $\pi_\alpha^{-1}(E_\alpha) = \prod_{\beta \in A} E_\beta$ for $E_\alpha \in \mathcal{F}_\alpha$ where $E_\beta = X_\beta$ for $\beta \neq \alpha$. With another technical lemma and under the assumption that $X_1, \ldots, X_n$ are separable metric spaces ($A = \{1, \ldots, n\}$-countable), we have as a corollary that

$$\mathcal{B}_{\mathbb{R}^n} = \bigotimes_{i=1}^{n} \mathcal{B}_{\mathbb{R}}$$

from which we can prove the following proposition:

Proposition 6.0.3 (Proposition 2.4 [8]) Let $(Y, \mathcal{A})$ and $(X_\alpha, \mathcal{F}_\alpha)$ for $\alpha \in A$ be measurable spaces, $X = \prod_{\alpha\in A} X_\alpha$, $\mathcal{F} = \otimes_{\alpha\in A}\mathcal{F}_\alpha$ be the product σ-algebra on X, and $\pi_\alpha \colon X \to X_\alpha$ be the canonical projections. Then $f \colon X \to Y$ is $(\mathcal{A}, \mathcal{F})$-measurable iff $f_\alpha = \pi_\alpha \circ f$ is $(\mathcal{A}, \mathcal{F}_\alpha)$-measurable for all $\alpha \in A$.

Let $(Y, \mathcal{A}) = (\Omega, \mathcal{F})$ and $(X, \otimes_{\alpha\in A}\mathcal{F}_\alpha) = (\mathbb{R}^n, \otimes_{i=1}^n \mathcal{B}_\mathbb{R}) = (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ and the claim follows.

Indeed, by **Proposition 6.0.3** a random vector is simply an n-tuple $\mathbf{X} = (X_1, \ldots, X_n)$ of random variables and a realisation of $\mathbf{X}$ is denoted by an n-tuple $(x_1, \ldots, x_n)$.

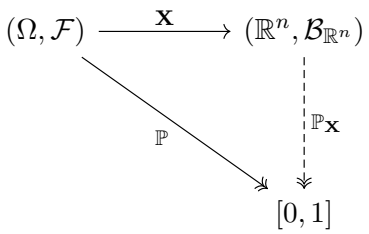
An important thing to mention here is that, in greater generality, one can consider a random vector of random variables $X_i \colon (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$ where each $(E_i, \mathcal{E}_i)$ is a Borel space. By **Theorem 4.4.4**, the product space

$$\left(E = \prod_{i=1}^n E_i,\ \mathcal{E} = \bigotimes_{i=1}^n \mathcal{E}_i\right)$$

is also a Borel space, and hence $\mathbf{X} = (X_1, \ldots, X_n)$ maps into the Borel space $(E, \mathcal{E})$.

6.1 Probability Distribution of $\mathbf{X}$

Similar to the case of a (univariate) real random variable, every real random vector is a measurable map that induces a unique probability distribution $\mathbb{P}_\mathbf{X}$ as the push-forward measure $\mathbf{X}_\sharp\mathbb{P}$ of $\mathbb{P}$ via $\mathbf{X}$:



The **probability distribution of a real random vector** $\mathbf{X} = (X_1, \ldots, X_n) \colon \Omega \to \mathbb{R}^n$ (under $\mathbb{P}$) is the map $\mathbb{P}_\mathbf{X} \colon \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ defined for all $B \in \mathcal{B}_{\mathbb{R}^n}$ by:

$$\begin{aligned}\mathbb{P}_\mathbf{X}(B) &:= \mathbb{P}\big(\mathbf{X}^{-1}(B)\big) = \mathbb{P}(\mathbf{X} \in B)\\ &= \mathbb{P}(\{\omega \in \Omega \colon \mathbf{X}(\omega) \in B\}).\end{aligned}$$

We also call $\mathbb{P}_\mathbf{X}$ the **joint probability distribution** of the random variables $X_1, \ldots, X_n$.

6.1.1 Joint CDF

We follow the beaten path once more by discussing the ways in which we can specify the probability distribution of a random variable. We already have that $\mathcal{B}_{\mathbb{R}^n} = \otimes_{i=1}^n \mathcal{B}_\mathbb{R}$ because $\mathbb{R}^n$ is a separable metric space. A technical lemma states that if each $\mathcal{F}_\alpha$ in $(X_\alpha, \mathcal{F}_\alpha)$ is generated by $\mathcal{C}_\alpha$, A is countable and $X_\alpha \in \mathcal{C}_\alpha$ for all $\alpha \in A$, then $\mathcal{F} = \otimes_{\alpha\in A}\mathcal{F}_\alpha$ is generated by

$$\mathcal{F}_2 = \left\{\prod_{\alpha\in A} E_\alpha \colon E_\alpha \in \mathcal{C}_\alpha\right\}.$$

Let $X_\alpha = \mathbb{R}$, $\mathcal{F}_\alpha = \mathcal{B}_\mathbb{R}$ for all $\alpha \in A = \{1, \ldots, n\}$. If we consider $\mathcal{C}_\alpha = \{(-\infty, x] \colon x \in \mathbb{R}\}$ for all $\alpha \in A$, it follows that an alternative specification (that again, always exists) of our $\mathbb{P}_\mathbf{X}$ is the cumulative distribution function of $\mathbf{X}$:

Definition 6.1.1 **All** random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ admit a **joint cumulative distribution function** $F_{\mathbf{X}} \colon \mathbb{R}^n \to [0,1]$ defined for any $\boldsymbol{x} \in \mathbb{R}^n$ by:

$$\begin{aligned} F_{\mathbf{X}}(\boldsymbol{x}) &\coloneqq \mathbb{P}_{\mathbf{X}}((-\infty, x_1] \times \ldots \times (-\infty, x_n]) \\ &= \mathbb{P}(\{\omega \in \Omega \colon X_1(\omega) \leqslant x_1, \ldots, X_n(\omega) \leqslant x_n) \\ &= \mathbb{P}(\{\mathbf{X} \leqslant \boldsymbol{x}\}), \end{aligned}$$

where $(a_1, \ldots, a_n) = \boldsymbol{a} \leqslant \boldsymbol{x} = (x_1, \ldots, x_n)$ iff $a_i \leqslant x_i$ for all $i = 1, \ldots, n$.

Example 6.1.2 (Bivariate Joint Distribution) Let $(X, Y) \colon \Omega \to \mathbb{R}^2$ be a random vector. Suppose that $F \colon \mathbb{R}^2 \to [0,1]$ is a function satisfying the following:

1. $F(-\infty, y) = F(-\infty, -\infty) = F(x, -\infty) = 0$, $F(\infty, \infty) = 1$
2. $(x, y) \leqslant (u, v) \implies F(x, y) \leqslant F(u, v)$
3. Continuity from above: $\lim_{u,v \downarrow 0} F(x + u, y + v) = F(x, y)$
4. If $(x, y) \leqslant (u, v)$ then $F(u, v) - F(u, y) - F(x, v) + F(x, y) \geqslant 0$

These are sufficient conditions to construct a probability space and random variable such that F is the cumulative distribution function of (X, Y).

Proof Sketch. We can define a set function ν on the semi-algebra of half-open rectangles by $\mu((x, u] \times (y, v]) \coloneqq F(u, v) - F(u, y) - F(x, v) + F(x, y)$ and extend it via the Carathéodory extension theorem to a Borel measure (which we call μ). Then the random variable (X, Y) on $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \mu)$ has distribution function F.

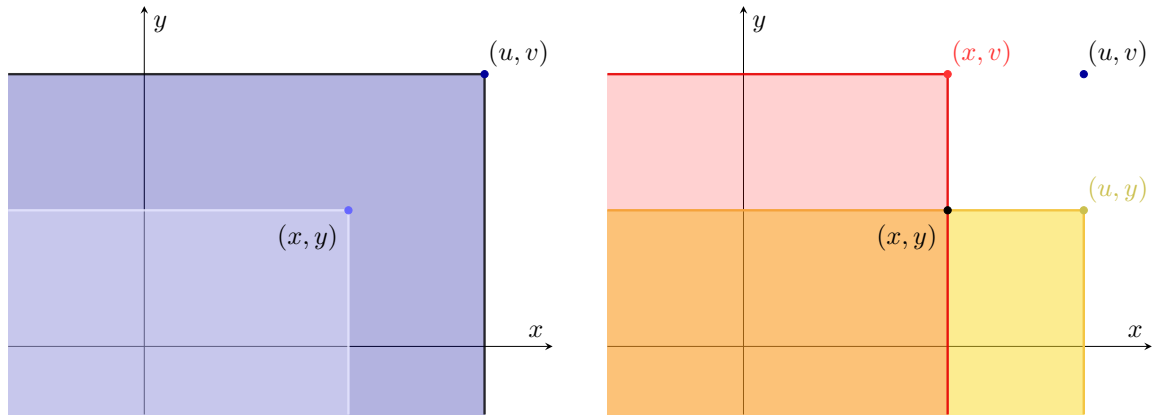
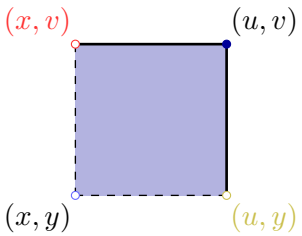


Figure 6.1: A visualisation of the doubly-subtracted region $(-\infty, x] \times (-\infty, y]$, represented by the orange overlap of the red and yellow regions, needing to be compensated for by adding $F(x, y)$ to $F(u, v) - F(u, y) - F(x, v)$.

The resulting half-open rectangle looks like:



■

I believe that this argument extends analogously to an n-dimensional random vector.

6.2 Types of Random Vectors

This section is largely analogous to the earlier discussion of types of random variables.

Definition 6.2.1 A real random vector $\mathbf{X}\colon \Omega \to \mathbb{R}^n$ is **discrete** if[a] there exists an at most countably infinite set Γ such that $\mathbb{P}_{\mathbf{X}}(\Gamma) = 1$.

[a] This is an equivalent way to say $\operatorname{supp}(\mathbb{P}_{\mathbf{X}})$ is at most countably infinite.

Remarks 6.2.2 If $\mathbf{X}\colon \Omega \to \mathbb{R}^n$ is discrete, then:

- We say that $X_1, \dots, X_n$ are **jointly discrete**,
- By the RNT, $\mathbb{P}_{\mathbf{X}}$ admits a **joint probability mass function** $p_{\mathbf{X}}\colon \mathbb{R}^n \to [0,1]$ defined for all $\boldsymbol{x} \in \mathbb{R}^n$ by
$$p_{\mathbf{X}}(\boldsymbol{x}) = \mathbb{P}_{\mathbf{X}}(\{x\}) = \mathbb{P}(\{\mathbf{X} = \boldsymbol{x}\}).$$
Thus, for any $B \in \mathcal{B}_{\mathbb{R}^n}$:
$$\begin{aligned}\mathbb{P}_{\mathbf{X}}(B) &= \mathbb{P}(\{\mathbf{X} \in B\}) \\ &= \mathbb{P}(\{\omega \in \Omega\colon \mathbf{X}(\omega) \in B\}) \\ &= \mathbb{P}\Big(\bigcup_{\boldsymbol{x}\in B} \{\omega \in \Omega\colon \mathbf{X}(\omega) = \boldsymbol{x}\}\Big) \\ &= \sum_{\boldsymbol{x}\in B} p_{\mathbf{X}}(\boldsymbol{x}).\end{aligned}$$

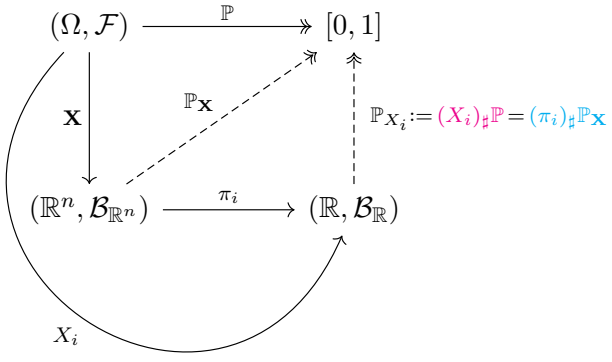
Definition 6.2.3 A real random vector $\mathbf{X}$ is **absolutely continuous** if its probability distribution $\mathbb{P}_{\mathbf{X}}$ is absolutely continuous with respect to the Lebesgue measure $\lambda_{\mathbb{R}^n}$.

Remarks 6.2.4 If $\mathbf{X} = (X_1, \dots, X_n)\colon \Omega \to \mathbb{R}^n$ is absolutely continuous, then:

- We say that the components $X_1, \dots, X_n$ are **jointly absolutely continuous**.
- By the RNT, $\mathbb{P}_{\mathbf{X}}$ admits a ($\lambda_{\mathbb{R}}$-a.e. unique) density function $f_{\mathbf{X}}\colon \mathbb{R}^n \to \mathbb{R}_+$ s.t. $\forall B \in \mathcal{B}_{\mathbb{R}^n}$:
$$\mathbb{P}_{\mathbf{X}}(B) = \int_B f_{\mathbf{X}}\, \mathrm{d}\lambda_{\mathbb{R}^n}.$$
This density $f_{\mathbf{X}}$ is called the **joint probability density of X**.

6.3 Marginal Distributions

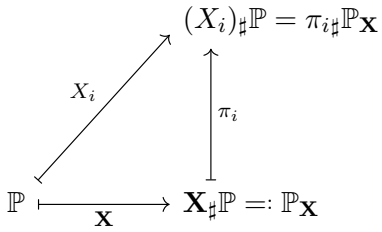
The i^{th} **marginal distribution of X** is the distribution of the i^{th} component of the random vector $\mathbf{X} = (X_1, \dots, X_n)$, denoted $\mathbb{P}_{X_i}$. Let $\pi_i\colon \mathbb{R}^n \to \mathbb{R}$ denote the projection map onto the i^{th} coordinate. Then $X_i := \pi_i \circ \mathbf{X}$.



From the diagram above, $\mathbb{P}_{X_i}$ is the pushforward $(X_i)_\sharp(\mathbb{P})\colon \mathcal{B}_\mathbb{R} \to [0,1]$ of $\mathbb{P}$ via X_i. This map is defined for any $B \in \mathcal{B}_\mathbb{R}$ by

$$\begin{aligned}((X_i)_\sharp\mathbb{P})(B) &= \mathbb{P}\big(X_i^{-1}(B)\big)\\ &= \mathbb{P}\big((\pi_i \circ \mathbf{X})^{-1}(B)\big)\\ &= \mathbb{P}\big(\mathbf{X}^{-1}(\pi_i^{-1}(B)\big)\\ &= \mathbb{P}_\mathbf{X}\big(\pi_i^{-1}(B)\big) = ((\pi_i)_\sharp\mathbb{P}_X)(B)\\ &= \mathbb{P}_\mathbf{X}\big(\underbrace{\mathbb{R}\times\ldots\times\mathbb{R}}_{(i-1)\text{ times}}\times B\underbrace{\times\mathbb{R}\times\ldots\times\mathbb{R}}_{(n-i)\text{ times}}\big)\end{aligned}$$

Note that the diagram also suggests that $(X_i)_\sharp(\mathbb{P})$ should be equal to the pushforward of $\mathbb{P}_\mathbb{X}$ via π_i (which is indeed a $(\mathcal{B}_{\mathbb{R}^n} = \otimes_{i=1}^n \mathcal{B}_\mathbb{R}, \mathcal{B}_\mathbb{R})$-measurable function with respect to the product σ-algebra on $\mathbb{R}^n$). This relationship reveals itself as the $(\pi_i)_\sharp\mathbb{P}_X$ term in the above equation above and I make this explicit in the following pushfoward diagram:



The process of marginalisation isolates the behaviour of a random vector's component from its joint distribution by "summing" over all values that can be assumed by the other components (excluding the i^{th}) of the random vector. This is represented by the Cartesian product $\mathbb{R}\times\ldots\times\mathbb{R}\times B\times\mathbb{R}\times\ldots\times\mathbb{R}$. In essence, this process collapses the random vector onto the single random variable X_i of interest.

- If $\mathbf{X}\colon \Omega \to \mathbb{R}^n$ is discrete with joint probability mass function $p_\mathbf{X}(x_1,\ldots,x_n)$, the marginal probability mass function of X_i is given by

$$p_{X_i}(x_i) = \sum_{x_j\colon j\neq i} p_\mathbf{X}(x_1,\ldots,x_n).$$

- If $\mathbf{X}\colon \Omega \to \mathbb{R}^n$ is absolutely continuous with density $f_\mathbf{X}(x_1,\ldots,x_n)$, then for each $i = 1,\ldots,n\colon X_i$ is absolutely continuous and its marginal probability density is given by

$$f_{X_i}(x_i) = \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f_\mathbf{X}(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)\,\mathrm{d}x_1\ldots\mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1}\ldots\mathrm{d}x_n.$$

Proof.

$$\begin{aligned}&\mathbb{P}(X_i \leqslant t)\\ &= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\int_{-\infty}^{t}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f_\mathbf{X}(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)\,\mathrm{d}x_1\ldots\mathrm{d}x_n\\ &= \int_{-\infty}^{t}\left(\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} f_\mathbf{X}(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)\,\mathrm{d}x_1\ldots\mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1}\ldots\mathrm{d}x_n\right)\mathrm{d}x_i\end{aligned}$$

Since $f_\mathbf{X}$ is non-negative and measurable on $\mathbb{R}^n$, we were permitted to use Tonelli's theorem to exchange the order of integration. The integrand on the last line is precisely $f_{X_i}(x_i)$. ■

In the spirit of marginalisation, for each i the joint CDF of $\mathbf{X}$ is related to the marginal CDF of X_i by

$$F_{X_i}(x_i) = \lim_{\substack{x_j \to +\infty \\ \forall j \neq i}} F_{\mathbf{X}}(x_1, \ldots, x_n).$$

6.4 Conditional Distributions

The opposite of marginalisation is the idea of conditioning — a concept that represents how we update our beliefs given prior information e.g. computing the value of an event occurring given the knowledge that some events have already happened. Some keywords I had at the time of writing this chapter were **conditional density**, and **disintegration theorem**. I answer what these are affirmatively in **Chapter 17**.

For the time being, I offer a very brief summary of the undergraduate-level treatment of conditioning because I use it in an **exercise 7** before I give the proper treatment in a later chapter.

When discussing marginalisation, we looked at a single component of a random vector in isolation. If instead, we wish to consider the values some components of a random vector will take given that the remaining components have already assumed a value in some set, then we're in the realms of conditioning.

For the following exposition, let $\mathbf{X} = (X_1, X_2) \colon \mathbb{R}^2 \to \mathbb{R}$ be a 2-dimensional real random vector.

6.4.1 Jointly Discrete

Suppose that X_1 and X_2 are jointly discrete with joint probability mass function $p(x_1, x_2)$, and marginal probability mass functions $p_1(x_1)$ and $p_2(x_2)$. The **discrete conditional probability function of X_1 given $\{X_2 = x_2\}$** is given[1] by

$$p(x_1 \mid x_2) = \frac{p_{\mathbf{X}}(x_1, x_2)}{p_2(x_2)}$$

provided that $p_2(x_2) > 0$.

6.4.2 Jointly Absolutely Continuous

In the case that X_1 and X_2 are jointly absolutely continuous, we can't define a conditional probability function of X_1 given $X_2 = x_2$ since both $\{X_1 = x_1\}$ and $\{X_2 = x_2\}$ are evens with probability zero. Instead, a useful and consistent definition for a conditional density function can be found if we're interested in probabilities of the form

$$\mathbb{P}(\{X_1 \leqslant x_1 \,|\, X_2 = x_2\}) =: F(x_1 \mid x_2)$$

which as a function of x_1 for a fixed x_2 is called the **conditional distribution function of X_1 given $\{X_2 = x_2\}$**.

If we could multiply $F(x_1|x_2)$ by each value of $\mathbb{P}(\{X_2 = x_2\})$, we would hopefully recover $F_{X_1}(x_1)$ but each value of $\mathbb{P}(\{X_2 = x_2\})$ is 0. We can do something analogous by instead multiplying by the density $f_{X_2}(x_2)$ and integrating:

$$F_{X_1}(x_1) = \int_{-\infty}^{+\infty} F(x_1|x_2) f_{X_2}(x_2) \, \mathrm{d}x_2$$

This term in blue can be thought of as the approximate probability that X_2 takes on a value within a small interval about x_2 and then the integral can be thought of as a generalised sum.

[1] This expression is analogous to the conditional probability expression for events $\mathbb{P}(A \,|\, B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$.

From previous considerations:

$$\begin{aligned} F_{X_1}(x_1) &= \int_{-\infty}^{x_1} f_{X_1}(t_1)\,\mathrm{d}t_1 \\ &= \int_{-\infty}^{x_1} \left(\int_{-\infty}^{\infty} f_{\mathbf{X}}(t_1, x_2)\,\mathrm{d}x_2 \right) \mathrm{d}t_1 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{x_1} f_{\mathbf{X}}(t_1, x_2)\,\mathrm{d}t_1\,\mathrm{d}x_2 \end{aligned}$$

For now, this is the best we can do. Watch this space in **Chapter 17**.

6.5 Independence of Random Variables

We shall build on our prior definition of mutually $\mathbb{P}$-independent sub-σ-algebras. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

Definition 6.5.1 The **σ-algebra $\sigma(\mathbf{X})$ generated by a random vector $\mathbf{X}\colon \Omega \to \mathbb{R}^n$** is the coarsest[2] σ-algebra on Ω that makes $\mathbf{X}$ measurable i.e.

$$\sigma(\mathbf{X}) := \{\mathbf{X}^{-1}(B)\colon B \in \mathcal{B}_{\mathbb{R}^n}\}.$$

Any collection of **random variables $X_1, \ldots, X_n$** defined on the same probablity space **are mutually $\mathbb{P}$-independent** if $\{\sigma(X_i)\}_{1 \leqslant i \leqslant n}$ are mutually $\mathbb{P}$-independent.

Example 6.5.2 Let's follow the definition for a random vector $\mathbf{X} = (X_1, X_2)$. Suppose that X_1 and X_2 are independent. This means that $\sigma(X_1) = \{X_1^{-1}(A)\colon A \in \mathcal{B}_{\mathbb{R}}\}$ and $\sigma(X_2) = \{X_2^{-1}(A)\colon A \in \mathcal{B}_{\mathbb{R}}\}$ are mutually $\mathbb{P}$-independent i.e. for any choice of $A \in \sigma(X_1)$ and $B \in \sigma(X_2)$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Since $A \in \sigma(X_1)$ and $B \in \sigma(X_2)$, there exist $B_1 \in \mathcal{B}_{\mathbb{R}} \ni B_2$ s.t. $A = X_1^{-1}(B_1)$ and $B = X_2^{-1}(B_2)$. This means that:

- $\mathbb{P}(A \cap B) = \mathbb{P}\big(X_1^{-1}(B_1) \cap X_2^{-1}(B_2)\big) = \mathbb{P}\big(\mathbf{X}^{-1}(B_1 \times B_2)\big) = \mathbb{P}_{(X_1, X_2)}(B_1 \times B_2)$
- $\mathbb{P}(A) \cdot \mathbb{P}(B) = \mathbb{P}\big(X_1^{-1}(B_1)\big) \cdot \mathbb{P}\big(X_2^{-1}(B_2)\big) = \mathbb{P}_{X_1}(B_1) \cdot \mathbb{P}_{X_2}(B_2)$.

Finally, we recover the familiar expression

$$\mathbb{P}_{(X_1, X_2)}(B_1 \times B_2) = \mathbb{P}_{X_1}(B_1) \cdot \mathbb{P}_{X_2}(B_2).$$

This example extends to the general case:

Theorem 6.5.3 Suppose that the components X_i of a real random vector $\mathbf{X} = (X_1, \ldots, X_n)$ are independent. Then the joint distribution $\mathbb{P}_{\mathbf{X}}$ of $\mathbf{X}$ is the[3] product measure $\bigotimes_{i=1}^n \mathbb{P}_{X_i}\colon \bigotimes_{i=1}^n \mathcal{B}_{\mathbb{R}} \to [0, 1]$ defined for every $B_i \in \mathcal{B}$ by

$$\left(\otimes_n \mathbb{P}_{X_i}\right)(B_1 \times \ldots \times B_n) = \mathbb{P}_{X_1}(B_1) \cdot \ldots \cdot \mathbb{P}_{X_n}(B_n).$$

As before, we can choose a generating set for each σ-algebra i.e. consider the half-open rays $(-\infty, x]$ that generate $\mathcal{B}_{\mathbb{R}}$. Then we have a representation of the above theorem in terms of CDFs! For simplicity, we consider a 2-dimensional real random vector $\mathbf{X} = (X_1, X_2)$.

Lemma 6.5.4 Let $\mathbf{X}$ have CDF $F_{\mathbf{X}}(x_1, x_2)$ and the marginals of X_1 and X_2 be $F_{X_i}(x_i)$ for $i = 1, 2$ respectively. Then X_1 and X_2 are (mutually $\mathbb{P}$-)independent iff

$$F_{\mathbf{X}}(x_1, x_2) = F_{X_1}(x_1) \cdot F_{X_2}(x_2)$$

for every pair of real numbers (x_1, x_2).

[2] Smallest.

[3] This is unique because of σ-finiteness of the constituent prodands.

If X_1 and X_2 are not independent, they are said to be dependent.
There are also example expressions for discrete and absolutely continuous random vectors $\mathbf{X}$:

- If $\mathbf{X}$ is discrete with joint probability mass function $p_{\mathbf{X}}(x_1, x_2)$, and marginal mass functions $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$ respectively, then X_1 and X_2 are independent iff for every pair of real numbers (x_1, x_2):
$$p_{\mathbf{X}}(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2).$$
- If $\mathbf{X}$ is absolutely continuous, admitting joint probability density function $f_{\mathbf{X}}(x_1, x_2)$, then X_1 and X_2 are absolutely continuous with respective marginal density functions $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. Then X_1 and X_2 are independent iff for every pair of real numbers (x_1, x_2):
$$f_{\mathbf{X}}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2).$$

Here are two very general but powerful theorems that will help to simplify our lives many times over for the remainder of these notes:

Theorem 6.5.5 (Theorem 4.6.11 [1]) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random vectors defined on the same probability space. Then $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent random vectors if and only if there exist non-negative real-valued functions $g_i(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$ such that the joint density f of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ factors into
$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = g_1(\boldsymbol{x}_1) \cdot \ldots \cdot g_n(\boldsymbol{x}_n).$$

Theorem 6.5.6 (Theorem 4.6.12 [1]) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be mutually independent random vectors defined on the same probability space. Let $g_i(\boldsymbol{x}_i)$ be a function of only $\boldsymbol{x}_i$ for $i = 1, \ldots, n$. Then the random vectors $U_i = g_i(\mathbf{X}_i)$ for $i = 1, \ldots, n$ are mutually independent.

CHAPTER 7

Averages, Dispersion, and Correlation

The word 'average' can mean many things. There are multiple ways to take an average of numbers; mean, median, mode, weighted averages etc. Without clarification, average in these notes means 'mean' or 'expected value.'

When discussing types of random variables, we've already seen the expectation of a random variable as a Lebesgue integral that quantifies its central tendency. We can also calculate measures of variability (variance, standard deviation etc.) to describe how spread out a random variable's distribution is:

7.1 Variance

The *variance* of a random variable is a measure of how concentrated around the mean, the distribution of a random variable is.

Definition 7.1.1 The **variance of a random variable X** is defined as

$$\operatorname{Var}(X) := \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$$

if $\mathbb{E}\big(X^2\big)$ exists and is finite (the reason for which is explained by the following corollary).

Corollary 7.1.2 By the linearity of expectation:

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

If the random variables under consideration are independent, we can sometimes simplify the work involved in finding expectations.

Theorem 7.1.3 Let Y_1 and Y_2 be independent random variables on the same probability space and let $g(Y_1)$ and $h(Y_2)$ be functions of **only** Y_1 and **only** Y_2 respectively. Then

$$\mathbb{E}(g(Y_1)h(Y_2)) = \mathbb{E}(g(Y_1))\mathbb{E}(h(Y_2))$$

provided the expectations exist.

Continuous Case. Let $f_{\mathbf{Y}}(y_1, y_2)$ be the joint density of $\mathbf{Y} = (Y_1, Y_2)$. The product $g(Y_1)h(Y_2)$ is a function of Y_1 and Y_2 — call it $i(Y_1, Y_2)$. Then

$$\begin{aligned}\mathbb{E}(g(Y_1)h(Y_2)) = \mathbb{E}(i(Y_1, Y_2)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} i(y_1, y_2) f_{\mathbf{Y}}(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2) f_{\mathbf{Y}}(y_1, y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y_1)h(y_2) f_{Y_1}(y_1) f_{Y_2}(y_2)\,\mathrm{d}y_1\,\mathrm{d}y_2 \quad \text{by independence} \\ &= \int_{-\infty}^{\infty} h(y_2) f_{Y_2}(y_2) \underbrace{\left(\int_{-\infty}^{\infty} g(y_1) f_{Y_1}(y_1)\,\mathrm{d}y_1\right)}_{=:\ \mathbb{E}(g(Y_1))} \mathrm{d}y_2 \\ &= \mathbb{E}(g(Y_1)) \int_{-\infty}^{\infty} h(y_2) f_{Y_2}(y_2)\,\mathrm{d}y_2 \\ &= \mathbb{E}(g(Y_1))\mathbb{E}(h(Y_2))\end{aligned}$$

■

7.2 Covariance

Intuitively, we think of the dependence of two random variables Y_1 and Y_2 as implying that one variable, say Y_1, either increases or decreases as Y_2 changes. Two measures of dependence are the 'covariance' between two random variables and the correlation coefficient of two random variables. Consider the following scatter plot of observations of (Y_1, Y_2):

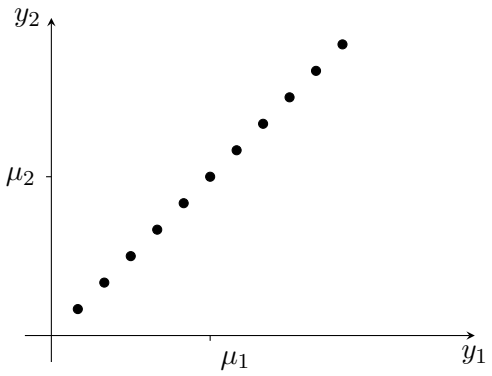
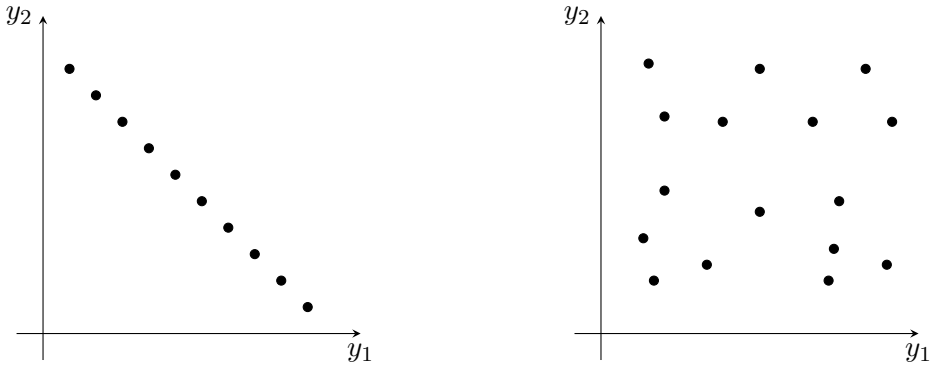


Figure 7.1: A figure of (seemingly?) dependent observations for (y_1, y_2).

- Suppose that $\mu_1 = \mathbb{E}(Y_1)$ and $\mu_2 = \mathbb{E}(Y_2)$.
- Measure the deviations $(y_1 - \mu_1)$ and $(y_2 - \mu_2)$.
- For any point (y_1, y_2) in the figure, both deviations assume the same sign so their product is positive.
- On average (over all observations), the product is large and positive.

Now consider the following scatter plots:



- Had the relationship sloped downward and to the right, as shown in the scatter plot on the left, all corresponding pairs of deviation would have had opposite signs and the average value of their product $(y_1 - \mu_1)(y_2 - \mu_2)$ would have been a large negative number.
- For the plot on the right, we can see that there's little dependence between Y_1 and Y_2. The corresponding deviations will have the same sign for some points and opposite signs for others. The average of these products will be some value around zero.

Definition 7.2.1 The average value of $(Y_1 - \mu_1)(Y_2 - \mu_2)$ provides a measure of the dependence between Y_1 and Y_2. This quantity $\mathbb{E}((Y_1 - \mu_1)(Y_2 - \mu_2))$ is called the **covariance of Y_1 and Y_2**, denoted $\mathrm{Cov}(Y_1, Y_2)$.

A convenient computational formula for covariance is:

$$\begin{aligned}\operatorname{Cov}(Y_1, Y_2) &= \mathbb{E}((Y_1 - \mu_1)(Y_2 - \mu_2)) \\ &= \mathbb{E}(Y_1Y_2 - Y_1\mu_2 - \mu_1Y_2 + \mu_1\mu_2) \\ &= \mathbb{E}(Y_1Y_2) - \mu_2\mathbb{E}(Y_1) - \mu_1\mathbb{E}(Y_2) + \mu_1\mu_2 \\ &= \mathbb{E}(Y_1Y_2) - \mu_2\mu_1 - \mu_1\mu_2 + \mu_1\mu_2 \\ &= \mathbb{E}(Y_1Y_2) - \mu_1\mu_2\end{aligned}$$

Corollary 7.2.2 (Independent Random Variables Are Uncorrelated) If Y_1 and Y_2 are random variables that are independent, then $\operatorname{Cov}(Y_1, Y_2) = 0$.

Proof.

$$\begin{aligned}\operatorname{Cov}(Y_1, Y_2) &= \mathbb{E}(Y_1Y_2) - \mu_1\mu_2 \\ &= \mathbb{E}(Y_1)\mathbb{E}(Y_2) - \mu_1\mu_2 \quad \text{by independence} \\ &= 0\end{aligned}$$

■

The converse is not true. Uncorrelated $\not\Longrightarrow$ Independence.

7.3 Correlation

It's difficult to employ covariance as an absolute measure of dependence because its value depends on the scale of measurements. This problem can be rectified by standardising its value: [4]

The **correlation coefficient** ρ is defined by

$$\rho := \frac{\operatorname{Cov}(Y_1, Y_2)}{\sigma_1\sigma_2}$$

where σ_i is the standard deviation of Y_i.

Proposition 7.3.1 The correlation coefficient $\rho \in [-1, 1]$.

Proof.

Exercise 1 (5.137 [6], 5.167 [7])

(a) Show that $(\mathbb{E}(Y_1Y_2))^2 \leqslant \mathbb{E}\big((Y_1)^2\big)\mathbb{E}\big((Y_2)^2\big)$.

Hint: Observe that $\mathbb{E}\big((tY_1 - Y_2)^2\big) \geqslant 0$ for any $t \in \mathbb{R}$, or equivalently,

$$t^2\mathbb{E}\big((Y_1)^2\big) - 2t\mathbb{E}(Y_1Y_2) + \mathbb{E}\big((Y_2)^2\big) \geqslant 0.$$

(b) Using the inequality in (a), show that $\rho^2 \leqslant 1$.

(a) $\mathbb{E}\big((tY_1 - Y_2)^2\big)$ is a non-negative quadratic in t which is equivalent to its discriminant being non-positive.

$$\therefore 0 \geqslant b^2 - 4ac = (2\mathbb{E}(Y_1Y_2))^2 - 4\mathbb{E}\big((Y_1)^2\big)\mathbb{E}\big((Y_2)^2\big)$$

from which the result follows.

(b) By definition,

$$\rho^2 = \frac{(\operatorname{Cov}(Y_1, Y_2))^2}{(\sigma_1)^2(\sigma_2)^2} = \frac{(\mathbb{E}((Y_1 - \mu_1)(Y_2 - \mu_2)))^2}{\operatorname{Var}(Y_1)\operatorname{Var}(Y_2)} \overset{(a)}{\leqslant} \frac{\mathbb{E}\big((Y_1 - \mu_1)^2\big)\mathbb{E}\big((Y_2 - \mu_2)^2\big)}{\operatorname{Var}(Y_1)\operatorname{Var}(Y_2)} = 1$$

The intermediate step takes for granted that $Z_i = Y_i - \mu_i$ satisfy the requirements of (a) i.e. that (Z_1, Z_2) admits a joint density and that each Z_i has finite variance (which is true because variance is translationally invariant).

■

- A value of $\rho = 0$ indicates the lack of linear relationship between the two variables (absence of correlation) but not necessarily independence.
- The limiting values indicate perfect negative ($\rho = -1$) or positive ($\rho = +1$) correlation — the sample values fall on a straight line.

Definition 7.3.2 [4]

- A **correlation coefficient** in general is a measurement (that's unchanged by both addition and multiplication of the random variable(s) by constants) which describes the tendency of two random variables X and Y to vary together.
- An **estimator of ρ** obtained from n samples values $(x_1, y_1), \ldots, (x_n, y_n)$ of the two random variables of interest is **Pearson's product moment correlation coefficient**, denote by r and given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}}.$$

- More generally, if a score is allotted to each pair of individuals, say a_{ij} for the x-group and b_{ij} for the y-group, a **generalised coefficient** of correlation may be defined as

$$\Gamma = \frac{\sum a_{ij}b_{ij}}{\sqrt{\left(\sum a_{ij}^2\right)\left(\sum b_{ij}^2\right)}}$$

  where $\sum$ is a summation over all values of i, j $(i \neq j)$ from 1 to n. This general coefficient includes:

  - Kendall's τ
  - Spearman's ρ
  - Pearson's PMCC r i.e. $a_{ij} = x_i - x_j$ and $b_{ij} = y_i - y_j$

7.4 Calculating Expectations and Variances of Linear Combinations

Let $Y_1, \ldots, Y_n$ and $X_1, \ldots, X_m$ be random variables with $\mathbb{E}(Y_i) = \mu_i$ and $\mathbb{E}(X_j) = \xi_j$. Define

$$U_1 = \sum_{i=1}^{n} a_i Y_i \quad \text{and} \quad U_2 = \sum_{j=1}^{m} b_j X_j$$

for constants $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$. Then

(a) $\mathbb{E}(U_1) = \sum_{i=1}^{n} a_i \mu_i$

(b) $\mathrm{Var}(U_1) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(Y_i) + 2 \sum_{1 \leqslant i < j \leqslant n} a_i a_j \mathrm{Cov}(Y_i, Y_j)$

(c) $\mathrm{Cov}(U_1, U_2) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \mathrm{Cov}(Y_i, X_j)$

Proof.

(a) Follows from linearity.

(b) Expand out the expression for $\mathrm{Var}(U_1)$

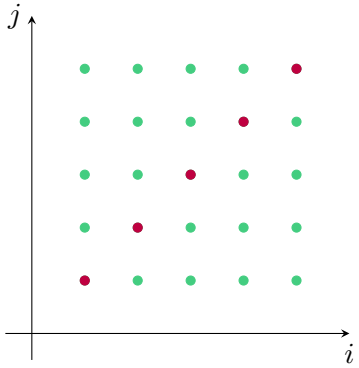
$$\begin{aligned}\mathrm{Var}(U_1) = \mathbb{E}\big((U_1 - \mathbb{E}(U_1))^2\big) &\overset{\text{(a)}}{=} \mathbb{E}\left(\left(\sum_{i=1}^n a_i Y_i - \sum_{i=1}^n a_i \mu_i\right)^2\right) \\ &= \mathbb{E}\left(\left(\sum_{i=1}^n a_i (Y_i - \mu_i)\right)^2\right)\end{aligned}$$

and note the following general expansion

$$\begin{aligned}\left(\sum_{i=1}^n a_i\right)^2 = \sum_{i=1}^n a_i \sum_{j=1}^n a_j = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \\ = \sum_{\substack{i=1 \\ i=j}}^n a_i a_j + \sum_{\substack{i=1 \\ i\neq j}}^n a_i a_j \\ = \left(\sum_{i=1}^n a_i^2\right) + 2\sum_{\substack{i=1 \\ i>j}}^n a_i a_j\end{aligned} \tag{7.1}$$

in order to conclude that

$$\begin{aligned}\mathrm{Var}(U_1) &= \mathbb{E}\left(\sum_{i=1}^n a_i^2 (Y_i - \mu_i)^2 + 2\sum_{i=1}^n \sum_{j>i} a_i a_j (Y_i - \mu_i)(Y_j - \mu_j)\right) \\ &= \sum_{i=1}^n a_i^2 \underbrace{\mathbb{E}\big((Y_i - \mu_i)^2\big)}_{=\mathrm{Var}(Y_i)} + 2\sum_{i=1}^n \sum_{j>i} a_i a_j \underbrace{\mathbb{E}((Y_i - \mu_i)(Y_j - \mu_j))}_{=\mathrm{Cov}(Y_i, Y_j)}.\end{aligned}$$



The diagonal elements correspond to when $i = j$. The off-diagonal entries are symmetric about $i = j$ so their sum is equal to twice the entries strictly above (resp. strictly below) the line $i = j$.

(c)

$$\begin{aligned}\mathrm{Cov}(U_1, U_2) &= \mathbb{E}((U_1 - \mathbb{E}((U_1))(U_2 - \mathbb{E}((U_2))) \\ &= \mathbb{E}\left(\left(\sum_{i=1}^n a_i (Y_i - \mu_i)\right)\left(\sum_{j=1}^m b_j (X_j - \xi_j)\right)\right) \\ &= \mathbb{E}\left(\sum_{i=1}^n \sum_{j=1}^m a_i b_j (Y_i - \mu_i)(X_j - \xi_j)\right) \\ &= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \underbrace{\mathbb{E}((Y_i - \mu_i)(X_j - \xi_j))}_{=\mathrm{Cov}(Y_i, X_j)}.\end{aligned}$$

■

CHAPTER 8

Discrete Probability Distributions

This chapter will go over some examples of different experiments, their associated discrete random variables of interest, and their respective probability distributions. Unless stated otherwise, the σ-algebra associated with any finite set is simply its power set.

8.1 Uniform

The simplest available distribution to mankind. A **uniform experiment** is formalised by a probability space $(\Omega = \{\omega_1, \dots, \omega_n\}, \mathcal{F} = 2^\Omega, \mathbb{P})$ in which all outcomes have equal selection probability i.e. $\mathbb{P}(\{\omega_i\}) = 1/\operatorname{card}(\Omega)$. Such a probability measure is called **uniform** on Ω.

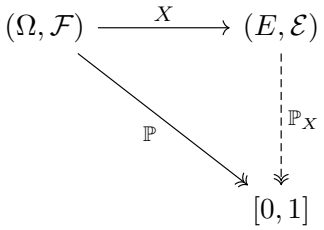
Once we carry out a uniform experiment, we observe some element $e_i = X(\omega_i)$ of a population $(E, \mathcal{E})$ that corresponds to the theoretical outcome $\omega_i \in \Omega$.

The associated random variable $X\colon \Omega \to E$ maps into the population E (e.g. students in a class $E = \{\text{Marshall}, \text{Chad}, \dots, \text{Shamrock}\}$). Note that in this case X is a bijection i.e. for every $i = 1, \dots, n$: $X(\omega_i) = e_i$.

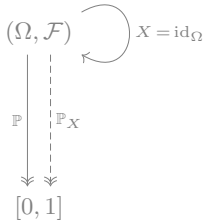
The probability distribution of X is determined entirely by its behaviour on each element $e_i \in E$:

$$\mathbb{P}_X(\{e_i\}) = \mathbb{P}\big(X^{-1}(\{e_i\})\big) = \mathbb{P}(\{\omega_i\}) = \frac{1}{n}$$

i.e. the mass function at each singleton is simply the reciprocal of the cardinality of the outcome space. Thus, we say that X has a **uniform distribution** if its law $\mathbb{P}_X$ is uniform on E.



Informally, we could also identify Ω and E via $X = \mathrm{id}$. In this case, our setup is as follows:



I'm not a big fan of this because it conflates the observable $X(\omega)$ with the theoretical underlying random outcome ω to which the observable corresponds.

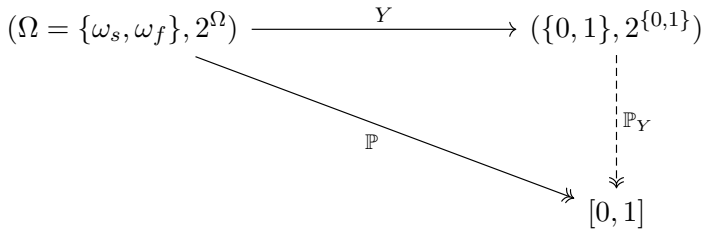
8.2 Bernoulli

A **Bernoulli experiment** is a random experiment with only two outcomes $\Omega = \{\omega_s, \omega_f\}$, success ω_s with a fixed probability $p \in [0, 1]$, or failure ω_f with probability $q = 1 - p$. The associated

random variable $Y : \Omega \to E = \{0, 1\}$ denotes the values we observe by:

$$Y(\omega) = \begin{cases} 1, & \text{if } \omega = \omega_s \\ 0, & \text{if } \omega = \omega_f. \end{cases}$$

The associated diagram is as follows:



Y is said to have a **Bernoulli p distribution**, denoted $Y \sim \text{Bern}(p)$, if its probability distribution is defined by:

$$\begin{aligned} \mathbb{P}_Y(\{y\}) = \mathbb{P}\big(Y^{-1}(\{y\})\big) &= \begin{cases} \mathbb{P}(\{\omega_s\}) & \text{if } y = 1 \\ \mathbb{P}(\{\omega_f\}) & \text{if } y = 0 \end{cases} \\ &= \begin{cases} p & \text{if } y = 1 \\ q = 1 - p & \text{if } y = 0. \end{cases} \end{aligned}$$

Uses:

- Coin flip — success/failure experiments.

The expected value of $Y \sim \text{Bern}(p)$ is given by the following sum over $y \in \text{supp}(\mathbb{P}_Y)$:

$$\mathbb{E}(Y) = \sum_y y \cdot \mathbb{P}_Y(\{y\}) = 1 \cdot \mathbb{P}(\{Y = 1\}) + 0 \cdot \mathbb{P}(\{Y = 0\}) = p.$$

8.3 An Important Point!

This is a good place to mention an important point: The Kolmogorov formulation of probability defines an abstract probability space to model the randomness of an experiment. A random variable X formalises a quantity that we may observe depending on the outcomes of said experiment, and its distribution is the pushforward of $\mathbb{P}$ via X — this distribution models the observables.

Consider the following examples:

- The experiment of flipping a coin is just the Bernoulli experiment above with success parameter $1/2$, $\omega_s = H$, and $\omega_f = T$. As before, define $X(\{H\}) = 1$, $X(\{T\}) = 0$.

  Thus, $X \sim \text{Bern}(1/2)$.

- Consider an experiment formalised by $(\tilde{\Omega} = \{\omega_1, \omega_2, \gamma_1, \gamma_2\}$, with probability measure $\tilde{\mathbb{P}}$ defined by $\tilde{\mathbb{P}}(\{\omega_i\}) = 1/4 = \tilde{\mathbb{P}}(\{\gamma_i\})$. Let Y be defined by $Y(\omega_i) = 1$, $Y(\gamma_i) = 0$. The law of Y is given by:

  - $\tilde{\mathbb{P}}_Y(\{1\}) = \tilde{\mathbb{P}}(Y^{-1}(\{1\})) = \tilde{\mathbb{P}}(\{\omega_1, \omega_2\}) = \tilde{\mathbb{P}}(\{\omega_1\}) + \tilde{\mathbb{P}}(\{\omega_2\}) = 1/2$
  - $\tilde{\mathbb{P}}_Y(\{0\}) = \tilde{\mathbb{P}}(Y^{-1}(\{0\})) = \tilde{\mathbb{P}}(\{\gamma_1, \gamma_2\}) = \tilde{\mathbb{P}}(\{\gamma_1\}) + \tilde{\mathbb{P}}(\{\gamma_2\}) = 1/2$

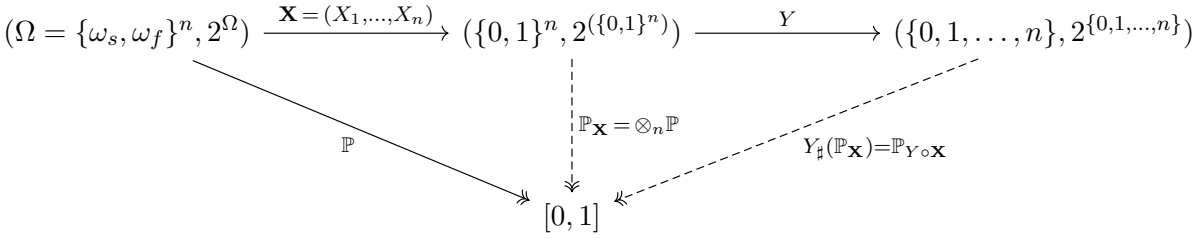
  Thus, $Y \sim \text{Bern}(1/2)$

This demonstrates that we have two different underlying probability spaces and random variables giving rise to the same law. From the perspective of the values we observe, which is the case in mathematical statistics, the underlying randomness is immaterial.

When context dictates that only the observables in $(E, \mathcal{E})$ matter, we may simply refer to $\mathbb{P}_X$ (defined on $\mathcal{E}$) which makes no explicit reference to $(\Omega, \mathcal{F}, \mathbb{P})$. This is reflected in the notation, for example, that $Y \sim \text{Bern}(p)$.

8.4 Binomial

A **binomial experiment** consists of the observation of a sequence of n independent[1] and identically distributed Bernoulli p random variables $X_i \sim \text{Bern}(p)$ for $i = 1, \ldots, n$. The random variable of interest Y is the number of "successes" in n trials. The overall picture is as follows:



- On the left, we perform the same Bernoulli experiment n times and independently which is represented by $\mathbf{X}$ mapping into the space of tuples of length n with entries as either 0 or 1. The resulting distribution of these samples/tuples is given by the product measure of the original distribution $\mathbb{P}$ of the underlying Bernoulli experiment space.
- On the right, we define the random variable Y as the sum of the entries in such a tuple $(X_1, \ldots, X_n)$.

The probability distribution of Y is given by

$$\begin{aligned}(Y_\sharp \mathbb{P}_{\mathbf{X}})(\{y\}) = \mathbb{P}_{\mathbf{X}}\big(Y^{-1}(\{y\})\big) &= (\otimes_n \mathbb{P})\left(Y^{-1}(\{y\})\right) \\ &= (\otimes_n \mathbb{P})\left(\left\{(x_1, \ldots, x_n) \in \{0,1\}^n : \sum_{i=1}^n x_i = y\right\}\right)\end{aligned}$$

Using the sample point method, we can identify the event $\{\omega \in \Omega \colon Y(\omega) = y\}$ of y successes in n trials as a disjoint union of elementary events. Each elementary event can be written as a string of y successes and $n - y$ failures. Since each Bernoulli trial is independent, the probability of this sample point is $p^y(1-p)^{n-y}$. The selection of y successes from a total of n trials is equivalent to partitioning the n objects into 2 groups, the y selected and $n - y$ remaining. The number of ways to do so is

$$\binom{n}{y, n-y} = \frac{n!}{y!(n-y)!} = \binom{n}{y}$$

We combine all of the above information and complete our calculation of the probability of the event in which Y assumes the value y:

[1] It's important to note that the assumption of each trial being independent and identically distributed is conceptually equivalent to a scenario in which one samples n elements from a population of N-large (relative to n) or infinitely many elements ($N = \infty$) without replacement.

$$
\begin{aligned}
(Y_\sharp \mathbb{P}_{\mathbf{X}})(\{y\}) = \ldots &= (\otimes_n \mathbb{P})\left(\left\{(x_1,\ldots,x_n) \in \{0,1\}^n : \sum_{i=1}^n x_i = y\right\}\right) \\
&= \sum_{(x_1,\ldots,x_n)\in Y^{-1}(\{y\})} \left(\prod_{i=1}^n \mathbb{P}_{X_i}(\{x_i\})\right) \\
&= \sum_{(x_1,\ldots,x_n)\in Y^{-1}(\{y\})} p^y(1-p)^{n-y} \\
&= \binom{n}{y} p^y (1-p)^{n-y}
\end{aligned}
$$

This is the probability mass function of a **binomial distribution** with parameters n and p.

We say that Y is **binomially distributed with parameters n and p** (denoted by $Y \sim \text{Binom}(n,p)$) if its law $\mathbb{P}_Y$ is as above.

Remarks (Sampling) In the case that the sample size n is relatively small when compared to the population size N from which we are sampling the $X_1,\ldots,X_n$, then the conditional probability of success on a later trial given the number of successes on previous trials will remain approximately constant. This suggests that the Bernoulli trials are approximately independent. Sampling problems of this type are approximately binomial.

In the case that n is large relative to N and we're sampling without replacement, the conditional probability of "success" on a later trial will be affected by previous draws. Thus, the trials are no longer independent. A more appropriate probability model to use is called the hypergeometric distribution.

Uses:

- To model counts.

Let $Y \sim \text{Binom}(n,p)$. Then

$$
\begin{aligned}
\mathbb{E}(Y) &= \sum_{k=0}^n k\mathbb{P}(\{Y=k\}) \\
&= \sum_{k=0}^n k\binom{n}{k} p^k (1-p)^{n-k} \\
&= np\sum_{k=1}^n \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \\
&= np\sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\
&= np
\end{aligned}
$$

8.5 (Discrete) Geometric

A geometric experiment exhibits some similarities to the binomial experiment. In particular, a **geometric experiment** entails observing/drawing an infinite sequence of identical and independent Bernoulli trials (modelled as iid $X_i \sim \text{Bern}(p)$ for $i = 1,2,\ldots$).

The random variable Y of interest is how many trials it takes to draw the first success, **including the first success**. Denote the shared outcome space of the X_i by Ω. Then the outcome space of Y is the Cartesian product $\Omega^\infty = \{(\omega_1,\omega_2,\ldots)\}$. We therefore define $Y : \Omega^\infty \to \mathbb{N}$ for a particular $\omega \in \Omega^\infty$ by[2]

$$Y(\omega) := \min\{i \geqslant 1 : X_i(\omega_i) = 1\},$$

[2] I'm not mentioning the σ-algebra over Ω^∞, the σ-algebra associated with $\mathbb{N}$ (though it's probably just $2^{\mathbb{N}}$), and therefore measurability of Y.

where ω_i is the outcome of the i^{th} trial.
The probability distribution of Y, for $y \in \mathbb{N}$, is given by

$$\begin{aligned}
\mathbb{P}_Y(\{y\}) &= \mathbb{P}\big(Y^{-1}(\{y\})\big) \\
&= \mathbb{P}(\{\omega \in \Omega^\infty \colon X_1(\omega_1) = 0, \ldots, X_{y-1}(\omega_{y-1}) = 0, X_y(\omega_y) = 1, X_{y+1}(\omega_{y+1}) \in \{0,1\}, \ldots\}) \\
&= \left(\prod_{i=1}^{y-1} \mathbb{P}(\{X_i = 0\})\right) \mathbb{P}(\{X_y = 1\}) \left(\prod_{i=y+1}^{\infty} \mathbb{P}(\{X_i \in \{0,1\}\})\right) \quad \text{by independence} \\
&= (1-p)^{y-1}p.
\end{aligned}$$

This is called the **geometric distribution**.

8.5.1 Memorylessness

Exercise 2 (3.55 [6]) Let Y denote a discrete random variable that has a geometric distribution with probability of success p.

1. Show that for a positive integer a,
$$\mathbb{P}(\{Y > a\}) = q^a$$
2. Show that for positive integers a and b,
$$\mathbb{P}(\{Y > a+b\} \,|\, \{Y > a\}) = q^b = \mathbb{P}(\{Y > b\}).$$
This result implies that, for example, $\mathbb{P}(\{Y > 7\} \,|\, \{Y > 2\}) = \mathbb{P}(\{Y > 5\})$. Why do you think this property is called the **memoryless property** of a discrete geometric distribution?

Solution:

1. For any positive integer a:
$$\begin{aligned}
\mathbb{P}(\{Y > a\}) &= 1 - \mathbb{P}(\{Y \leqslant a\}) \\
&= 1 - \sum_{y=1}^{a} \mathbb{P}(\{Y = y\}) \quad \text{since the events } \{Y = y\} \text{ are mutually exclusive} \\
&= 1 - \sum_{y=1}^{a} (1-p)^{y-1}p \\
&= 1 - p\sum_{y=0}^{a-1} (1-p)^{y} \\
&= 1 - p\left(\frac{1((1-p)^{(a-1)+1} - 1)}{(1-p) - 1}\right) \\
&= 1 - (-1)((1-p)^a - 1) \\
&= (1-a)^p \\
&= q^p \quad (\text{where } q = 1-p)
\end{aligned}$$
2. By the definition of conditional probability:
$$\mathbb{P}(A \,|\, B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Let $A = \{Y > a + b\}$ and $B = \{Y > a\}$. Then $A \cap B = A$ and

$$\begin{aligned}\mathbb{P}(\{Y > a+b\} \mid \{Y > a\}) = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} &= \frac{\mathbb{P}(\{Y > a+b\})}{\mathbb{P}(\{Y > a\})} \\ &= \frac{q^{a+b}}{q^a} \\ &= q^b \\ &= \mathbb{P}(\{Y > b\})\end{aligned}$$

This is called the memoryless property because we can, in terms of probabilities, forget all activity prior to the "current time" a and the subsequent probabilities would be the same had the process begun at time 0.

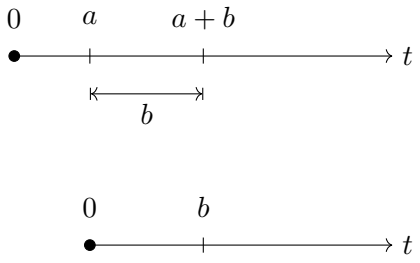


Figure 8.1: The time a already spent waiting for an event to occur i.e. $\{Y > a\}$ does not affect how much longer the wait will be. Our new origin when conditioning is a — we re-label it to 0 in the second visualisation.

Uses:

- Modelling the distribution of the waiting time until an event occurs.
  (As we shall see later, this is the discrete-time analogue of an exponential distribution.)

8.6 Negative Binomial

A negative binomial experiment can be considered a generalisation of the geometric experiment where we are instead interested in the random variable Y representing the number of trials it takes to achieve the r^{th} success. If we denote the number of cumulative successes after y trials as

$$S_y = \sum_{i=1}^{y} X_i,$$

then $Y = \min\{y \in \mathbb{N} \colon S_y = r\}$.

- Let event A denote exactly $r-1$ successes in the first $y-1$ trials

$$A = \{\omega \in \Omega^\infty \colon S_y(\omega) = r - 1\}$$

- Let B denote the event that the y^{th} trial is a success

$$B = \{\omega \in \Omega^\infty \colon X_y(\omega) = r - 1\}$$

There are $\binom{y-1}{r-1}$ ways to arrange $r-1$ successes in $y-1$ trials. Each sample point of $r-1$ successes and $(y-1)-(r-1)$ failures has probability $p^{r-1}(1-p)^{y-1-(r-1)}$. The y^{th} trial is a success so B

has probability p. The trials are independent so A and B are independent. Therefore,

$$\begin{aligned}\mathbb{P}_Y(\{y\}) = \mathbb{P}\big(Y^{-1}(\{y\})\big) &= \mathbb{P}(\{\omega \in \Omega^\infty \colon Y(\omega) = y\}) \\ &= \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) \cdot \mathbb{P}(B) = \binom{y-1}{r-1} p^{r-1}(1-p)^{y-r} \cdot p.\end{aligned}$$

This is the probability mass function of the **negative binomial distribution**.

There are two forms of the geometric and negative binomial experiments:

- Failures until the first success
- The number of trials including the first success

Uses:

- Modelling times to r^{th} success in a sequence of independent Bernoulli trials.
  (As we shall see later, this is the discrete-time analogue of a gamma distribution.)

8.7 Hypergeometric

Suppose that a population contains a finite number N of elements. In a hypergeometric experiment, we suppose that we're randomly sampling n elements without replacement from a finite population of N elements, with each element possessing one of two characteristics: red and black.

- r elements are red
- $b = N - r$ elements are black

We think of the sampling process as generating outcomes tied to indicator random variables for each draw. Each draw can be represented by a random variable X_i for $i = 1, \ldots, n$. However, these variables are dependent and not identically distributed. This is because sampling without replacement means that the conditional probability of "success" (drawing a red) on a later trial will be affected by previous draws. The random variable of interest Y is the number of red elements in the sample.

An appropriate sample space Ω consists of all possible subsets[3] of size n drawn from the population of N elements. There are $\binom{N}{n}$ elements in Ω.

For each outcome, let Y be the number of red elements drawn. Therefore, the probability of y red elements in the sample of n elements, $\mathbb{P}_Y(\{y\})$, is given by the number of subsets with exactly r elements divided by $\binom{N}{n}$. The total number of sample points containing y red elements is equal to the number of ways of choosing y red elements from r red elements multiplied by the number of ways we can select $n - y$ black elements from the $N - r$ black elements

$$\mathbb{P}_Y(\{y\}) = \mathbb{P}\big(Y^{-1}(\{y\}\big) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}.$$

This is the mass function of the **hypergeometric distribution**.

[3]This is in contrast to the binomial experiment which tracked the order of outcomes as a sequence of random variables. The order of the draws for the hypergeometric experiment does not matter.

8.8 Hypergeometric Approximates Binomial

Recall the hypergeometric experiment and how n is large relative to N. Under certain circumstances, one would expect the probabilities assigned by a hypergeometric distribution to approach those assigned by a binomial distribution as N grows large and n remains fixed. Stated more precisely:

Lemma 8.8.1 Let Y be a hypergeometrically distributed random variable with probability mass function $p_Y(y)$. For fixed n and y, as $N \to +\infty$ and $r = r(N)$ is such that $\frac{r}{N}$ is held constant at some value p, it follows that

$$\lim_{N\to\infty} p_Y(y) := \lim_{N\to\infty} \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \binom{n}{y} p^y (1-p)^{n-y}.$$

Proof. Omitted from the Wackerly textbook so I decided to derive it on my own.

$$\begin{aligned}
p_Y(y) &= \frac{r!}{y!(r-y)!}\frac{(N-r)!}{(n-y)!((N-r)-(n-y))!}\frac{n!(N-n)!}{N!} \\
&= \binom{n}{y}\frac{r!}{(r-y)!}\frac{(N-r)!}{((N-r)-(n-y))!}\frac{(N-n)!}{N!} \\
&\overset{(*)}{=} \binom{n}{y}\frac{r(r-1)\cdot\ldots\cdot(r-y+1)\cdot(N-r)(N-r-1)\cdot\ldots\cdot(N-r-(n-y)+1)}{N(N-1)\cdot\ldots\cdot(N-n+1)} \\
&\overset{(**)}{=} \binom{n}{y}\frac{r(r-1)\cdot\ldots\cdot(r-y+1)}{N(N-1)\cdot\ldots\cdot(N-y+1)}\cdot\frac{(N-r)(N-r-1)\cdot\ldots\cdot(N-r-(n-y)+1)}{(N-y)(N-y-1)\cdot\ldots\cdot(N-n+1)} \\
&= \binom{n}{y}\cdot\left(\prod_{a=0}^{y-1}\frac{r-a}{N-a}\right)\cdot\left(\prod_{b=0}^{n-y-1}\frac{N-r-b}{N-y-b}\right) \\
&\overset{n\to\infty}{\longrightarrow} \binom{n}{y}\left(\frac{r}{N}\right)^y\left(\frac{N-r}{N}\right)^{n-y} \\
&= \binom{n}{y}\left(\frac{r}{N}\right)^y\left(1-\frac{r}{N}\right)^{n-y} \\
&=: \binom{n}{y}p^y(1-p)^{n-y}
\end{aligned}$$

$(*)$: The denominator was broken up into two products of "length" y and $(n-y)$, respectively.

$(**)$: This final term is equal to $(N-y+y-n+1)$ which continues the pattern in the denominator.

■

CHAPTER 9

Point Processes (Random Scatters)

I wrote this section after **Chapter 22**, so I already had some familiarity with the *type* of objects seen by that point (kernels). There's little reason why the following exposition wouldn't make sense when read now, but I thought it was worth adding when I studied it.

This chapter is largely based on the **video lectures on Poisson point processes** by Prof. Nicolas Lanchier which follow chapter 9 from his book [2]. I've tried to fill in some details on the types of objects being considered.

9.1 Point Processes

Consider an experiment in which we observe a random scattering of at most countably many points in some state space $(S, \mathcal{S})$. The state space S is typically $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Upon realisation of the random scatter, each point is called[1] a "hit". Consider some bounded Borel subset $B \in \mathcal{B}(\mathbb{R}^d)$. The number of hits in B is a random variable we denote by $N(B)$. If one knows exactly the number of hits in each B i.e. if one knows the collection of random variables $\{N(B)\}_{B \in \mathcal{B}_{\mathbb{R}^d}}$, then one has a full description of the scatter.

> The perspective that Lanchier takes is to give this collection a name $\mathcal{N}$, and demand certain natural properties of its elements. Let $A, B \in \mathcal{B}_{\mathbb{R}^d}$. Then:
>
> - $N(\emptyset) = 0$
> - For non-overlapping A and B, $N(A \sqcup B) = N(A) + N(B)$
> - If $A \subseteq B$, then $N(A) \leqslant N(B)$.
> - Set-difference, inclusion-exclusion etc.
>
> All of these properties follow from one property; that N is σ-additive i.e.
>
> $$N\Big(\bigsqcup_{j \in \mathbb{N}} B_j\Big) = \sum_{j \in \mathbb{N}} N(B_j).$$
>
> These are the same properties one would expect for a positive measure, so we may think of N as more or less the same as a positive measure but taking values in $\mathbb{N}_0$.

I enjoy the perspective that such experiments can be modelled by point processes. We first begin with the definition of a random measure — a measure-valued random element:

Definition 9.1.1 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A **random measure** is a random variable Φ from Ω to the space[2] of counting measures $\mathbb{M}$ on a state space $(S, \mathcal{S})$. This means that each realisation of a point process is a counting measure $\Phi(\omega) \colon \mathcal{S} \to \overline{\mathbb{N}_0}$.

Let $(S = \mathbb{R}^d, \mathcal{S} = \mathcal{B}_{\mathbb{R}^d})$. In a sense, we can view Φ as a function N of two variables

$$N \colon \Omega \times \mathcal{B}_{\mathbb{R}^d} \to \overline{\mathbb{N}_0},$$

and fixing each variable gives us a map in its own right.

- Fixing $\omega \in \Omega$, we can call our map $N(\omega) \colon \mathcal{B}_{\mathbb{R}^d} \to \overline{\mathbb{N}_0}$. The act of fixing ω tells us that a particular random scatter has been realised. Then for each B, $N(\omega)(B)$ counts how many points in this realisation are in B.

  This map $N(\omega, \cdot) = \Phi(\omega)(\cdot)$ is a counting measure.

[1] In the particular case of a Poisson point process, that we shall soon describe, we call these hits "Poisson points."
[2] What's the associated σ-algebra?

- Fixing $B \in \mathcal{B}_{\mathbb{R}^d}$, we can call our map $N(B)(\omega)\colon \Omega \to \overline{\mathbb{N}_0}$. Thus, we have a fixed region B, and for each ω, $N(B)(\omega)$ is the random count of how many points fall in the set B.

  This map $N(\cdot, B) =: N(B)(\cdot)$ is a random variable.

The formal statement of the above discussion is that a random measure may also be defined via a(n a.e. locally finite[3]) transition kernel.

Definition 9.1.2 A transition kernel N from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$ is a map $N\colon \Omega \times \mathcal{B}_{\mathbb{R}^d} \to \mathbb{R}$ s.t.

- for each $\omega \in \Omega$, $N(\omega, \cdot)$ is a measure on $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$.
- for each $B \in \mathcal{B}_{\mathbb{R}^d}$, $N(\cdot, B)$ is $(\mathcal{F}, \mathcal{B}_{\mathbb{R}^d})$-measurable.

This transition kernel defines a random measure because we may think of N as mapping each $\omega \in \Omega$ to $N(\omega, \cdot) \in \mathbb{M}$.

Definition 9.1.3 A **point process** is an integer-valued transition kernel N from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(S, \mathcal{S})$. In our case, $S = \mathbb{R}^d$ and so our point processes are maps $N\colon \Omega \times \mathcal{B}_{\mathbb{R}^d} \to \overline{\mathbb{N}_0}$.

The transition kernel perspective of modelling a random scattering experiment by a point process lends itself well to Lanchier's interpretation because the kernel gives us a collection $\mathcal{N}$ of random variables indexed by the bounded (Borel) subregions of the state space $S = \mathbb{R}^n$.

9.2 Poisson Point Processes

A Poisson point process is a point process N that models a random scattering experiment satisfying a few natural assumptions:

Definition 9.2.1 A point process N, represented by $\mathcal{N} = \{N(B)\colon \mathcal{B}_{\mathbb{R}^d} \ni B\text{-bounded}\}$, is called a **Poisson point process with intensity** μ if it satisfies the following four criteria:

(1) For any pairwise disjoint collection of bounded subregions $\{B_i\}_1^n$, the number of hits in these regions are mutually $\mathbb{P}$-independent i.e.

$$i \neq j,\ B_i \cap B_j = \emptyset \implies N(B_1), N(B_2), \ldots, N(B_n) \text{ are independent.}$$

(2) Homogeneity — In distribution, the number of points in a set depends only on its size.

The distribution of $N(B)$ depends on B only through its Lebesgue measure $\lambda(B)$.

(3) If you take a Borel set with very small Lebesgue measure, the probability of seeing 1 point in this set is of the order $\mu \cdot \lambda(B)$ (small but scales with $\mu \cdot \lambda(B)$).

$$\frac{\mathbb{P}(\{N(B) = 1\})}{\lambda(B)} \to \mu \quad \text{i.e.} \quad \mathbb{P}(\{N(B) = 1\}) = \mu \cdot \lambda(B) + o(\lambda(B)) \text{ as } \lambda(B) \to 0.$$

We view μ as a measure of the density of Poisson points. This is also called an intensity parameter or a rate per unit volume.

(4) There is no aggregation of points in a small Borel subset of $\mathbb{R}^d$ i.e. the probability of observing more than 1 hit in said small set is negligible compared to the size of the set.

$$\frac{\mathbb{P}(\{N(B) > 1\})}{\lambda(B)} \to 0 \quad \text{i.e.} \quad \mathbb{P}(\{N(B) > 1\}) = o(\lambda(B)) \text{ as } \lambda(B) \to 0.$$

We denote this by $\mathcal{N} = \mathcal{P}(\mu)$.

[3]The locally finite condition means that for $\mathbb{P}$-a.e. $\omega \in \Omega$, the measure $N(\omega, \cdot)$ are finite for every bounded $B \in \mathcal{B}_{\mathbb{R}^d}$.

Theorem 9.2.2 The process $\mathcal{N} = \mathcal{P}(\mu)$ iff

(a) For any pairwise-disjoint collection $\{B_i\}_1^n \subseteq \mathcal{B}_{\mathbb{R}^d}$, $N(B_1), \ldots, N(B_n)$ are independent.

(b) For any bounded $B \in \mathcal{B}(\mathbb{R}^d)$, $N(B) \sim \text{Poisson}(\mu \cdot \lambda(B))$.

The Poisson distribution will be defined in the proof.

Proof. **Definition 9.2.1 (1)** is equivalent to **Theorem 9.2.2 (a)**.

For the forward implication, we now prove **Theorem 9.2.2 (b)**. Fix a set $B \in \mathcal{B}(\mathbb{R}^d)$. The idea is to construct a partition of B into cells of equal Lebesgue measure — first into 2, then into 4, ..., then into 2^n. For a single division into 2 parts, by the intermediate value theorem, we can take a hyperplane of constant x_p for some p between 1 and d (inclusive) and move it along B until we attain some unique position where the two subsets are of equal Lebesgue measure. We note, importantly, that each subset is a Borel set in its own right as the intersection of B and a half-space (which is a Borel set).

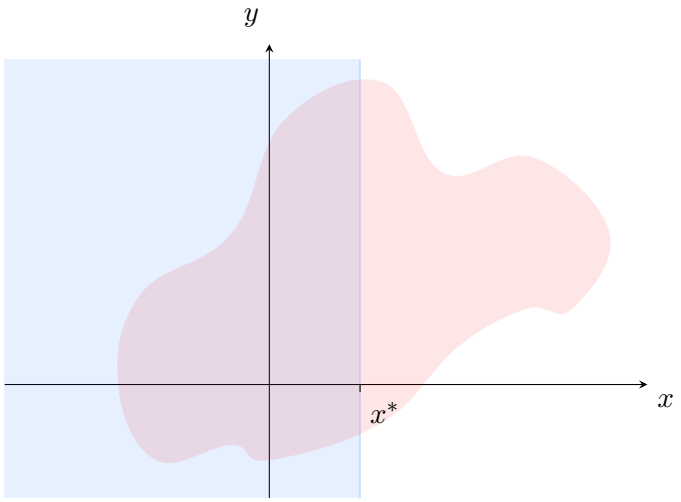


Figure 9.1: A lower dimensional $B \in \mathcal{B}(\mathbb{R}^2)$ for ease of visualisation. The intersection of B and $[-\infty, x^*] \times \mathbb{R}$ form the cell B_1 — its relative complement in B forms the other cell B_2.

Proceeding inductively, we can repeat this for each cell, giving us a partition $\{B_{n,1}, B_{n,2}, \ldots, B_{n,2^n}\}$ of B for every $n \in \mathbb{N}$ such that $\lambda(B_{n,i}) = (\lambda(B))/2^n$.

A skeleton for the rest of the proof is as follows:

- Since the $B_{n,i}$ are disjoint, the $N(B_{n,i})$ are independent from assumption (1).
- Since all the $B_{n,i}$ have the same Lebesgue measure, assumption (2) tells us that the $N(B_{n,i})$ are identically distributed.
- For n large enough, since the $\lambda(B_{n,i})$ are small, assumptions (3) and (4) tell that that the $N(B_{n,i})$ can only assume the values 0 (most of the time) and 1 (sometimes).
- In summary, the $N(B_{n,i})$ are independent and identically distributed Bernoulli random variables with specific success probability based on the intensity.
- Therefore, $N(B)$ is a sum of i.i.d. Bernoulli random variables so it has a binomial distribution. In the limit, we'll discover the distribution of $N(B)$.

For n large enough, there is at most one Poisson point in each cell $B_{n,i}$ by **Definition 9.2.1 (3)** and **Definition 9.2.1 (4)**. Now let

$$\Omega_n = \{N(B_{n,i}) > 1 \text{ for some } i = 1, \dots, 2^n\} = \bigcup_{i=1}^{2^n} \{N(B_{n,i}) > 1\}.$$

The probability of this event is given by

$$\begin{aligned}
\mathbb{P}(\Omega_n) = \mathbb{P}\left(\bigcup_{i=1}^{2^n} \{N(B_{n,i}) > 1\}\right) &\leqslant \sum_{i=1}^{2^n} \mathbb{P}(\{N(B_{n,i}) > 1\}) && \text{by Boole's inequality} \\
&= 2^n \mathbb{P}(\{N(B_{n,1}) > 1\}) && \text{by identical distribution} \\
&= 2^n \cdot o(\lambda(B_{n,1})) && \text{by } \textbf{Definition 9.2.1 (3)} \\
&= 2^n \cdot o(\lambda(B)/2^n) \\
&\overset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

The final limit is because that final quantity is negligible compared to $\lambda(B)$-fixed.

Note that for each fixed n, Ω_n and Ω_n^{c} are subsets of positive probability. By the law of total probability, it follows that the probability of there being k Poisson points in the set B is

$$\begin{aligned}
\mathbb{P}(\{N(B) = k\}) &= \mathbb{P}(\{N(B) = k\} \cap (\Omega_n \sqcup \Omega_n^{\mathsf{c}})) \\
&= \mathbb{P}(\{N(B) = k\} \,|\, \Omega_n)\mathbb{P}(\Omega_n) + \mathbb{P}(\{N(B) = k\} \,|\, \Omega_n^{\mathsf{c}})\mathbb{P}(\Omega_n^{\mathsf{c}}).
\end{aligned}$$

- In the limit as $n \to \infty$, the first term vanishes because $\mathbb{P}(\{N(B) = k\} \,|\, \Omega_n) \in [0, 1]$ and $\mathbb{P}(\Omega_n) \to 0$ as $n \to \infty$.
- For the second term, let $a_n \in [0, 1]$ and $b_n \to 1$. Then

$$|a_n b_n - a_n| = |a_n(b_n - 1)| = |a_n||b_n - 1| \leqslant |b_n - 1| \longrightarrow 0.$$

  Therefore, $a_n b_n \longrightarrow \lim_{n\to\infty} a_n$. Let $a_n = \mathbb{P}(\{N(B) = k\} \,|\, \Omega_n^{\mathsf{c}})$ and $b_n = \mathbb{P}(\Omega_n^{\mathsf{c}})$.

Now we note that k Poisson hits in B is equivalent to k Poisson hits over the partition of B:

$$\begin{aligned}
\mathbb{P}(\{N(B) = k\}) &= \lim_{n\to\infty} \mathbb{P}(\{N(B) = k\} \,|\, \Omega_n^{\mathsf{c}}) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\left\{\sum_{i=1}^{2^n} N(B_{n,i}) = k\right\} \mid \Omega_n^{\mathsf{c}}\right)
\end{aligned}$$

By **Definition 9.2.1 (2)**, each $N(B_{n,i})$ depends only on $B_{n,i}$ through $\lambda(B_{n,i}) = \lambda(B)/2^n$, and the $B_{n,i}$ are pairwise-disjoint, so the $N(B_{n,i})$ are independent and identically distributed. Given the information that Ω_n^{c} provides, the $N(B_{n,i})$ may only assume the values 0 or 1. By **Definition 9.2.1 (3)**, they are i.i.d. Bernoulli random variables with success parameter $\mu \cdot \lambda(B_{n,i})$. It follows that their sum is Binomial with parameter $2^n \mu \cdot \lambda(B_{n,1})$:

$$\begin{aligned}
\mathbb{P}(\{N(B) = k\}) &= \lim_{n\to\infty} \mathbb{P}\left(\text{Binomial}\left(2^n, \mu\tfrac{\lambda(B)}{2^n}\right) \,\middle|\, \Omega_n^{\mathsf{c}}\right)\mathbb{P}(\Omega_n^{\mathsf{c}}) \\
&= \lim_{n\to\infty} \binom{2^n}{k} \left(\mu\frac{\lambda(B)}{2^n}\right)^k \left(1 - \mu\frac{\lambda(B)}{2^n}\right)^{2^n - k} \\
&= \lim_{n\to\infty} \frac{2^n(2^n - 1) \cdot \ldots \cdot (2^n - (k-1))}{k!} \frac{(\mu \cdot \lambda(B))^k}{(2^n)^k} \left(1 - \mu\frac{\lambda(B)}{2^n}\right)^{2^n - k}
\end{aligned}$$

There are k terms in the numerator, all of the order 2^n as n grows large. These cancel out with the $(2^n)^k$ in the denominator. For the term in red, we may write it as the product

$$\left(1 - \mu\frac{\lambda(B)}{2^n}\right)^{2^n} \frac{1}{(\ldots)^k}$$

where the second term converges to 1 as n grows large because k is fixed. The behaviour of the first term is not so clear as the base converges to 1 but the exponent is not fixed and grows without bound to $+\infty$. To deal with this, let's take its natural logarithm to obtain

$$2^n \log\left(1 - \tfrac{\mu\cdot\lambda(B)}{2^n}\right).$$

The natural logarithm term, by looking at the behaviour of the Taylor series, behaves like $-(\mu \cdot \lambda(B))/2^n$ and so the whole term behaves like $-\mu \cdot \lambda(B)$. Thus, by the monotonicity of the natural logarithm, the original term

$$\left(1 - \mu\frac{\lambda(B)}{2^n}\right)^{2^n} \overset{n\to\infty}{\longrightarrow} e^{-\mu\cdot\lambda(B)}.$$

Finally, we put this limit back into the probability calculation:

$$\mathbb{P}(\{N(B) = k\}) = \frac{(\mu \cdot \lambda(B))^k}{k!}e^{-\mu\cdot\lambda(B)} =: \mathbb{P}(\{Y = k\})$$

where $Y \sim \text{Poisson}(\mu \cdot \lambda(B))$ which is defined by:

Definition 9.2.3 A discrete random variable Y that admits a density

$$p_Y(k) = \frac{(\mu\lambda(B))^k}{k!}e^{-\mu\cdot\lambda(B)}$$

is said to be **Poisson-distributed with rate parameter μ**. We denote this by $Y \sim \text{Poisson}(\mu)$.

Therefore, for every bounded $B \in \mathcal{B}(\mathbb{R}^d)$, $N(B) \sim \text{Poisson}(\mu \cdot \lambda(B))$.

For the reverse implication, suppose that **Theorem 9.2.2 (b)** holds true and let $B \in \mathcal{B}_{\mathbb{R}^d}$ be s.t. $\lambda(B)$ is small. Abusing notation slightly by replacing functions $f \in o(\dots)$ by $o(\dots)$, it follows that

$$\begin{aligned}
\mathbb{P}(\{X(B) = 1\}) &= \frac{(\mu \cdot \lambda(B))^1}{1!}e^{-\mu\cdot\lambda(B)} \\
&= (\mu \cdot \lambda(B))e^{-\mu\cdot\lambda(B)} \\
&= (\mu \cdot \lambda(B))(1 - \mu \cdot \lambda(B) + o(\mu \cdot \lambda(B)) \\
&= \mu \cdot \lambda(B) - \mu^2(\lambda(B))^2 + o(\lambda(B)) \\
&= \mu \cdot \lambda(B) + o(\lambda(B)).
\end{aligned}$$

Dividing through by $\lambda(B)$ demonstrates **Definition 9.2.1 (3)**. For the final property:

$$\begin{aligned}
\mathbb{P}(\{X(B) > 1\}) &= 1 - (\mathbb{P}(\{X(B) = 0\}) + \mathbb{P}(\{X(B) = 1\}) \\
&= 1 - (e^{-\mu\cdot\lambda(B)} + \mu \cdot \lambda(B)e^{-\mu\cdot\lambda(B)}) \\
&= 1 - (1 + \mu \cdot \lambda(B))e^{-\mu\cdot\lambda(B)} \\
&= 1 - (1 - \mu^2(\lambda(B))^2) + o(\lambda(B)) \\
&= o(\lambda(B))
\end{aligned}$$

Again, dividing through by $\lambda(B)$ proves **Definition 9.2.1 (4)**. ■

Uses:

- The Poisson distribution provides a good model for the probability distribution of the **count** Y of rare events that occur in a fixed, bounded space, time, volume, or any other dimension where μ is the average value (intensity, or rate per unit time) of Y.

Example 9.2.4 The Poisson distribution can be used to model a random variable representing the number of:

- telephone calls handled by a switchboard in a time interval,
- accidents in a given unit of time, or
- errors made by a typist in typing a page.

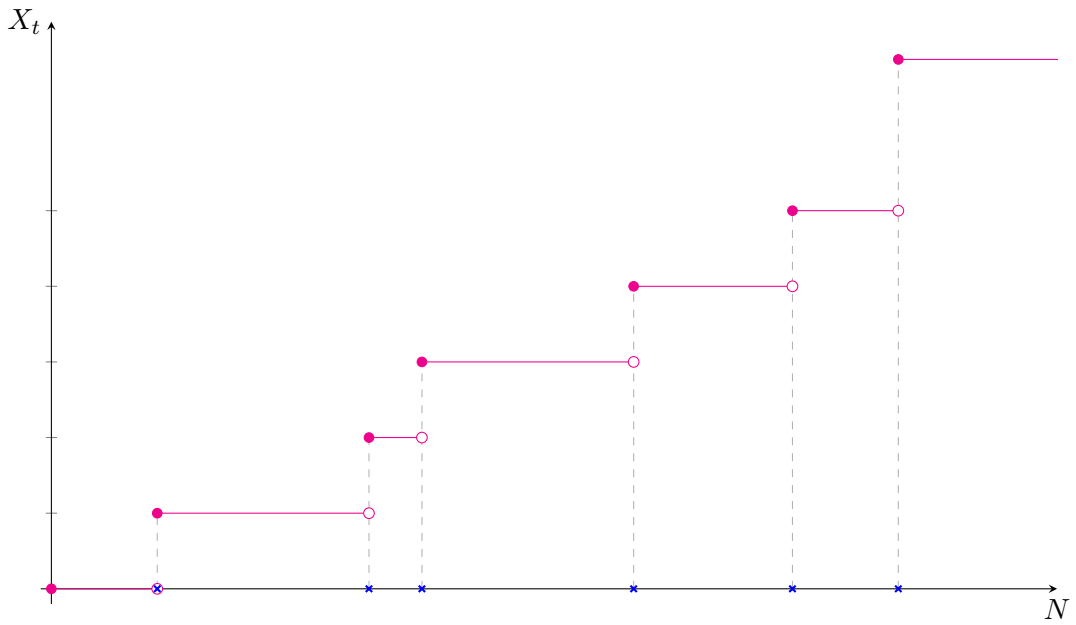
9.3 Poisson Processes

Consider a Poisson point process in 1 dimension. We may interpret the one-dimensional space as an axis of time $[0, +\infty)$. This point process counts the number of hits in any Borel subset of $[0, +\infty)$. Since we're in $[0, +\infty)$, we may order these Poisson points and reinterpret them as the times of occurrence of the process. We denote this collection of random variables by $\{N([0,t])\}_{t\in[0,+\infty)}$. We may plot such a Poisson point process in time with some intensity (or, really, rate per unit time) μ:

Given this Poisson point process, we may define a **continuous-time stochastic process**[4] i.e. a collection of random variables $\{X_t\}$ indexed by $t \in [0, +\infty)$, taking values in $\mathbb{N}_0$. For any $t \in (0, +\infty)$, define

$$X_t := N([0,t]),$$

and for $t = 0$, $X_0 := 0$. This stochastic process, at each time $t \in [0, +\infty)$ counts the number of Poisson points in the interval $[0, t]$. We may visualise the occurrences of events as jumps:



It's visually clear that given a Poisson point process N, we can construct a Poisson process, and we can also recover the Poisson point process from the Poisson process $\{X_t\}$ via the equality $X_t = N([0,t])$. Thus, despite being different mathematical objects, a Poisson process with rate μ is "mathematically equivalent" to a Poisson point process N with intensity μ in one dimension.

Definition 9.3.1 A stochastic process $\{X_t\}$ is a **Poisson process with rate μ** iff

[4] I'll define these rigorously in a later chapter. The definition for now is sufficient for the purposes of motivating the connection between a Poisson point process in one dimension, and a Poisson (stochastic) process in time $[0, +\infty)$.

- Increments are independent i.e. for all $t_1 < t_2 < \ldots < t_{2n}$:

$$X_{t_2} - X_{t_1}, X_{t_4} - X_{t_3}, \ldots, X_{t_{2n}} - X_{t_{2n-1}} \text{ are independent.}$$

- Increments are Poisson i.e. for all $s < t$, $X_t - X_s \sim \text{Poisson}(\mu(t-s))$.

Proof. For the forward implication:

- The random variables $X_{t_{2n}} - X_{t_{2n-1}} = X([0, t_{2n}]) - X([0, t_{2n-1}]) = X([t_{2n-1}, t_{2n}])$. Now note that the intervals $\{[t_{2i-1}, t_{2i}]\}_{i=1}^n$ are pairwise-disjoint, so **Definition 9.2.1 (1)** tells us the random variables $X_{t_{2n}} - X_{t_{2n-1}}$ are independent.
- It follows from **Theorem 9.2.2** that for a Poisson process, since any increment can be written as $X_t - X_s = X([s,t])$, and $[s,t]$ is a bounded Borel subset of $[0, +\infty)$, then we conclude that $X([s,t]) \sim \text{Poisson}(\mu \cdot \underbrace{\lambda([s,t])}_{=t-s})$.

I believe the reverse implication involves some kind of approximation or monotone class argument, building up from intervals $[a,b]$ to any Borel subset B. ■

9.3.1 A Bridge to Absolutely Continuous Distributions

I want to go into more details about Poisson processes but I think some background in conditional expectation/probability, and stochastic processes will be useful. As a placeholder, I will state some known facts that will be useful going forward:

- For a Poisson process, we define the **hitting time** at state i by $\tau_i := \inf\{t \geqslant 0 \colon X_t = i\}$, then the **inter-arrival times** are defined by $T_i := \tau_i - \tau_{i-1}$.
- The τ_i are random variables, and the T_i are independent and identically distributed **exponential random variables**, denoted $T_i \underset{\text{i.i.d.}}{\sim} \mathcal{E}(\mu)$.

Definition 9.3.2 An absolutely continuous[5] real-valued random variable Y is said to have an **exponential distribution with rate parameter $\lambda > 0$**, denoted $Y \sim \mathcal{E}(\mu)$, if its density function is:

$$f_Y(y) = \begin{cases} \mu e^{-\mu \cdot y}, & 0 \leqslant y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Proof. Recall that $X_0 = 0$ and for each $t \in (0, +\infty)$, $X_t := N([0,t]) \sim \text{Poisson}(\mu \cdot t)$. The first jump/hit/arrival occurs after time t if and only if there is no jump by time t. Thus

$$\mathbb{P}(\{T_1 > t\}) = \mathbb{P}(\{X_t = 0\}) = \frac{(\mu t)^0}{0!} e^{-\mu \cdot t} = e^{-\mu \cdot t}.$$

[5] With respect to the Lebesgue measure on $\mathbb{R}$.

Now note that the CDF of $Z \sim \mathcal{E}(\mu)$ is given by

$$\begin{aligned} F_Z(z) &= \int_{(-\infty,z]} f_Y \mathbb{1}_{[0,+\infty)} \, \mathrm{d}\lambda \\ &= \int_{-\infty}^{z} \mu e^{-\mu \cdot y} \mathbb{1}_{[0,+\infty)}(y) \, \mathrm{d}y \\ &= \int_{0}^{z} \mu e^{-\mu \cdot y} \, \mathrm{d}y \\ &= \mu \, \tfrac{1}{-\mu} e^{-\mu \cdot y} \Big|_0^z \\ &= -e^{-\mu \cdot y} \Big|_0^z \\ &= 1 - e^{\mu \cdot z} \end{aligned}$$

With this information, we can conclude that

$$F_{T_1}(t) = \mathbb{P}(\{T_1 \leqslant t\}) = 1 - \mathbb{P}(\{T_1 > t\}) = 1 - e^{-\mu \cdot t} = F_Z(t) \quad \text{where } Z \sim \mathcal{E}(\mu).$$

Thus, the waiting time T_1 until the first hit/arrival has an exponential distribution with rate μ. ■

Uses:

- Exponentially distributed random variables are used to model, for example, the length of life of an electronic component e.g. a fuse.

As we shall see later on in **Proposition 10.5.1**, random variables that are exponentially distributed exhibit a property called memorylessness i.e. if $Y \sim \mathcal{E}(\mu)$, and $a, b > 0$, then

$$\mathbb{P}(\{Y > a + b\} \,|\, \{Y > a\}) = \mathbb{P}(\{Y > b\}).$$

This fact is used to prove that the T_i are identically distributed.

I won't do this now because it requires taking more care with conditional probability. When I do, it will be in a later chapter dedicated to stochastic processes.

- Waiting times are in a sense "dual" with the count of events. Let T_1 denote the time until the first event, T_2 the time from the first until the second event, etc. Then

  $$\{X_t := X([0,t]) \geqslant k\} \iff \{T_1 + \ldots + T_k \leqslant t\}.$$

  This relates Poisson counts to sums of waiting times. The proof of this relationship can be found in **Exercise 3**.

- The **Gamma distribution** with **rate parameter** naturally arises by summing i.i.d. exponentially distributed random variables (each representing the waiting time until a rare Poisson event occurs) e.g. if

  $$T_1, T_2, \ldots \overset{\text{i.i.d.}}{\sim} \mathcal{E}(\mu)$$

  are the inter-arrival times in a Poisson process, then the time of the n^{th} arrival/hit $W = T_1 + \ldots + T_n$ in the same Poisson process is Gamma(n, μ)-distributed.

I define these distributions in the next chapter, but I'll expand on Poisson processes in a later chapter.

CHAPTER 10

Absolutely Continuous Distributions

This chapter will go over some examples of different types of experiments, their associated absolutely continuous random variables of interest and their respective probability distributions.

10.1 Uniform Distribution

The probability of a random variable assuming a value between any two real numbers θ_1 and θ_2 is proportional to the reciprocal of the length of the interval $[\theta_1, \theta_2]$ between them. Sub-intervals (of $[\theta_1, \theta_2]$) with the same length have the same relative frequencies. We say that Y **is uniformly distributed between** θ_1 **and** θ_2, denoted by $Y \sim \text{Unif}(\theta_1, \theta_2)$, if the density function of the law $\mathbb{P}_Y$ of Y is

$$f_Y(y) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1}, & \theta_1 \leqslant y \leqslant \theta_2 \\ 0, & \text{otherwise.} \end{cases}$$

10.2 Normal Distribution

A random variable X is said to have a **normal probability distribution** if, for $\sigma > 0$ and μ s.t. $|\mu| < \infty$, $\mathbb{P}_X$ admits a density function defined for $x \in \mathbb{R}$ by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

In this case, we write $X \sim \mathcal{N}(\mu, \sigma^2)$.

Properties Let $X \sim \mathcal{N}(\mu, \sigma^2)$. It follows that $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

Proof. As per the discussion of densities in **Section 5.6.1**, we can compute the expectation of an absolutely continuous random variable X by the formula

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)\, \mathrm{d}\lambda(x).$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, we have that:

$$\begin{aligned}
\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x)\, \mathrm{d}\lambda(x) &= \int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} x \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\sigma^2}}\, \mathrm{d}x \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (u\sqrt{2\sigma^2} + \mu) \exp(-u^2)\, \mathrm{d}u \quad \text{via the substitution } u = \tfrac{x-\mu}{\sqrt{2\sigma^2}} \\
&= \frac{\sqrt{2\sigma^2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} u \exp(-u^2)\, \mathrm{d}u + \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-u^2)\, \mathrm{d}u \\
&= 0 + \frac{\mu}{\sqrt{\pi}} \sqrt{\pi} = \mu
\end{aligned}$$

The last line uses (without proof, but one can easily find the polar co-ordinate proof) the well-known Gaussian integral over $\mathbb{R}$. The first integral capitalises on the symmetry of $\exp(-u^2)$ which

tells us that $u\exp(-u^2)$ is an odd-function; integrating it over a symmetric domain about the origin (like $\mathbb{R}$) gives us 0. ■

10.2.1 Standard Normal

Every normally distributed random variable Y with mean μ and variance σ^2 can be transformed into a standard normal random variable Z via

$$Z \coloneqq \frac{Y-\mu}{\sigma}.$$

Z represents the position of a point relative to the mean of a normal random variable, with the distance measured in terms of how many standard deviations away it is from the mean.

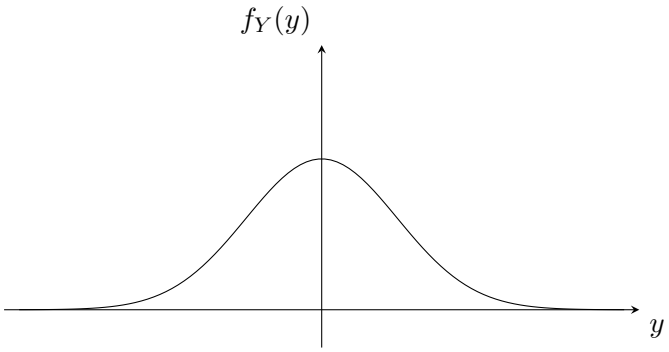


Figure 10.1: The density of $Z \sim \mathcal{N}(0,1)$. The y-intercept is $1/\sqrt{2\pi} \approx 0.3989$.

Properties $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

Uses:

- The symmetry and bell shape of normal distributions are useful properties for modelling other mound-shaped distributions which often account for naturally-occurring phenomena.
- The density has some good analytical properties.
- As seen later, central limit theorems offer a way to asymptotically approximate any population distribution with a normal distribution.

10.2.2 Link: Normal Approximates Binomial

For a fixed value of p and for large n, one can approximate the binomial distribution by a normal distribution when the number of successes is within some appropriate range of np. More precisely:

Theorem 10.2.1 (Theorem 5 [9, p. 214]) Suppose $0 < p < 1$; put $q = 1 - p$, and

$$x_{n,k} = \frac{k-np}{\sqrt{npq}}, \quad 0 \leqslant k \leqslant n.$$

Clearly $x_{n,k}$ depends on both n and k, but it will be written as x_k below.

Let A be an arbitrary but fixed positive constant. Then in the range of k such that $|x_k| \leqslant A$, we have that

$$\binom{n}{k} p^k q^{n-k} \sim \frac{1}{\sqrt{2\pi npq}} e^{-x_k^2/2}.$$

The convergence is uniform with respect to k in the range specified above.

Proof Sketch. Pick $A > 0$. For k in the specified range, we have that $|x_k| \leqslant A$ where A is some fixed positive constant. Now we do some work to get inequalities that are only satisfied by k in the specified range.

- The inequality $|x_k| \leqslant A$ rearranges to

$$np - A\sqrt{npq} \leqslant k \leqslant np + A\sqrt{npq}$$

and one can imagine that since A, p, and q are fixed, as n gets very large the dominating term becomes np on both ends of this inequality so $k \sim np$.

  - From this, we can conclude the other approximation

$$k \sim np \implies n - k \sim n - np = n(1-p) = nq.$$

- By Stirling's formula[1], we can asymptotically approximate the factorial $n!$ in the binomial density to obtain

$$\binom{n}{k} p^k q^{n-k} \sim \frac{1}{\sqrt{2\pi npq}} \varphi(n,k) \quad \text{where} \quad \varphi(n,k) = \left(\tfrac{np}{k}\right)^k \left(\tfrac{nq}{n-k}\right)^{n-k}.$$

By taking logarithms, expanding the appropriate Taylor series where the expansions are valid for sufficiently large n, we obtain that

$$\log \varphi(n,k) \sim -\frac{n^2 pq x_k^2}{2k(n-k)}.$$

Now we use the original inequalities we derived to replace k and $n-k$. Finally, we end up with

$$\log \varphi(n,k) \sim -\frac{x_k^2}{2} \implies \varphi(n,k) \sim e^{-x_k^2/2}.$$

Thus concludes the proof sketch.

■

Now we state the main theorem of this subsection — the De Moivre-Laplace Theorem.

Theorem 10.2.2 (Theorem 6 [9, pp. 215–216]) Let $S_n \sim \text{Binom}(n,p)$. For any two constants a and b, $-\infty < a < b < +\infty$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{S_n - np}{\sqrt{npq}} \leqslant b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, \mathrm{d}x.$$

Historically speaking, this was the first known particular case of "the" central limit theorem, but more on this later in **Section 20.5**.

10.3 Gamma Distribution

As we've already seen in **Chapter 9**, the Poisson, Exponential, and Gamma distributions arise naturally from Poisson processes. I will expand on these in a later chapter but for now I present these distributions via their densities, describe their shapes, and make a brief comment on a formula connecting the Poisson and Gamma densities.

[1] Given by $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$, and denoted $n! \sim \ldots$.

Definition 10.3.1 Let $\alpha, \beta > 0$. A random variable Y is said to have a **Gamma distribution with parameters α and β**, denoted $Y \sim \text{Gamma}(\alpha, \beta)$, if its density function is given by:

$$f_Y(y) = \begin{cases} \dfrac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} & \text{if } 0 \leqslant y < \infty, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Gamma(\alpha) = \displaystyle\int_0^\infty y^{\alpha-1}e^{-y}\,\mathrm{d}y$.

The way this Gamma distribution has been defined involves a **scale** parameter β. This is in contrast to the **rate** parameter μ from the exposition on Poisson processes. They are reciprocals of each other i.e.

$$\beta = 1/\mu.$$

The density of $Y \sim \text{Gamma}(\alpha, \mu)$, where μ is a rate parameter is given by

$$f_Y(y) = \frac{\mu^\alpha y^{\alpha-1}e^{-\mu \cdot y}}{\Gamma(\alpha)} \mathbb{1}_{[0,+\infty)}(y).$$

Properties Let $\alpha, \beta > 0$ and $Y \sim \text{Gamma}(\alpha, \beta)$. Then $\mathbb{E}(Y) = \alpha\beta$ and $\text{Var}(Y) = \alpha\beta^2$.

Proof. The result follows from a simple integration by parts.

$$\begin{aligned}
\mathbb{E}(Y) &= \int_{\mathbb{R}} y f_Y(y)\,\mathrm{d}\lambda(y) \\
&= \int_{-\infty}^{\infty} y \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)} \mathbb{1}_{[0,\infty)}(y)\,\mathrm{d}y \\
&= \frac{1}{\beta^\alpha\Gamma(\alpha)} \int_0^\infty y^\alpha e^{-y/\beta}\,\mathrm{d}y \\
&= \frac{1}{\beta^\alpha\Gamma(\alpha)} \Big(\underbrace{y^\alpha(-\beta)e^{-y/\beta}\Big|_0^\infty}_{=0} - \int_0^\infty \alpha y^{\alpha-1}(-\beta)e^{-y/\beta}\,\mathrm{d}y \Big) \\
&= \alpha\beta \int_0^\infty \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\,\mathrm{d}y \\
&= \alpha\beta
\end{aligned}$$

since the final integral is that of the density f_Y over its support, and is hence equal to 1. ■

Unfortunately, if $\alpha \in \mathbb{R}_{>0} \setminus \mathbb{N}$ and $0 < c < d < \infty$, then it's impossible to give a closed-form expression for

$$\int_c^d \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\,\mathrm{d}y.$$

10.3.1 Shape

The number $\alpha > 0$ is called the **shape parameter** of the Gamma distribution.

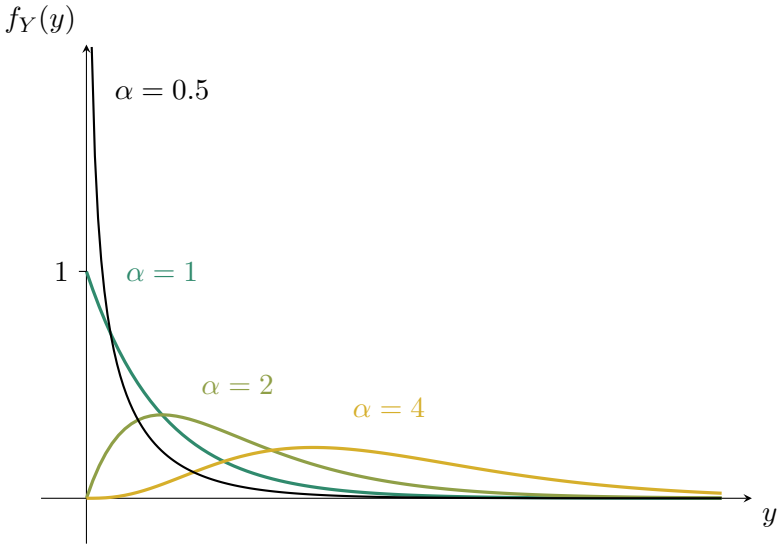


Figure 10.2: Plots for the densities of $Y \sim \Gamma(\alpha, \beta = 1)$ for $\alpha = 0.5$, 1, 2, and 4.

The plot for $\alpha = 1$ looks different in shape to the others and is considered a density in its own right (parameterised by β), called the **exponential density**.

> Curiously, the Gamma density for $0 < \alpha < 1$ has an asymptote at $y = 0$ where it veers off to $+\infty$. After a bit of searching, this family doesn't seem to have a name of its own.

Increasing α leads to a more peaked distribution nearer to 0.

10.3.2 Scale Parameter

Definition 10.3.2 A **scale parameter** is a parameter that specifies the spread of a distribution. More precisely, let $F(y; s, \theta)$ denote a family of cumulative distribution functions. If the parameter s is such that

$$F(y; s, \theta) = F(y/s; 1, \theta)$$

then s is called a scale parameter.

Corollary 10.3.3 By differentiating with respect to y, the corresponding density statement says that s is a scale parameter if

$$f(y; s, \theta) = \frac{1}{s} f(y; 1, \theta).$$

Example 10.3.4 In the case of the gamma density, let's verify that β is indeed a scale parameter.

Proof. $\frac{1}{\beta} f(y/\beta; \alpha, 1) = \frac{1}{\beta} \frac{1}{1^\alpha \Gamma(\alpha)} \left(\frac{y}{\beta}\right)^{\alpha-1} e^{-(y/\beta)/1} = \frac{1}{\beta} \frac{1}{\beta^{\alpha-1}} \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} = f(y; \alpha, \beta)$ ■

The effect of different values of β for fixed values of $\alpha = 1$ and $\alpha = 2$ respectively, are illustrated below:

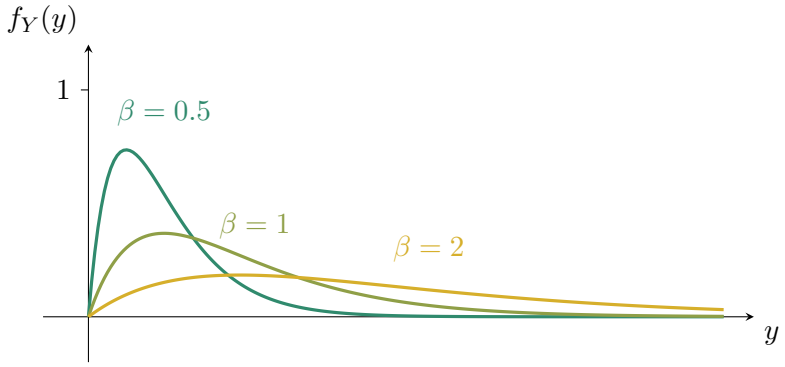


Figure 10.3: Let $\alpha = 2$. Plots for the densities of $Y \sim \Gamma(\alpha, \beta)$ for $\beta = 0.5$, 1, and 2.

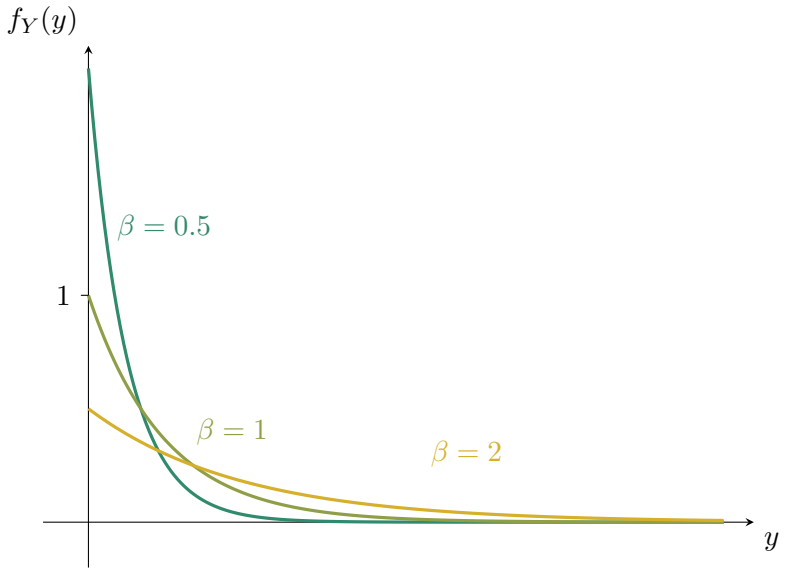


Figure 10.4: Let $\alpha = 1$. Plots for the densities of $Y \sim \Gamma(\alpha, \beta)$ for $\beta = 0.5$, 1, and 2. These are plots for the exponential density.

Notice that increasing β spreads the density out more over its support $[0, +\infty)$. Since $\alpha = 1 \in \mathbb{N}$, we can appeal to the Poisson process interpretation of the parameter β. Recall that the rate of a Poisson process is $\mu = 1/\beta$. It's sensible to remark that increasing $\beta > 0$ decreases $\mu > 0$, so the intensity of the process is lower which ostensibly means a longer time required until a Poisson hit/arrival. Equally, the scale parameter β measures the average time until the first Poisson arrival/occurrence since $\mathbb{E}(Y) = \beta$, where $Y \sim \Gamma(\alpha = 1, \beta)$.

10.3.3 Link: Poisson and Gamma

In the special case that $\alpha = n \in \mathbb{N}$, the distribution function of a gamma distributed random variable can be expressed as a sum of Poisson probabilities.

We default back to the Poisson process example for insight. Let $\{X_t\}_{t\in[0,+\infty)}$ be a Poisson process with rate μ. Then $X_0 = 0$ and $X_t \sim \text{Poisson}(\mu \cdot t)$ for $t > 0$. The waiting time until the n^{th} Poisson arrival is the random variable $W_n \sim \text{Gamma}(n, \mu)$. Equivalently, we may write it with a scale parameter satisfying $\mu = 1/\beta$. The n^{th} arrival happening after time t is equivalent to there only being $n - 1$ Poisson arrivals in the interval $[0, t]$. We can write this relationship as follows:

$$\{W_n > t\} \iff \{X_t \leqslant n - 1\}.$$

Exercise 3 (Adapted from 4.79 [6]) Given the above setting, take probabilities of the "dual" events to relate the distribution functions of $W_n \sim \text{Gamma}(n, \mu)$ and $X_t \sim \text{Poisson}(\mu \cdot t)$ with the formula

$$\mathbb{P}(\{W_n > t\}) = \int_t^\infty \frac{\mu^n}{\Gamma(n)} y^{n-1} e^{-\mu y} \, \mathrm{d}y = \sum_{x=0}^{n-1} \frac{(\mu \cdot t)^x e^{-\mu \cdot t}}{x!} = \mathbb{P}(\{X_t \leqslant n-1\}).$$

Proof. Since $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$. Now we derive a recursive integration formula ($\diamond$) by integrating by parts.

$$\begin{aligned}
I(n) &= \int_\lambda^\infty y^{n-1} e^{-y} \, \mathrm{d}y \\
&= y^{n-1}(-1)e^{-y}\Big|_\lambda^\infty - \int_\lambda^\infty (n-1) y^{n-2}(-e^{-y}) \, \mathrm{d}y \\
&= \lambda^{n-1} e^{-\lambda} + (n-1) I(n-1) \\
&= \lambda^{n-1} e^{-\lambda} + (n-1)\Big(\lambda^{n-2} e^{-\lambda} + (n-2) I(n-2)\Big) \\
&= \Big(\lambda^{n-1} e^{-\lambda} + (n-1)\lambda^{n-2}\Big) e^{-\lambda} + (n-1)(n-2) I(n-2) \\
&= \Big(\lambda^{n-1} e^{-\lambda} + (n-1)\lambda^{n-2} + (n-1)(n-2)\lambda^{n-3}\Big) e^{-\lambda} + (n-1)(n-2)(n-3) I(n-3) \\
&= \dots \\
&= \Big(\lambda^{n-1} e^{-\lambda} + (n-1)\lambda^{n-2} + \dots + (n-1)(n-2)\dots(n-(n-2))\lambda^{n-(n-1)}\Big) e^{-\lambda} \\
&\quad + (n-1)! I(1) \\
&= \Big(\lambda^{n-1} e^{-\lambda} + (n-1)\lambda^{n-2} + \dots + (n-1)(n-2)\dots(2)\lambda^1\Big) e^{-\lambda} + (n-1)! e^{-\lambda} \\
&= \Big(\lambda^{n-1} + (n-1)\lambda^{n-2} + \dots + (n-1)(n-2)\dots(2)\lambda^1 + (n-1)! + 1\Big) e^{-\lambda} \\
&= \sum_{k=0}^{n-1} \frac{(n-1)! \lambda^k e^{-\lambda}}{k!}.
\end{aligned}$$

In order to use ($\diamond$), we massage our desired integral into the appropriate form with a substitution:

$$\begin{aligned}
\mathbb{P}(\{W_n > t\}) &= \int_t^\infty \frac{\mu^n}{\Gamma(n)} y^{n-1} e^{-\mu y} \, \mathrm{d}y \\
&= \frac{1}{\Gamma(n)} \int_{\mu t}^\infty u^{n-1} e^{-u} \, \mathrm{d}u \quad \text{via } u = \mu y \\
&= \frac{1}{\Gamma(n)} \sum_{k=0}^{n-1} \frac{(n-1)! (\mu \cdot t)^k e^{-\mu \cdot t}}{k!} \quad \text{by } (\diamond) \text{ with } \lambda = \mu t \\
&= \mathbb{P}(\{X_t \leqslant n-1\})
\end{aligned}$$

■

10.4 Gamma($\nu/2, \beta = 2$), The Chi-Squared Distribution ($\nu \in \mathbb{N}$)

Let $\nu \in \mathbb{N}$. A random variable Y is said to have a **chi-squared distribution[2] with ν degrees of freedom**, denoted by $Y \sim \chi^2_\nu$, iff $Y \sim \text{Gamma}(\alpha = \frac{\nu}{2}, \beta = 2)$.

Properties

- $\mathbb{E}(Y) = \nu$

[2]Tables for the Gamma distribution are not as readily available as tables for χ^2 distributions which are available for many values of ν. Using a transformation means one can use χ^2 table values for Gamma distributions.

- $\mathrm{Var}(Y) = 2\nu$

Uses:

- Commonplace in statistical inference. Namely, this is the distribution of the sum of the squares of ν independent standard normal random variables.

10.5 Gamma($\alpha = 1, \beta$), The Exponential Distribution

When $\alpha = 1$ in a Gamma distribution, we have what is known as the exponential distribution. Y is said to have an **exponential distribution with parameter** $\beta > 0$, denoted $Y \sim \mathcal{E}(\beta)$, if its density function is:

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta} & \text{if } 0 \leqslant y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Properties

- $\mathbb{E}(Y) = \beta$
- $\mathrm{Var}(Y) = \beta^2$

10.5.1 Memorylessness

Suppose that the length of time a component has already operated doesn't affect its chance of operating for at least b additional time units. This is formalised by the following property called the **memorylessness** of the exponential distribution:

Proposition 10.5.1 (**Example 4.10 [6]** Memorylessness of $\mathcal{E}(\beta)$) Let $a, b > 0$. If $\mathbb{P}(\{Y > a\}) > 0$, prove that

$$\mathbb{P}(\{Y > a + b\} \,|\, \{Y > a\}) = \mathbb{P}(\{Y > b\}).$$

Proof. First note that

$$\mathbb{P}(\{Y > c\}) = \int_c^{+\infty} f_Y(y)\,\mathrm{d}y = \int_c^{+\infty} \frac{1}{\beta} e^{-y/\beta}\,\mathrm{d}y = \frac{1}{\beta} \frac{1}{-\frac{1}{\beta}} e^{-y/\beta} \Big|_c^{+\infty} = e^{-c/\beta}.$$

From the definition of conditional probability:

$$\begin{aligned} \mathbb{P}(\{Y > a + b\} \,|\, \{Y > a\}) = \frac{\mathbb{P}(\{Y > a + b\} \cap \{Y > a\})}{\mathbb{P}(\{Y > a\})} = \frac{\mathbb{P}(\{Y > a + b\})}{\mathbb{P}(\{Y > a\})} &= \frac{e^{-(a+b)/\beta}}{e^{-b/\beta}} \\ &= e^{-b/\beta} \\ &=: \mathbb{P}(\{Y > b\}) \end{aligned}$$

■

10.5.2 Link: Exponential and Geometric

The exponential distribution is the continuous-time analogue of the geometric distribution.

Exercise 4 (4.75 [6]) Let $Y \sim \mathcal{E}(\beta)$. Define a random variable X in the following way: $X = k$ if $k - 1 \leqslant Y < k$ for $k = 1, 2, \ldots$.

(a) Find $\mathbb{P}(\{X = k\})$ for each $k = 1, 2, \ldots$.

(b) Show that the answer to (a) can be written as:

$$\mathbb{P}(\{X = k\}) = (e^{-1/\beta})^{k-1}(1 - e^{-1/\beta}), \quad k = 1, 2, \ldots$$

and that X has a geometric distribution with success probability $p = 1 - e^{-1/\beta}$.

Proof.

$$\begin{aligned}
\mathbb{P}(\{X = k\}) = \mathbb{P}(\{\omega \in \Omega \colon k - 1 \leqslant Y < k\}) &= \int_{k-1}^{k} f_Y(y)\,\mathrm{d}y \\
&= \int_{k-1}^{k} \frac{1}{\beta} e^{-y/\beta}\,\mathrm{d}y \\
&= \frac{1}{\beta}\frac{1}{-\frac{1}{\beta}}\, e^{-y/\beta}\Big|_{k-1}^{k} \\
&= \frac{1}{e^{(k-1)/\beta}} - \frac{1}{e^{k/\beta}} \\
&= (e^{-1/\beta})^{k-1} \underbrace{(1 - e^{-1/\beta})}_{= p}
\end{aligned}$$

■

10.6 Beta Distribution

Two parameters $\alpha, \beta > 0$. Defined over $[0, 1]$. $Y \sim \text{Beta}(\alpha, \beta)$ if the density function is

$$f_Y(y) = \begin{cases} \dfrac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha, \beta)} & \text{if } 0 \leqslant y \leqslant 1 \\ 0 & \text{otherwise,} \end{cases}$$

where

$$B(\alpha, \beta) \coloneqq \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,\mathrm{d}y = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Despite being defined over $[0, 1]$, we can fit a beta density function to a random variable Y on any interval e.g. $[c, d]$ by defining a new random variable $Y^* = \frac{Y-c}{d-c}$.

Uses:

- ???

10.6.1 Link: Beta and Binomial

The cumulative distribution function for a beta random variable is commonly called the incomplete beta function and is defined by

$$F_Y(y) = \int_0^y \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)}\,\mathrm{d}t = I_y(\alpha, \beta).$$

When $\alpha, \beta \in \mathbb{Z}$, $I_y(\alpha, \beta)$ is related to the binomial probability function: For $y \in (0, 1)$ and $\alpha, \beta \in \mathbb{Z}$:

$$F_Y(y) = \int_0^y \frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)}\,\mathrm{d}t = \sum_{i=\alpha}^{n} \binom{n}{i} y^i (1-y)^{n-i}$$

where $n = \alpha + \beta - 1$. That term in blue is a sum of probabilities associated with a binomial random variable with $n = \alpha + \beta - 1$ trials and $p = y$ probability of "success."

$$\mathbb{E}(Y) = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

10.7 Chebyshev's Theorem

Theorem 10.7.1 (Chebyshev's Theorem) Let Y be a random variable with finite mean μ and variance σ^2. Then for any $k > 0$,

$$\mathbb{P}(\{|Y-\mu| < k\sigma\}) \geqslant 1 - \frac{1}{k^2} \quad \text{or} \quad \mathbb{P}(\{|Y-\mu| \geqslant k\sigma\}) \leqslant \frac{1}{k^2}$$

This theorem enables us to find bounds on probabilities that are ordinarily tedious to calculate.

Proof. Let $f_Y(y)$ denote the density of Y. Then

$$\begin{aligned}
\mathrm{Var}(Y) = \sigma^2 &= \int_{-\infty}^{\infty} (y-\mu)^2 f_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\mu-k\sigma} (y-\mu)^2 f_Y(y)\,\mathrm{d}y + \underbrace{\int_{\mu-k\sigma}^{\mu+k\sigma} (y-\mu)^2 f_Y(y)\,\mathrm{d}y}_{\geqslant 0} + \int_{\mu+k\sigma}^{\infty} (y-\mu)^2 f_Y(y)\,\mathrm{d}y \\
&\geqslant \int_{-\infty}^{\mu-k\sigma} (y-\mu)^2 f_Y(y)\,\mathrm{d}y + \int_{\mu+k\sigma}^{\infty} (y-\mu)^2 f_Y(y)\,\mathrm{d}y
\end{aligned}$$

The middle integral on the 2nd line is bounded below by 0 and the other two integrals have $(y-\mu)^2$ in the integrand bounded below by $k\sigma^2$.

$$\begin{aligned}
\mathrm{Var}(Y) = \sigma^2 &\geqslant (k\sigma)^2 \int_{-\infty}^{\mu-k\sigma} f_Y(y)\,\mathrm{d}y + (k\sigma)^2 \int_{\mu+k\sigma}^{\infty} f_Y(y)\,\mathrm{d}y \\
&=: (k\sigma)^2 \cdot \mathbb{P}(\{Y \leqslant \mu - k\sigma\}) + (k\sigma)^2 \cdot \mathbb{P}(\{Y \geqslant \mu + k\sigma\}) \\
&=: (k\sigma)^2 \cdot \mathbb{P}(\{|Y-\mu| \geqslant k\sigma\})
\end{aligned}$$

The desired inequality follows and the other can be obtained by taking the complement. ■

10.8 Expectations of Discontinuous Functions and Mixed Probability Distributions

- We may be interested in $\mathbb{E}(g(Y))$ where g is discontinuous.
- Y itself may have a distribution function that is continuous over some intervals and such that some isolated points have positive probabilities.

A mixed distribution can be uniquely written as

$$F_Y(y) = c_1 F_1(y) + c_2 F_2(y)$$

where

- $F_1(y)$ is a step distribution function
- $F_2(y)$ is a continuous distribution function
- c_1 is the accumulated probability of all discrete points
- c_2 is the accumulated probability of all continuous portions/intervals
- $c_1 + c_2 = 1$

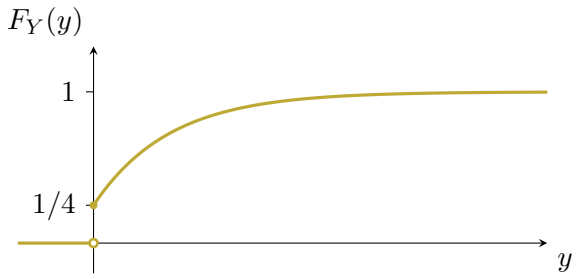
Example 10.8.1 Y is equal to the length of life of electronic components. The components frequently fail immediately upon insertion into the system with probability 1/4. If it doesn't fail immediately, the distribution for its length of life has the exponential density function

$$f(y) = \begin{cases} e^{-y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Since $Y = 0$ is the only discrete point, $c_1 = 1/4$. This means that $c_2 = 1 - (1/4) = 3/4$. It follows that Y is a mixture of the distribution of two random variables X_1 and X_2 where X_1 has probability 1 at point 0, and X_2 has the given exponential density $f(y)$. The distribution functions of X_1 and X_2 are F_1 and F_2 respectively:

- $F_1 = \begin{cases} 0, & y < 0 \\ 1, & y \geqslant 0. \end{cases}$
- $F_2 = \begin{cases} 0, & y < 0 \\ \displaystyle\int_0^y e^{-t}\,\mathrm{d}t = 1 - e^{-y}, & y \geqslant 0. \end{cases}$

Therefore, $F_Y(y) = \dfrac{1}{4}F_1(y) + \dfrac{3}{4}F_2(y)$ and its graph is:



10.8.1 Expectation of a Mixed Random Variable

As before, let Y have the mixed distribution function

$$F_Y(y) = c_1F_1(y) + c_2F_2(y).$$

Let $g(Y)$ denote a function of Y. Then

$$\mathbb{E}(g(Y)) = c_1\mathbb{E}(g(X_1)) + c_2\mathbb{E}(g(X_2)).$$

10.9 Summary

- Density functions provide models for population frequency distributions
  - This yields a mechanism for inferring characteristics of the population based on measurements contained in a sample taken from said population.
- Four types of absolutely continuous random variable were presented — uniform, gamma (special cases were χ^2_ν and exponential), normal and beta.

10.10 Location-Scale Families

In the preceding examples of discrete and absolutely continuous distributions, we've (occasionally) made (implicit) reference to the parameters that determine their properties. Such parameters fall under different names:

Definition 10.10.1 A **location parameter** is a parameter that determines the "location" or shift of a frequency distribution in the sense of defining a central or typical value such as the mean or mode.

e.g. A common example of a location parameter is μ in the normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Definition 10.10.2 A **scale parameter** is a parameter that specifies the spread of a distribution.

e.g. The example of σ^2 in $\mathcal{N}(\mu, \sigma^2)$ is the *squared scale* of the normal distribution.

e.g. Let $Y \sim \mathcal{E}(\beta)$ so the density of the exponential random variable is defined by

$$f_Y(y) = \tfrac{1}{\beta} e^{-y/\beta} \mathbb{1}_{[0,\infty)}(y).$$

The parameter $\beta > 0$ is the scale parameter.

Definition 10.10.3 The inverse of a scale parameter is often called a **rate parameter**.

e.g. For $Y \sim \mathcal{E}(\beta)$, we could equivalently have let $\lambda = 1/\beta$ and written the density as

$$f_Y(y) = \lambda e^{-\lambda y} \mathbb{1}_{[0,\infty)}(y).$$

In this case, λ is the rate parameter of the distribution.

Probability distributions can be grouped into families with common functional forms.

Definition 10.10.4 A family of probability distributions $\{\mathbb{P}_{\mu,\sigma} \colon \mu \in \mathbb{R}, \sigma > 0\}$ parameterised by a location parameter μ and a non-negative scale parameter $\sigma > 0$ is called a **location-scale family** if there exists a fixed probability measure $\mathbb{P}_0$ s.t for each (μ, σ):

$$\mathbb{P}_{\mu,\sigma} = (T_{\mu,\sigma})_{\#}\mathbb{P}_0$$

where $T_{\mu,\sigma}(x) = \sigma x + \mu$ is an affine transformation.

In the language of random variables, let Z be the random variable with distribution $\mathbb{P}_0$. Then for (μ, σ), we have that $X = \sigma Z + \mu$ has distribution $\mathbb{P}_{\mu,\sigma} = (T_{\mu,\sigma})_{\#}\mathbb{P}_0$.

When constructing a location-scale family, one typically takes $\mathbb{P}_0$ to be the distribution of a random variable Z in **standard measure** — having zero mean and unit variance.

Lemma 10.10.5 A location-scale family $\{\mathbb{P}_{\mu,\sigma} \colon \mu \in \mathbb{R}, \sigma > 0\}$ is closed under affine transformations.

Proof. Suppose that $X \sim \mathbb{P}_{\mu,\sigma}$ for some $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{>0}$. Let $0 \neq a \in \mathbb{R} \ni b$ and Y be an affine translate of X i.e. $Y = aX + b$. By definition of the location-scale family, $\exists Z \sim \mathbb{P}_0$ s.t. $X = \sigma Z + \mu$. Therefore,

$$Y = a(\sigma Z + \mu) + b = (a\sigma)Z + (\mu + b)$$

and we can recognise the distribution of Y as the pushforward of $\mathbb{P}_0$ under the affine (note that $\sigma > 0$ and $a \neq 0$) transformation $T_{a\sigma,\mu+b}$ so Y is a member of the location-scale family. ■

e.g. $\mathcal{U}[a, b]$: Let $Z \sim \mathcal{U}[0, 1]$ and identify μ and σ by letting $X \sim \mathcal{U}[a, b]$ and noticing that we can scale Z by $(b - a)$ first and then translate by a. Therefore, $X = (b - a)Z + a$.

e.g. $\mathcal{N}(\mu, \sigma^2)$: Let $Z \sim \mathcal{N}(0, 1)$ and follow similar logic to the above example.

Chapter 11

Moment-Generating Functions

Moments are values used to characterise the probability distributions of random variables.

- The **k^{th} moment of a random variable X about the origin** is denoted μ'_k and defined by
$$\mu'_k := \mathbb{E}\Big(X^k\Big).$$
  - e.g. $\mu'_1 = \mathbb{E}(X) = \mu$
- The **k^{th} moment of a random variable X about its mean** (also called the **k^{th} central moment of X**) is denoted μ_k and defined as
$$\mu_k := \mathbb{E}\Big((X-\mu)^k\Big).$$
  - e.g. $\mu_2 = \mathbb{E}\big((X-\mu)^2\big) = \mathrm{Var}(X) = \sigma^2$.

Under certain conditions (**Theorem 14.5.2**) moments can uniquely determine the probability distribution of X.

The **moment-generating function of a random variable X** (or **MGF**), denoted by $M_X(t)$, is defined as
$$M_X(t) := \mathbb{E}\big(e^{tX}\big).$$
We say that a moment-generating function for X exists if the expectation above exists in a neighbourhood of the origin i.e. $\exists b > 0$ such that $|t| < b \implies M_X(t) < \infty$.

Example 11.0.1 Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with μ and σ finite. The moment-generating function of X is given by:
$$\begin{aligned}
M_X(t) &:= \mathbb{E}(\exp(tX)) \\
&= \int_{\mathbb{R}} \exp(tx) f_X(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}} \exp(tx) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\
&= \dots \\
&= \exp\left(t\mu + \frac{t^2}{2}\sigma^2\right) \underbrace{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-(\mu+t\sigma^2))^2}{2\sigma^2}\right)\mathrm{d}x}_{= \int_{\mathbb{R}} f_Y(y)\,\mathrm{d}y = 1,\ \text{where } Y \sim \mathcal{N}(\mu+t\sigma^2, \sigma^2)}
\end{aligned}$$
Finally, we conclude that
$$M_X(t) = \exp\left(t\mu + \frac{t^2}{2}\sigma^2\right).$$

An invaluable theorem (that we'll again see later as **Theorem 14.5.1**) tells us that if two random variables have the same MGF, then they have the same distribution.

> **Theorem 11.0.2** (Theorem 6.1 [6]) Let $M_X(t)$ and $M_Y(t)$ denote the moment-generating functions of random variables X and Y respectively. If both MGFs exist, and for all t in some neighbourhood of 0: $M_X(t) = M_Y(t)$, then X and Y have the same probability distribution.

We'll use this theorem to talk a bit more about location-scale families from the end of last chapter:

Lemma 11.0.3 If $Z \sim \mathbb{P}_0$ where $\mathbb{P}_0$ is a member of the location-scale family $\{\mathbb{P}_{\mu,\sigma} \colon \mu \in \mathbb{R}, \sigma > 0\}$, then the distribution of $Y = aZ + b$, where $a \neq 0$ and $b \in \mathbb{R}$, is also a member of the location-scale family.

Proof.

$$M_Y(t) \coloneqq \mathbb{E}(\exp(tY)) = \mathbb{E}(\exp(t(aZ+b))) = e^{tb}\mathbb{E}(\exp(atZ)) =: e^{tb}M_Z(at)$$

Notice that the functional form of the moment-generating function does not change and so this corresponds to a probability distribution in the location-scale family. ■

Example 11.0.4 Let $Z \sim \mathcal{N}(0,1)$ and consider $X = aZ + b$. The MGF of Z is given by

$$M_X(t) \overset{11.0.3}{\coloneqq} e^{tb}M_Z(ta) \overset{11.0.1}{=} e^{tb}\left(\exp\left(\mu t + \tfrac{(ta)^2}{2}\sigma^2\right)\Big|_{\mu=0,\,\sigma=1}\right) = \exp\left(t\mu + \tfrac{t^2}{2}\sigma^2\right)$$

This matches up with our expression from **Example 11.0.1** with $a = \sigma$ and $\mu = b$.

Example 11.0.5 The MGF of $X \sim \text{Gamma}(\alpha, \beta)$ where $\alpha, \beta > 0$ is

$$\begin{aligned} M_X(t) &\coloneqq \mathbb{E}(\exp(tX)) \\ &= \int_{\mathbb{R}} e^{tx}\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\mathbb{1}_{[0,+\infty)}(x)\,\mathrm{d}x \\ &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}\exp(-x(-t+\tfrac{1}{\beta}))\,\mathrm{d}x \end{aligned}$$

This integral only converges if $(-t+\frac{1}{\beta}) > 0$ i.e. $t < \frac{1}{\beta}$. Now we use the substitution $u = x(-t+\frac{1}{\beta})$ so that $\mathrm{d}u = (-t+\frac{1}{\beta})\,\mathrm{d}x$ and:

$$\begin{aligned} M_X(t) &= \frac{1}{\beta^\alpha\Gamma(\alpha)}\int_0^\infty \left(\left(-t+\tfrac{1}{\beta}\right)^{-1}u\right)^{\alpha-1}\exp(-u)\left(-t+\tfrac{1}{\beta}\right)^{-1}\mathrm{d}u \\ &= \frac{1}{\beta^\alpha\Gamma(\alpha)}((-t+\tfrac{1}{\beta})^{-1})^\alpha \underbrace{\int_0^\infty u^{\alpha-1}\exp(-u)\,\mathrm{d}u}_{=:\Gamma(\alpha)} \\ &= \frac{1}{(1-\beta t)^\alpha} \end{aligned}$$

11.1 Technical Points

This subsection is a paraphrased version of this **Cross Validated answer** from Stack Exchange user `cardinal`.

The right conditions under which we can say something about the moments of X based on its MGF are contained in the following proposition:

Proposition 11.1.1 If there exist $t_n < 0$ and $t_p > 0$ such that $m(t_n) < \infty$ and $m(t_p) < \infty$, then the moments of all orders of X exist and are finite.

The contrapositive of this proposition says that if any of the moments of X are infinite or do not exist, we can immediately conclude that the MGF is not finite in an open neighbourhood of the origin.

Lemma 11.1.2 Suppose that such t_n and t_p exist. Then for any $t_0 \in [t_n, t_p]$, $M_X(t_0) < \infty$.

Proof. For any such t_0, $\exists \lambda \in [0,1]$ s.t. $t_0 = t_n + \lambda(t_p - t_n) = (1-\lambda)t_n + \lambda t_p$. By the convexity of $\exp(\cdot)$, it follows that

$$e^{t_0 X} = e^{((1-\lambda)t_n + \lambda t_p)X} \leqslant (1-\lambda)e^{t_n X} + \lambda e^{t_p X}.$$

We conclude by using the monotonicity of expectation:

$$\mathbb{E}\big(e^{t_0 X}\big) \leqslant \lambda \underbrace{\mathbb{E}\big(e^{t_n X}\big)}_{<\infty} + (1-\lambda)\underbrace{\mathbb{E}\big(e^{t_p X}\big)}_{<\infty} < \infty.$$

■

Proposition 11.1.3 The MGF $M_X(t)$ is finite in an open neighbourhood (t_n, t_p) of the origin if and only if the tails of the distribution of X are exponentially bounded i.e. $\mathbb{P}(\{|X| > x\}) \leqslant Ce^{t_0 x}$ for some $C > 0$ and $t_0 > 0$.

If $M_X(t)$ is finite in some open neighbourhood of the origin, then it determines the distribution of X i.e. it's the only distribution with the moments $\mu'_k = \mathbb{E}\big(X^k\big)$.

11.2 Generating Moments

Suppose that the mgf is well-defined on a neighbourhood $N = (t_n, t_p)$ of the origin. Consider $\delta = \min\{-t_n, t_p\}$. Then $M_X(t)$ exists and is finite in $(-\delta, \delta)$. Since $\exp(tX)$ has a Taylor series, we may write for $t \in (-\delta, \delta)$:

$$M_X(t) \coloneqq \mathbb{E}\big(e^{tX}\big) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right)$$

We can exchange the expectation and infinite sum via:

Theorem 11.2.1 (Dominated Convergence Theorem) Suppose that $\{f_n\}_n \subseteq L^1(X, \mathcal{M}, \mu)$ is a sequence such that:

- $f_n \to f$ a.e.
- The f_n are uniformly bounded i.e. there exists a non-negative $g \in L^1(X, \mathcal{M}, \mu)$ such that $\forall n \colon |f_n| \leqslant g$ a.e.

Then $f \in L^1$ and

$$\int \lim_{n\to\infty} f_n = \lim_{n\to\infty} \int f_n.$$

Let $f_n = \sum_{k=0}^{n} \frac{(tX)^k}{k!}$. Then the f_n are uniformly bounded above by $e^{|tX|}$. Also note that $e^{|tX|} \leqslant e^{tX} + e^{-tX}$. Then $e^{|tX|}$ is indeed integrable in $(-\delta, \delta)$ because monotonicity of the integral implies that

$$\mathbb{E}\Big(e^{|tX|}\Big) \leqslant \mathbb{E}\big(e^{tX} + e^{-tX}\big) = \mathbb{E}\big(e^{tX}\big) + \mathbb{E}\big(e^{-tX}\big) =: M_X(t) + M_X(-t) < \infty.$$

Therefore,

$$\begin{aligned} M_X(t) &= \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right) \\ &= \sum_{n=0}^{\infty} \mathbb{E}\left(\frac{(tX)^n}{n!}\right) \quad \text{by the Dominated Convergence Theorem} \\ &= \sum_{n=0}^{\infty} \frac{\mathbb{E}(X^n)t^n}{n!} \quad \text{by linearity of } \mathbb{E}(\,\cdot\,) \end{aligned}$$

With this expression, it's clearer how we'll extract/generate the moments of X — successive term by term differentiation. Can we do so? We've shown that $M_X(t)$ coincides with a power series on $(-\delta, \delta)$. A power series that converges can be infinitely differentiated term-by-term within its radius of convergence.

Therefore, we can differentiate term-by-term to obtain:

$$\begin{aligned}\frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t) &= \frac{\mathrm{d}^k}{\mathrm{d}t^k} \sum_{n=0}^{\infty} \frac{\mathbb{E}(X^n)t^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{\mathrm{d}^k}{\mathrm{d}t^k} \frac{\mathbb{E}(X^n)t^n}{n!} \\ &= \sum_{n=0}^{\infty} \mathbb{E}(X^n) \frac{n \cdot (n-1) \cdot \ldots \cdot (n-(k-1))t^{n-k}}{n!} \\ &= \sum_{n=0}^{\infty} \mathbb{E}(X^n) \frac{t^{n-k}}{(n-k)!}.\end{aligned}$$

At $t = 0$, we clearly recover the k^{th} term of the series $\mathbb{E}\big(X^k\big)$ which is precisely the k^{th} moment of X, μ'_k.

Therefore, the k^{th} moment of X is the coefficient of $t^k/k!$ in the series expansion of $M_X(t)$.

11.3 Alternative Derivation

Another method to demonstrate that $M_X(t)$ can be differentiated term-by-term relies on the following theorem:

Theorem 11.3.1 (Term-by-Term Differentiation) Let $\{f_n\}_{n\geqslant 1} \subseteq \mathcal{C}^1([a,b];\mathbb{R})$ and suppose that

- $S_n(x_0) = \sum_{i=1}^n f_i(x_0)$ converges pointwise as $n \to \infty$
- The series of derivatives $S'_n(x) = \sum_{i=1}^n f'_i(x)$ converges uniformly on $[a,b]$ as $n \to \infty$.

Then the series $S_n(x)$ converges uniformly on $[a,b]$ to some function $S(x)$ and

$$\frac{\mathrm{d}}{\mathrm{d}x} \sum_{n=1}^{\infty} f_n(x) = \sum_{n=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}x} f_n(x).$$

Now we must figure out whether $f_n(t) = \mathbb{E}(X^n)t^n/n!$ satisfies the above assumptions. The series $\sum_{i=0}^n f_i(t)$ clearly converges pointwise to $M_X(t)$ for all $t \in (t_n, t_p) \supseteq (-\delta, \delta)$. Does the series of derivatives converge uniformly on any compact subset of $(-\delta, \delta)$? To figure this out, we can use the Weierstrass M-test:

Theorem 11.3.2 (Weierstrass M-Test) For a sequence of functions $\{f_n\}_{n\geqslant 1}$ on $A \subseteq \mathbb{R}$, if $\exists \{M_k\}_{k\geqslant 1} \subseteq \mathbb{R}$ such that $\forall x \in A\colon |f_n(x)| < M_n$ and $\sum_{k=1}^{\infty} M_k$ converges, then $\sum_{n=1}^{\infty} f_n$ converges uniformly.

In our case, we can note that for any compact subset $K \subseteq (-\delta, \delta)$, we have that $|t| \leqslant \max_{t\in K} |t| =: U$ and so we can bound above:

$$|f_n(t)| = \left| \frac{\mathbb{E}(X^n t^n)}{n!} \right| \leqslant \frac{U^n}{n!} \mathbb{E}(|X|^n) =: M_n.$$

Since the MGF exists in $(-\delta, \delta)$, the absolute moments of X are all finite. I want to appeal to some fact about X that limits the growth of $\mathbb{E}(|X^n|)$ so the $M_n \overset{n\uparrow\infty}{\longrightarrow} 0$ and I can deduce that $\sum_{n=0}^{\infty} M_n$ converges.

CHAPTER 12

Multivariable Distributions

12.1 Multinomial Distribution

Skipped.

12.2 Bivariable Normal Distribution

Also skipped. I will come back to these.

CHAPTER 13

Population and Sampling

Definition 13.0.1

- A **sampling unit** is an entity to be selected by a sampling procedure.
- A **sampling procedure** is a method by which one selects a sample of units from a population with the purpose of increasing representativeness of the sample.

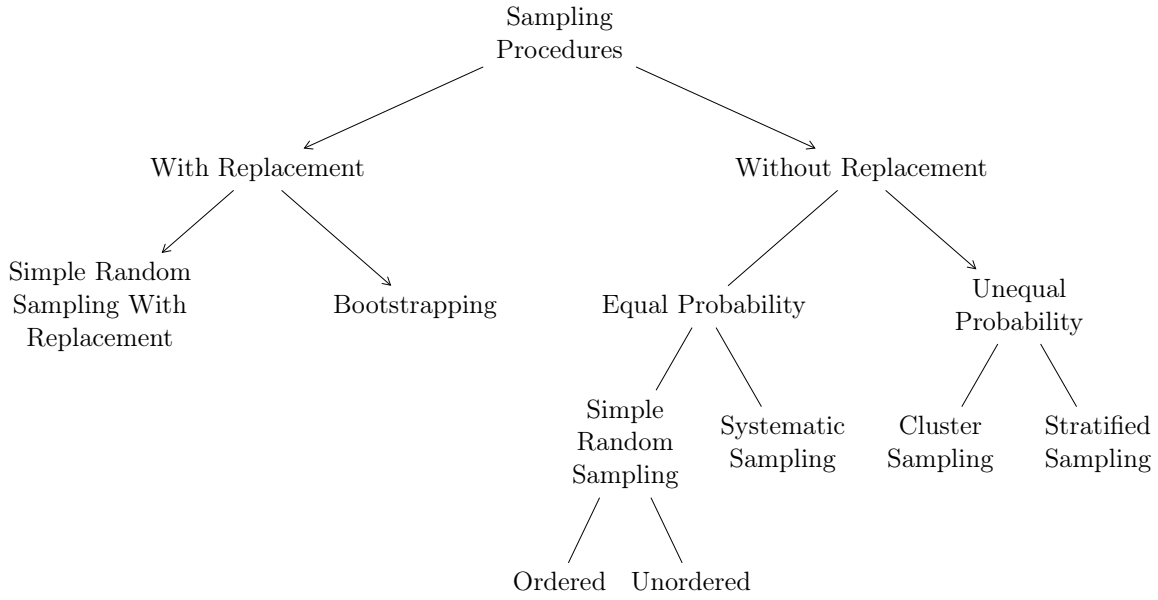


Figure 13.1: Some examples of sampling procedures, categorised by whether units are replaced or not.

Our sampling procedure affects everything downstream. The observed results from an experiment can carry two types of error:

- **Bias** is the effect of depriving a sample its representativeness of a population. This manifests by introducing a **systematic**[1] **error** in the sampling procedure. Such procedures are called *biased* sampling procedures. Systematic errors like favouring certain elements in the population, or taking items from a wrong population yield **biased samples**.
- All error that isn't systematic is designated as **random error**. Such errors may distort any one observation at any given point in the sampling procedure but this typically "balances out on average."

13.1 What Really Is A Population?

Many books disagree on (and occasionally don't define) what a 'population' is — this is not entirely unexpected but is incredibly annoying *within* the field of mathematical statistics. If different authors use the same word to represent different mathematical objects, then **'population' is clearly an overloaded term**. I want a definition that is both

[1] 'Systematic' means non-random in the setting of sampling processes.

- *unambiguous* (referring to a single type of mathematical object), and
- *compatible with general parlance* in academic and non-academic settings.

1. In general parlance, people refer to a population as a collection representing the totality of (similar) objects with which we're concerned.

   Thus, it follows that a sample is defined[2] as a proper subset of the population Π.

2. Some mathematical statistics textbooks, like Mathematical Statistics by Shao, refer to the population as the probability measure of a probability space.

> In statistical inference and decision theory, the data set is viewed as a realization or observation of a random element defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ related to the random experiment. The probability measure $\mathbb{P}$ is called the population. The data set or the random element that produces the data is called a sample from $\mathbb{P}$.
>
> [10, pp. 91–92]

> I believe this *random element* $\mathbf{X}$ represents the result of observing a collection of elements from a population — we call $\mathbf{X}$ a sample. The term random *element* encompasses both samples of size $n > 1$, and a sample of size 1; the former being a random vector, the latter a random variable.

Are those two definitions of 'population' compatible? The term 'sample' is also somewhat overloaded. In the second definition, a sample is a random element (but it's also permissible to use the term sample for the set, or collection, of realisations of the random variables). The population in the second definition is a probability measure but for it to be consistent with the first definition (our general usage of the term), it would also need to be a super**set** of a sample.

My Intent

I want to keep the first definition intact (a population is a set) while living with the duality of sample referring to a random element or its realisation — context will dictate which 'sample' is meant. I think this is reasonable.

The following two definitions best reflect what I wish for the word population to represent. The latter emphasises a real population and the former is an umbrella term for either a real or conceptual population.

[2] This is the definition I encountered at the beginning of my studies/these notes.

> A (real or hypothetical) totality of objects or individuals under consideration, of which the statistical attributes may be estimated by the study of a sample or samples drawn from it.
>
> OED 2010

> [...] population, which, understood physically, consists of individuals with various observable quantities.
>
> [11, p. 3]

The above definitions reflect what I believe best fits the word 'population' in the context of statistics, and I will denote a population by Π. Say we flip a coin 10 times in an experiment. The population, in the above sense, is the set $\{H, T\}$ from which we sample with replacement.

13.2 Inadequacies of a Single Space

Earlier commentary begs the obvious question; why are we doing something new? Why does the way we've defined probability spaces, and random variables on them fall short (or perhaps need re-working) when it comes to formalising sampling?

Remarks 13.2.1 Let's consider formalising the random experiment of flipping a fair coin $n > 1$ times.

- Let's begin with $\Omega = \{H, T\}$, $\mathcal{F} = 2^\Omega$, and define $\mathbb{P}$ by $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2$. Then we may define the i^{th} coin flip by the random variable $X_i(\omega) = \mathbb{1}_{\{\omega=H\}}(\omega)$, and so $X_i \sim \text{Bern}(1/2)$. These flips do not affect one another so they should be independent. How do we express this with only the information the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ offers us? Trick question — this probability space doesn't have enough information to support a description of how multiple variables interact.
- Instead, we must extend our definition of the outcome space to $\tilde{\Omega} := \Omega^n = \{H, T\}^n$, and define $\tilde{\mathbb{P}}$ to be uniform on $\tilde{\Omega}$. If we define each $X_i \colon \tilde{\Omega} = \{H, T\}^n \to \{0, 1\}$ by $X_i(\omega_1, \ldots, \omega_n) = \mathbb{1}_{\{\omega_i=H\}}(\omega_1, \ldots, \omega_n)$, then the mutual $\tilde{\mathbb{P}}$-independence of the X_i presents itself through the factoring of
$$\tilde{\mathbb{P}} = \otimes_n \mathbb{P}.$$

The above remarks inform the following definition:

Definition 13.2.2 A **random sample** of size n is a collection of random variables
$$X_i \colon (\Omega, \mathcal{F}) \to (E, \mathcal{E})$$
defined on the same probability space $(\Omega, \mathcal{F}, \tilde{\mathbb{P}})$ that are mutually $\tilde{\mathbb{P}}$-independent, and identically distributed with distribution $\mathbb{P}$ s.t. $\tilde{\mathbb{P}} = \otimes_n \mathbb{P}$. We denote this by
$$X_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}.$$

Equivalence of Absolute Continuity of Random Vector and Its Independent Components

I include in this section, the proof of which may be skipped, an important fact that is often taken for granted when discussing a random sample $\mathbf{X}$, and permits us to speak of the (joint) density of $\mathbf{X}$.

Theorem 13.2.3 For a collection of mutually independent random variables $X_1, \ldots, X_n$, let $\mathbf{X} = (X_1, \ldots, X_n)$. Then

$$\mathbb{P}_{\mathbf{X}} \ll \lambda_{\mathbb{R}^n} \iff \forall i = 1, \ldots, n\colon\ \mathbb{P}_{X_i} \ll \lambda_{\mathbb{R}}.$$

Proof. Recall that $\lambda_{\mathbb{R}^n}$ is the completion $\overline{\otimes_n \lambda_{\mathbb{R}}}$.

$\Longrightarrow$ Fix $i \in \{1, \ldots, n\}$ and let $A \in \mathcal{B}_{\mathbb{R}}$ be s.t. $\lambda_{\mathbb{R}}(A) = 0$. Assume that $\mathbb{P}_{\mathbf{X}} \ll \lambda_{\mathbb{R}^n}$. We wish to show that $\mathbb{P}_{X_i}(A) = 0$. Note that

$$\mathbb{P}_{X_i}(A) = \mathbb{P}_{\mathbf{X}}\big(\pi_i^{-1}(A)\big) = \mathbb{P}_{\mathbf{X}}(\mathbb{R} \times \ldots \times \mathbb{R} \times A \times \mathbb{R} \times \ldots \times \mathbb{R})$$

and that $\lambda_{\mathbb{R}^n}$ coincides with the product measure $\otimes_n \lambda_{\mathbb{R}}$ on $\mathcal{B}_{\mathbb{R}^n}$ so

$$\begin{aligned}
\lambda_{\mathbb{R}^n}(\pi_i^{-1}(A)) &= (\overline{\otimes_n \lambda_{\mathbb{R}}})(\pi_i^{-1}(A)) \\
&= (\otimes_n \lambda_{\mathbb{R}})(\pi_i^{-1}(A)) \quad \because \text{ they coincide on } \mathcal{B}_{\mathbb{R}^n} \\
&= \lambda_{\mathbb{R}}(A) \cdot \prod_{\substack{1 \leqslant j \leqslant n \\ j \neq i}} \lambda_{\mathbb{R}}(\mathbb{R}) = 0.
\end{aligned}$$

Since $\mathbb{P}_{\mathbf{X}} \ll \lambda_{\mathbb{R}^n}$, it follows that $0 = \mathbb{P}_{\mathbf{X}}\big(\pi_i^{-1}(A)\big)$. Thus, $\mathbb{P}_{X_i}(A) = 0$ and so $\mathbb{P}_{X_i} \ll \lambda_{\mathbb{R}}$.

$\Longleftarrow$ For the reverse implication, suppose that $\forall i \in \{1, \ldots, n\}\colon\ \mathbb{P}_{X_i} \ll \lambda_{\mathbb{R}}$. We wish to prove that

$$\mathbb{P}_{\mathbf{X}} \ll \lambda_{\mathbb{R}^n}.$$

Two intermediate factoids will help us reach the above conclusion.

Lemma 13.2.4 For $i = 1, \ldots, n$, let μ_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$, and let ν_i be a measure on the same measurable space. If for each i, $\nu_i \ll \mu_i$, then

$$\bigotimes_{i=1}^{n} \nu_i \ll \bigotimes_{i=1}^{n} \mu_i.$$

Proof. I need to think about this...  ■

Lemma 13.2.5 Let μ be a measure on $(\Omega, \mathcal{F})$ and denote by $\overline{\mu}$ its completion defined on $\overline{\mathcal{F}}$. Suppose that ν is another measure on $(\Omega, \mathcal{F})$ s.t. $\nu \ll \mu$. Then $\nu \ll \overline{\mu}$.

Proof. Let $B \in \overline{\mathcal{F}}$ s.t. $\overline{\mu}(B) = 0$.

- Since $B \in \overline{F}$, there exist $A \in \mathcal{F}$ and $F \in \mathcal{N}_\mu$ (i.e. $\exists N \in \mathcal{F}$ s.t. $F \subseteq N$ and $\mu(N) = 0$) s.t. $B = (A \cup F) \subseteq (A \cup N)$.
- Note that $0 = \overline{\mu}(B) = \overline{\mu}(A \cup F) \coloneqq \mu(A)$.

Now note that $\nu(B) = \nu(A \cup F) \leqslant \nu(A \cup N) \leqslant \nu(A) + \nu(N)$. Since $\nu \ll \mu$, and $0 = \mu(A) = \mu(N)$, it follows that $\nu(B) = 0$. Thus, $\nu \ll \overline{\mu}$. ■

We can now see that

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}} &= \bigotimes_{i=1}^{n} \mathbb{P}_{X_i} \quad \text{by independence} \\
&\ll \otimes_n \lambda_{\mathbb{R}} \quad \text{by } \textbf{Lemma 13.2.4}
\end{aligned}$$

and conclude from **Lemma 13.2.5** that

$$\mathbb{P}_{\mathbf{X}} \ll \overline{\otimes_n \lambda_{\mathbb{R}}} =: \lambda_{\mathbb{R}^n}.$$

■

In a sampling experiment, we can interpret the law/distribution of population elements deterministically (e.g. a real, tangible population of people at a fixed point in time is not random — the people exist), and we can attribute all the randomness to the act itself of sampling from said population. Thus, it seems sensible to have some kind of framework that supports these two different probability measures. This is something that is **not** done by the example above with $\Omega = \{H, T\}^n$. Indeed, that example conflates the randomness of the sampling experiment with the distribution of the "population" by encoding both into the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Below are two examples of sampling frameworks that highlight the separation of experimental randomness from population law:

Example 13.2.6 This example really made it click for me. Say that a village is running a local election. We're in the year 1695, and we wish to conduct an experiment to sample a voter from a population Π of voters in the village, each identified by a unique ID number. The method to select this voter is to fill a massive beer barrel with identical balls (labelled with each voter ID) and some mechanism to randomly eject a ball. This mechanism isn't influenced at all by the distribution of voters in the population. Instead, it works entirely according to the random mechanism and this randomness may be modelled abstractly as its own probability space.

Example 13.2.7 It's often the case that sampling from a particular distribution is done by computer — one employs an algorithm to generate pseudo-uniformly random numbers from [0,1], and then a deterministic map S transforms[3] these numbers into a sample $x_1, \ldots, x_n$. Such algorithms/sampling methods are not truly random but we can use these numbers x_i to make decisions about the law of the population.

13.3 The Two-Space Framework

The following framework for sampling, introduced by Yiping Cheng [11], encapsulates all of the points discussed so far. We opt for two probability spaces:

- $(\Pi, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$ is called the **population probability space**. The non-deterministic true distribution of the population elements is encoded by $\mathbb{P}_\Pi$.
- $(\Lambda, \mathcal{F}_\Lambda, \mathbb{P}_\Lambda)$ is called the **experiment probability space**. The randomness of the sampling procedure is captured by $\mathbb{P}_\Lambda$.

The following list summarises the framework:

-1. The population probability space is fixed — it is what it is in nature.

0. We begin with some inferential goal (to estimate a mean, test a hypothesis etc.)

1. We choose a sampling procedure (e.g. random sampling, or systematic sampling) that's appropriate for said goal.
    - The output of this procedure is a tuple of length n from the set of all possible tuples of Π^n — an element of Π^n is called a **sample**.
    - If we could apply this sampling procedure perfectly to the full population, then the arising distribution of tuples $\mathbb{P}_{\Pi^n}$ is called the **intended sampling distribution**. In truth, $\mathbb{P}_{\Pi^n}$ is a theoretical artefact that serves as the goal of this entire process.
        - We denote by $\mathcal{F}_{\Pi^n}$ the corresponding σ-algebra over which $\mathbb{P}_{\Pi^n}$ is defined, and call $(\Pi^n, \mathcal{F}_{\Pi^n}, \mathbb{P}_{\Pi^n})$ the **sampling probability space**.

2. We typically have the form of $\mathbb{P}_{\Pi^n}$ in mind e.g. if we're performing random sampling, then $\mathbb{P}_{\Pi^n} = \otimes_n \mathbb{P}_\Pi$ despite not knowing $\mathbb{P}_\Pi$.

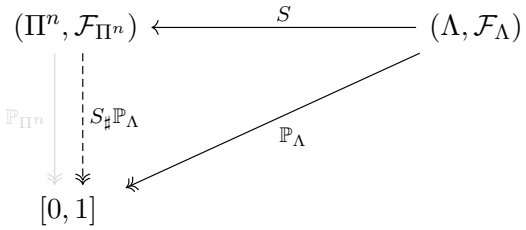
[3] In the language of distributions, S pushes forward this pseudo-uniform randomness to a distribution of our choosing. In the idealised world of mathematics, the existence of a canonical source of randomness is a deep fact, and one that will be mentioned in this section.

3. The experiment probability space is set up to separate the randomness in the experimental procedure of sampling, from the deterministic mappings that shape said randomness into complex distributions that model our population.

    - The fundamental bridge between the randomness of the experiment and the intended distribution $\mathbb{P}_{\Pi^n}$ of the samples is a **deterministic** $(\mathcal{F}_\Lambda, \mathcal{F}_{\Pi^n})$-measurable function, dubbed the **sampler mapping**

    $$S\colon \Lambda \to \Pi^n.$$

    The sampler mapping transforms each experimental outcome $e \in \Lambda$ into a population sample $S(e) \in \Pi^n$.

    The measurability of S is so that probabilities are well-defined when pushing $\mathbb{P}_\Lambda$ forward.

    - If all has been set up correctly, the sampler mapping will faithfully transform the randomness of sampling $\mathbb{P}_\Lambda$ into our intended population distribution $\mathbb{P}_{\Pi^n}$ on Π^n i.e.

    $$S_\sharp \mathbb{P}_\Lambda = \mathbb{P}_{\Pi^n}.$$

In summary, we hope that correctly constructing an appropriate sampling framework (which includes S, and $(\Lambda, \mathcal{F}_\Lambda, \mathbb{P}_\Lambda)$) will yield our intended population distribution $\mathbb{P}_{\Pi^n}$.

13.4 Simple Random Sampling With Replacement (SRSWR)

This subsection deals with incorporating the procedure of simple random sampling with replacement (from a finite population Π) into the two-space framework.

Simple random sampling with replacement is a sampling procedure in which one successively selects n units from a finite population, after each draw returning the selected unit to the population, in such a way that that each sample of size n has equal selection probability. Since we replace each drawn unit, the population remains the same and so we can regard each individual as independent of any other, and that each draw is governed by the same population law.

Other books (e.g. [4]) define simple random sampling with replacement as a sampling procedure in which every member of the population has an equal chance of being chosen and successive drawings are independent, as, for example, in sampling with replacement. This is a consequence of the above if we let $n = 1$, and so we may assume the population law is uniform.

To incorporate this into the two-space framework, our assumptions tell us that the intended sampling distribution $\mathbb{P}_{\Pi^n}$ is equal to the product measure $\otimes_n \mathbb{P}_\Pi$ which itself is a uniform measure on Π^n. A natural choice for $\mathcal{F}_{\Pi^n}$ is the product σ-algebra $\otimes_n \mathcal{F}_\Pi$ which is the smallest[4] σ-algebra s.t. the coordinate maps S_i are measurable.

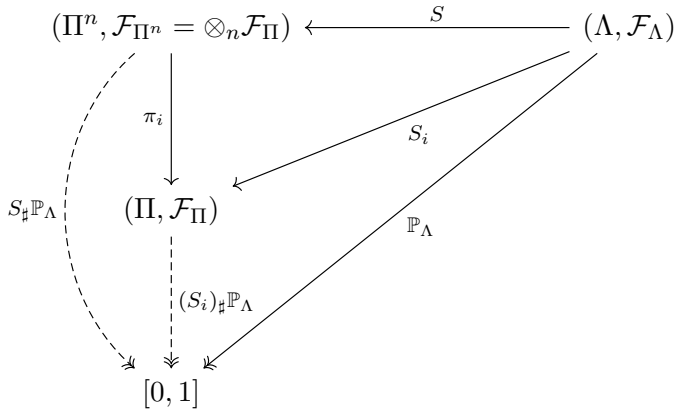
> S is a map from $\Lambda \to \Pi^n$ and so it is of the form $S = (S_1, \ldots, S_n)$. Furthermore, the assumed $(\mathcal{F}_\Lambda, \mathcal{F}_{\Pi^n} = \otimes_n \mathcal{F}_\Pi)$-measurabilty of S implies, by satisfying the conditions of **Proposition 6.0.3** (Proposition 2.4 [8]), that each coordinate map $S_i\colon \Lambda \to \Pi$ is $(\mathcal{F}_\Lambda, \mathcal{F}_\Pi)$-measurable.

These assumptions of simple random sampling with replacement force our framework to obey the equality

$$S_\sharp \mathbb{P}_\Lambda = \bigotimes_1^n \mathbb{P}_\Pi.$$

[4] For any σ-algebra $\mathcal{F}_{\Pi^n}$ smaller than $\otimes_n \mathcal{F}_\Pi$, we don't have enough information required to speak of the coordinates of S as random variables.

Visually, we may represent the maps by:



Now we can make some comments about the S_i:

- Let $A \in \mathcal{F}_\Pi$.

$$\begin{aligned}
((S_i)_\sharp \mathbb{P}_\Lambda)(A) &= \mathbb{P}_\Lambda\big(S_i^{-1}(A)\big) \\
&= \mathbb{P}_\Lambda\big((\pi_i \circ S)^{-1}(A)\big) \\
&= \mathbb{P}_\Lambda\big(S^{-1}(\pi_i^{-1}(A))\big) \\
&= (S_\sharp \mathbb{P}_\Lambda)(\pi_i^{-1}(A)) \\
&= (\otimes_n \mathbb{P}_\Pi)(\pi_i^{-1}(A)) \quad \text{by assumption} \\
&= (\otimes_n \mathbb{P}_\Pi)(\underbrace{\Pi \times \ldots \times \Pi}_{(i-1) \text{ times}} \times A \times \Pi \times \ldots \times \Pi) \\
&= \mathbb{P}_\Pi(\Pi) \cdot \ldots \cdot \mathbb{P}_\Pi(\Pi) \cdot \mathbb{P}_\Pi(A) \cdot \mathbb{P}_\Pi(\Pi) \cdot \ldots \cdot \mathbb{P}_\Pi(\Pi) \\
&= \mathbb{P}_\Pi(A)
\end{aligned}$$

  Therefore, the S_i are identically distributed with distribution $\mathbb{P}_\Pi$.

- It also follows immediately that the S_i are mutually $\mathbb{P}_\Lambda$-independent because for any collection $\{A_i\}_{i=1}^n \subseteq \mathcal{F}_\Pi$:

$$\begin{aligned}
\mathbb{P}_\Lambda(S_1 \in A_1, \ldots, S_n \in A_n) = (S_\sharp \mathbb{P}_\Lambda)(A_1 \times \ldots \times A_n) &= (\otimes_n \mathbb{P}_\Pi)(A_1 \times \ldots \times A_n) \\
&= \prod_{i=1}^n \mathbb{P}_\Pi(A_i) \\
&= \prod_{i=1}^n ((S_i)_\sharp \mathbb{P}_\Lambda)(A_i) \\
&= \prod_{i=1}^n \mathbb{P}_\Lambda(S_i \in A_i).
\end{aligned}$$

Thus, the S_i constitute a random sample in the traditional sense of **Definition 13.2.2**. Furthermore, we note that the condition $S_\sharp \mathbb{P}_\Lambda = \otimes_n \mathbb{P}_\Pi$ is the two-space framework equivalent of $\underset{\text{i.i.d.}}{\sim} \mathbb{P}_\Pi$ and we give it a name:

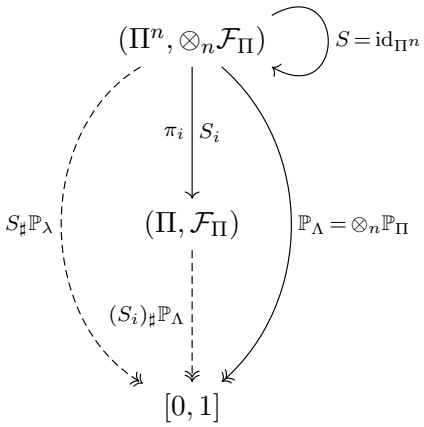
Definition 13.4.1 The sampler mapping S is **simple** if for every collection $\{A_j\}_{j=1}^n \subseteq \mathcal{F}_\Pi$, we have that

$$(S_\sharp \mathbb{P}_\Lambda)(A_1 \times \ldots A_n) = \mathbb{P}_\Lambda(\{e \in \Lambda \colon S_1(e_1) \in A_1, \ldots, S_n(e_n) \in A_n\}) = \prod_{j=1}^n \mathbb{P}_\Pi(A_j).$$

It's clear from this definition that simplicity of a sampler mapping depends on the experiment probability space $(\Lambda, \mathcal{F}_\Lambda, \mathbb{P}_\Lambda)$ — something we have so far taken for granted exists and represents the randomness of the sampling procedure. The reason for this is slightly underhanded; there is one choice trivial of experiment probability space that always works, and in this particular example for which a simple sampler mapping always exists:

Trivial Example ($S = \mathrm{id}$)

Let $(\Lambda = \Pi^n, \mathcal{F}_\Lambda = \otimes_n \mathcal{F}_\Pi, \mathbb{P}_\Lambda = \otimes_n \mathbb{P}_\Pi)$ and S be the identity mapping id_{Π^n}. The framework collapses to the familiar setting of a single probability space on which we have the random variables $S_i = \pi_i \circ \mathrm{id}_{\Pi^n} = \pi_i$ defined (and are equal to the canonical projections).



This is a special case of the general random sampling framework above so the S_i still constitute a random sample in the traditional sense. In particular, the coordinate maps are the projection maps $S_i = \pi_i$ and we can denote these by X_i.

13.5 Canonical Randomness

Before covering a non-trivial example that incorporates simple random sampling into the two-space framework, I think it's appropriate to put an important theorem here that powers the upcoming examples — any probability measure on a nice enough (standard Borel) space can be realised as the push-forward of a canonical source of randomness $\lambda_{[0,1)}$ defined on $([0,1), \mathcal{B}_{[0,1)})$ via some (appropriately) measurable map h.

More precisely:[5]

Theorem 13.5.1 (Theorem 5.4 [12]) Let $(E, \mathcal{E})$ be a Borel space.[a] Let μ be a σ-finite measure on $(E, \mathcal{E})$ and put $b = \mu(E)$, possibly $+\infty$. Then, there exists a mapping h from $[0, b)$ into E, measurable relative to $\mathcal{B}_{[0,b)}$ and $\mathcal{E}$ such that

$$\mu = h_\sharp \lambda$$

where λ is the Lebesgue measure on $[0, b)$.

[a] Çinlar calls this a standard measurable space.

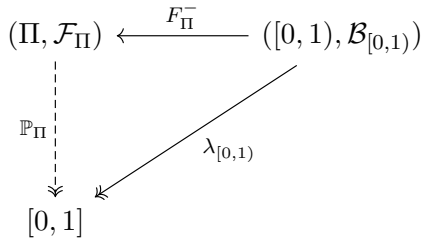
For our purposes:

- $(E, \mathcal{E})$ is $(\Pi, \mathcal{F}_\Pi)$ i.e. we assume our population probability space is a Borel space,
- μ is $\mathbb{P}_\Pi$, so $\mathbb{P}_\Pi(\Pi) = 1$ and so $b = 1$.

We denote h by F_Π^- and call this map the **generalised quantile function**[6] **of** $\mathbb{P}_\Pi$ that pushes forward λ to $\mathbb{P}_\Pi$ i.e. $\mathbb{P}_\Pi = (F_\Pi^-)_\sharp(\lambda)$.

[5] Çinlar's book instead uses the notation $\lambda \circ h^{-1}$ for the pushforward of λ via h. I passionately disagree with writing h^{-1} because it's too suggestive for the inverse of h, especially when it's generally not the case that h is invertible.

[6] Or **inverse probability transform**.



Total speculation on my part but thinking about the real case tells me there should probably be a theorem that states conditions under which we can upgrade the representation theorem above to an isomorphism i.e. there exists a measurable function $g\colon E \to [0,1]$ s.t. $g_\sharp(\mu) = \lambda$, and one has that g and h are "inverses" of each other in some sense. I'd call this map g the ***forward* probability integral transform of** $\mathbb{P}_\Pi$, and denote it by F_Π. Thus, F_Π would push $\mathbb{P}_\Pi$ forward to λ i.e.

$$(F_\Pi)_\sharp \mathbb{P}_\Pi = \lambda_{[0,1)}.$$

13.5.1 Constructive Proof For h (Π at most countable)

In the case that Π is at most countably infinite, one can explicitly construct a h in the statement of **Theorem 13.5.1**. We do so by partitioning $[0,1)$ into intervals with Lebesgue measure corresponding to appropriate probabilities.

First, observe that $(\Pi, \mathcal{F}_\Pi)$ is an at most countably infinite Borel space. By **Theorem 4.4.5** (the measurable classification theorem), it's trivial in the sense that $\mathcal{F}_\Pi = 2^\Pi$. Furthermore, it inherits an ordering from some subset of $\mathbb{R}$ and so we may write

$$\Pi = \{x_1, x_2, \ldots\}.$$

We call x_i the **i^{th} population unit**.

At the end of this, we want a map h that satisfies

$$\mathbb{P}_\Pi(\{x_i\}) = (h_\sharp \lambda_{[0,1)})(\{x_i\}) = \lambda_{[0,1)}(h^{-1}(\{x_i\}))$$

so it seems sensible that we wish for h to map an interval of length $\mathbb{P}_\Pi(\{x_i\})$ to each x_i. Let $s_0 = 0$ and for $i \in \mathbb{N}$:

$$s_i = \sum_{j=1}^{n} \mathbb{P}_\Pi(\{x_i\}).$$

Note that $0 = s_0 \leqslant s_1 \leqslant s_2 \leqslant \ldots$ and $s_i \to 1$ defines a partition $I_i = \{[s_{i-1}, s_i)\}_{i \in \mathbb{N}}$ of $[0,1)$. Now we define the map $h\colon [0,1] \to \Pi$ by $h(e) = x_i$ for all $e \in I_i$, and define $h(1) = x_k$ for any k (doesn't matter which because the singleton $\{1\}$ contributes zero measure).

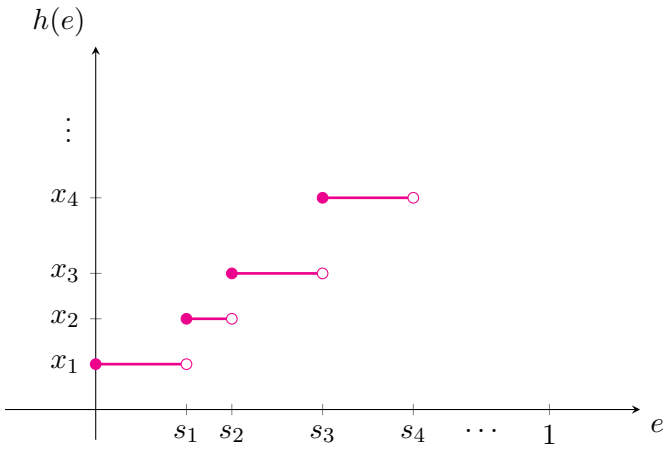


Figure 13.2: A plot of an example h for a particular (non-uniform) discrete probability measure $\mathbb{P}_\Pi$. Note that the partition widths sum to unity.

Now let $A \in \mathcal{F}_\Pi$ i.e. let A be any subset of Π so it looks like

$$A = \bigsqcup_{i\colon x_i \in A} \{x_i\}.$$

Let J_A denote the indexing set $\{i\colon x_i \in A\}$. Then we may write

$$\begin{aligned}
(h_\sharp \lambda_{[0,1)})(A) = \lambda_{[0,1)}(h^{-1}(A)) = \lambda_{[0,1)}\Big(\bigsqcup_{i\in J_A} I_i\Big) &= \sum_{i\in J_A} \lambda_{[0,1)}(I_i) \\
&= \sum_{i\in J_A} (s_i - s_{i-1}) \\
&= \sum_{i\in J_A} \mathbb{P}_\Pi(\{x_i\}) \\
&= \mathbb{P}_\Pi\Big(\bigsqcup_{i\in J_A} \{x_i\}\Big) \quad \text{by } \sigma\text{-additivity} \\
&= \mathbb{P}_\Pi(A)
\end{aligned}$$

13.5.2 Non-trivial Example (SRSWR)

Now we're in a position to apply the concept of canonical randomness to SRSWR in the two-space framework.

We'll define Λ as the n-fold product of $[0,1]$ and endow it with a probability distribution $\mathbb{P}_\Lambda$ that mimics a uniform distribution on the components of Λ. Then we'll use the generalised quantile function h from the at-most countably infinite case to map into the population probability space. Note that the same construction for h works if the population is only finite, and the setup of SRSWR only works for a finite population for there is no uniform probability measure on an infinite population (set)!

- The population probability space is as before $(\Pi, \mathcal{F}_\Pi, \mathbb{P}_\Pi)$ where Π is the finite population from which we'll draw our samples.
- $\Lambda = [0,1)^n$.
- Let $\mathcal{F}_\Lambda = \mathcal{B}_{[0,1)^n} = \otimes_n \mathcal{B}_{[0,1]}$ and $\mathbb{P}_\Lambda = \lambda_{[0,1)^n} = \overline{\otimes_n \lambda_{[0,1)}}$. This choice of experiment probability space tells us that the distribution of each component of a tuple $(e_1, \ldots, e_n) = e \in [0,1)^n$ in Λ can be considered to be uniformly distributed[7] on $[0,1]$.

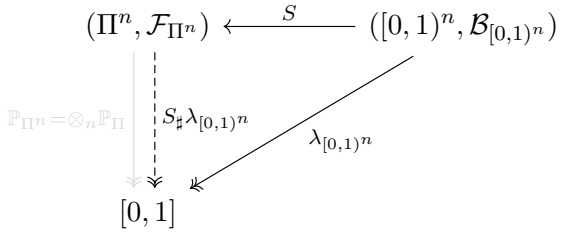
[7] Perhaps I've not yet explicitly written the idea that the Lebesgue measure on $[0,1]$ corresponds to the probability

- We define the sampler mapping $S\colon \Lambda \to \Pi^n$ by

$$S(e) = (h(e_1), \ldots, h(e_n)),$$

where $h\colon [0,1] \to \Pi$ is defined as before by $h(e) = x_i$ for all $e \in I_i$ where $I_i = [s_{i-1}, s_i)$ is an interval of Lebesgue measure $\mathbb{P}_\Pi(\{x_i\})$, and the $s_i = \sum_{j=1}^{n} \mathbb{P}_\Pi(\{x_i\})$ are the endpoints of a partition of $[0,1)$ with $s_0 = 0$ and $s_i \to 1$.

The diagram for this setup is:



It's sufficient to compute the induced probability measure $S_\sharp \mathbb{P}_\Lambda$ for every measurable rectangle $A \in \mathcal{F}_{\Pi^n} = \otimes_n \mathcal{F}_\Pi$ i.e. for every set of the form $A = A_1 \times \ldots \times A_n$ with $A_i \in \mathcal{F}_\Pi$ for every i.

$$\begin{aligned}
(S_\# \mathbb{P}_\Lambda)(A) = \mathbb{P}_\Lambda\big(S^{-1}(A)\big) &= \mathbb{P}_\Lambda\big(S^{-1}(A_1 \times \ldots \times A_n)\big) \\
&= \lambda_{[0,1)^n}(S^{-1}(A_1 \times \ldots \times A_n)) \\
&= \lambda_{[0,1)^n}\left((h^{-1}(A_1) \times \ldots \times h^{-1}(A_n)\right) \\
&= \prod_{i=1}^{n} \lambda_{[0,1)}(h^{-1}(A_i)) \\
&= \prod_{i=1}^{n} \mathbb{P}_\Pi(A_i) \quad \text{since } \mathbb{P}_\Pi = h_\sharp \lambda_{[0,1)} \\
&= \Big(\bigotimes_1^n \mathbb{P}_\Pi\Big)(A_1 \times \ldots \times A_n) \\
&= \Big(\bigotimes_1^n \mathbb{P}_\Pi\Big)(A) \\
&= \mathbb{P}_{\Pi^n}(A)
\end{aligned}$$

which is equal to our intended probability distribution.

Remarks The theorem in this section on a canonical source of randomness finds a natural application in how sampling is typically implemented in computer algorithms. Random number generators produce numbers (approximately) uniformly between 0 and 1, and then we can obtain a sample from any other distribution via a deterministic transformation (like our sampler mapping S).

13.6 Simple Random Sampling

Simple random sampling differs only from SRSWR only in the sense that we do not replace each unit selected after each draw. Thus, the underlying population is uniformly distributed. One can view the same procedure from two perspectives — keep track of the ordered tuples of units, or simply view the process as selecting some unordered subset of the population. It's the same

distribution of a uniformly distributed random variable on $[0,1]$ but it should be clear that for any $X \sim \text{Unif}([a,b])$, any subset $A \subseteq [a,b]$ has probability

$$\mathbb{P}_X(A) = \int_A \frac{1}{b-a}\,\mathrm{d}x = \frac{1}{b-a}\lambda(A).$$

Then we simply note that if $U \sim \text{Unif}([0,1])$, then $b - a = 1$ so $\mathbb{P}_U(A) = \lambda(A)$.

procedure either way; just the book-keeping that's different, and the former allows us to speak meaningfully of the i^{th} draw.

However, one important to point to keep in mind is that simple random sampling (without replacement) is materially different between "small" and "large" populations. If Π is a population that is significantly larger than the sample size n, removing any single observation from Π on any given draw makes little meaningful impact on the distribution of elements in Π. Thus, for a large enough population, there isn't very much of a distinction between simple random sampling without and with replacement.

In order to make some meaningful comments on each draw of a simple random sampling procedure (without replacement), I'll work with ordered tuples so let

$$\Lambda = \{(i_1, \dots, i_n) \in \{1, \dots, N\}^n \colon i_j\text{-distinct}\}.$$

Note that the number of such ordered samples is equal to $N(N-1)\dots(N-(n-1))$, and so $\mathbb{P}_\Lambda$ is the uniform probability measure on $\mathcal{F}_\Lambda = 2^\Lambda$ is defined by[8]

$$\mathbb{P}_\Lambda(\{(i_1, \dots, i_n)\}) = \frac{1}{\operatorname{card}(\Lambda)} = \frac{1}{\frac{N!}{(N-n)!}}.$$

The intended distribution $\mathbb{P}_{\Pi^n}$ is uniform by how simple random sampling is defined. **However**, the draws are not independent because we do not replace the units once drawn and we are not[9] considering a limiting case where the sample size is very small compared to $\operatorname{card}(\Pi)$.

Since our population is finite, $\mathcal{F}_{\Pi^n} = 2^{\Pi^n}$. The sampler mapping in this case $S \colon \Lambda \to \Pi^n$ is defined by

$$S(i_1, \dots, i_n) = (x_{i_1}, \dots, x_{i_n}).$$

and the pushforward is defined for any $\{(x_{i_1}, \dots, x_{i_n})\} \in \mathcal{F}_{\Pi^n}$ by

$$\begin{aligned}(S_\sharp \mathbb{P}_\Lambda)(\{(x_{i_1}, \dots, x_{i_n})\}) &= \mathbb{P}_\Lambda\big(S^{-1}(\{(x_{i_1}, \dots, x_{i_n})\})\big) \\ &= \mathbb{P}_\Lambda(\{(i_1, \dots, i_n)\}) \\ &= 1/\operatorname{card}(\Lambda)\end{aligned}$$

which is the uniform distribution on Π^n.

Now to discuss the j^{th} unit observed. Let π_j denote the natural projection onto the j^{th} coordinate. Then the j^{th} draw is $\pi_j \circ S =: S_j \colon \Lambda \to \Pi$ defined by

$$S_j(i_1, \dots, i_n) = x_{i_j}.$$

Definition 13.6.1 The **inclusion probability of the k^{th} unit** is the probability that x_k will be included in the sample realised by the sampling procedure.

The distribution of S_j is given by $((S_j)_\sharp \mathbb{P}_\Lambda)(\{x_k\})$ which looks to me like the k^{th} inclusion probability in the j^{th} slot. Fix $i_j \in \{1, \dots, n\}$ and a population unit x_k.

$$\begin{aligned}((S_j)_\sharp \mathbb{P}_\Lambda)(\{x_k\}) = \mathbb{P}_\Lambda\Big(S_j^{-1}(\{x_k\})\Big) &= \mathbb{P}_\Lambda(\{e \in \Lambda = \{i_1, \dots, i_n\} \colon S_j(e) = x_k\}) \\ &= \mathbb{P}_\Lambda(\{e \in \Lambda \colon i_j = k\})\end{aligned}$$

At this point, we count how many ordered outcomes in Λ put the index k (i.e. the k^{th} unit) in the j^{th} slot. With the j^{th} coordinate fixed, the remaining indices must be distinct and drawn without

[8] Had we considered unordered tuples i.e. subsets of Π size n, we could let $\Lambda = \{A \subseteq \Pi \colon \operatorname{card}(A) = n\}$, and then $\mathbb{P}_\Lambda$ would be defined by

$$\mathbb{P}_\Lambda(A) = \frac{1}{\#\text{ways to select } n \text{ elements from } N} = \frac{1}{\binom{N}{n}}.$$

[9] At least not yet.

replacement from the remaining $N-1$ indices i.e. there are $(N-1)(N-2)\dots((N-1)-(N-(n-1)))$ such ordered tuples which means that

$$\begin{aligned}((S_j)_\sharp\mathbb{P}_\Lambda)(\{x_k\}) = \dots &= \mathbb{P}_\Lambda(\{e \in \Lambda \colon i_j = k\}) \\ &= \frac{(N-1)\dots((N-1)-((n-1)-1))}{\mathrm{card}(\Lambda)} \\ &= \frac{(N-1)\dots(N-(n-1))}{N(N-1)\dots(N-(n-1))} = \frac{1}{N}\end{aligned}$$

i.e. that each draw is identically distributed.

Unlike SRSWR, the draws are not independent so the S_j do not constitute a random sample in the traditional sense of **Definition 13.2.2**.

CHAPTER 14

Functions of Random Variables

Henceforth, unless stated otherwise, $X_1, \ldots, X_n$ will denote a random sample.

14.1 Statistics, Estimators and Estimates

When a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ is drawn from a population, we can calculate some summary $T(x_1, \ldots, x_n)$ of the observed values $x_1, \ldots, x_n$.

Example 14.1.1 For an observed sample $x_1, \ldots, x_n$, the sample mean $\overline{\mu}$ is defined as

$$T(x_1, \ldots, x_n) = \overline{\mu} \coloneqq \frac{1}{n} \sum_{i=1}^{n} x_i.$$

We can view any such summary as a realisation of the function $T \circ \mathbf{X}$ of the random sample $\mathbf{X}$. If T is measurable in a way that's compatible with the measurability of $\mathbf{X}$ i.e. that $T \circ \mathbf{X}$ is measurable, then we call T a **statistic**.[1]

Definition 14.1.2 Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$, where $X_i \overset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, is a random sample, where θ is some fixed and unknown population parameter. Let T be a statistic. If the corresponding random element $T \circ \mathbf{X}$ is used to estimate a population parameter θ, then:

- we call $T \circ \mathbf{X}$ an **estimator**,
- and the observed value $\widehat{\theta} = T(x_1, \ldots, x_n)$ of such an estimator $T \circ \mathbf{X}$ based on an observed sample $x_1, \ldots, x_n$ is called an **estimate**.

Remarks 14.1.3

- We also require that T is not a function of any unknown parameters (including θ). Otherwise, we wouldn't be able to compute $T \circ \mathbf{X}$ since θ is unknown.
- The probability distribution $\mathbb{P}_{T \circ \mathbf{X}}$ of the corresponding random element $T \circ \mathbf{X}$ is called **the sampling distribution of $T \circ \mathbf{X}$**.

Example 14.1.4 Examples of $T \circ \mathbf{X}$.

- The sample mean $\overline{X}$ is the arithmetic average of the values in a random sample, denoted by

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}.$$

  We use the sample mean to estimate the mean $\theta = \mu$ of the population.

- The sample standard deviation S is the positive square root of the sample variance S^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

  We use it to estimate the population standard deviation.

[1] This will be made more precise later on:

- The general measure-theoretic definition is **Definition 16.2.1**.
- The statistics/estimation-geared definition (which is a specific case of the former) is **Definition 21.0.1**.

- $X_{(n)} = \max\{X_1, \ldots, X_n\}$
- $X_{(1)} = \min\{X_1, \ldots, X_n\}$
- The range $R \coloneqq X_{(n)} - X_{(1)}$

These definitions suppress the random sample e.g. $\overline{X}$ is truly $\overline{X}(X_1, \ldots, X_n)$. Observed values are lowercase.

14.1.1 How "Good" Is An Estimator?

I imagine that learning the distribution of an estimator will go some way to helping quantify how "good" an estimator is.

Consider the problem of estimating a population's mean μ. Intuitively, one draws a random sample of n observations $x_1, \ldots, x_n$ from the population and employs the sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

as an estimate for μ. How good is this estimate? The answer depends on the behaviour of the random sample $X_1, \ldots, X_n$ and its effect on the (sampling) distribution of $\overline{X}$.

- One measure of the goodness of the estimator $T \circ \mathbf{X} = \overline{X}$ is the error of its estimation — the difference between the estimate $T(x_1, \ldots, x_n)$ and the parameter being estimated (e.g. the difference between $\overline{x}$ and μ).
- Since $T \circ \mathbf{X}$ is a random variable, we can't assign a deterministic number to the error in estimation for each sample.
    - However, if we can determine the probability distribution $\mathbb{P}_{T\circ\mathbf{X}}$ of the estimator (e.g. $\mathbb{P}_{\overline{X}}$), then we can use it to bound the probability that the estimation error falls within some tolerance.

There are 3 main methods for finding the probability distribution of a function $T(X_1, \ldots, X_n)$ of random variables:

14.2 The 3 Methods

Each method works well for different examples. Consider random variables $Y_1, \ldots, Y_n$ and a function $U(Y_1, \ldots, Y_n)$ denoted simply by U. The 3 methods are summarised as follows:

1. The Method of Distribution Functions:
    - Conditions: Typically used when the $Y_1, \ldots, Y_n$ are jointly absolutely continuous random variables i.e. $\mathbf{Y} = (Y_1, \ldots, Y_n)$ admits a density.
    - Method: Find $F_U(u) = \mathbb{P}(\{U \leqslant u\})$ i.e. find the region in $y_1, \ldots, y_n$ space for which $U \leqslant u$ and then integrate over this region. Then differentiate $F_U(u)$ to find the density of U.
2. The Method of Transformations
    - Conditions: When given the density of a random variable Y, this method results in a general expression for the density of $U = h(Y)$ for some strictly increasing or strictly decreasing function of Y.
    - Method: If Y_1 and Y_2 have a joint distribution, we can use the univariate result to find the joint density of Y_1 and $U = h(Y_1, Y_2)$. By integrating over Y_1, we find the marginal density of U which is our objective.

3. The Method of Moment-Generating Functions

   - Conditions: Based on a uniqueness theorem: If two random variables have identical moment-generating functions, the two random variables possess the same probability distribution.
   - Method: Compute the moment-generating function $M_U(t)$ for $U = h(Y_1, \ldots, Y_n)$ and compare it against the moment-generating functions of common discrete or absolutely continuous random variables (from prior chapters).

These methods are computational, best illustrated with theory and examples.

14.3 Method of Distribution Functions

The following examples (minus the theorems) are from the lecture video series by Professor Unnikrishna Pillai which may be found **here on YouTube**. I'll continue to use my own notation in these notes, being as clear as I possibly can.

14.3.1 $Z = X + Y$

Suppose that $\mathbf{X} = (X, Y)$ admits a density $f_{\mathbf{X}}(x, y)$ and let $Z = X + Y$.

The general idea is to write the unknown quantity $F_Z(z)$ in terms of known quantities relating to the probability distribution of $\mathbf{X}$ e.g. $f_{X,Y}(x, y)$. First note that

$$\begin{aligned} F_Z(z) = \mathbb{P}(\{Z \leqslant z\}) &= \mathbb{P}(\{X + Y \leqslant z\}) \\ &= \mathbb{P}(\{\omega \in \Omega \colon X(\omega) + Y(\omega) \leqslant z\}) \\ &= \mathbb{P}\big(\mathbf{X}^{-1}(\{(x, y) \in \mathbb{R}^2 \colon x + y \leqslant z\})\big) \\ &= \mathbb{P}_{\mathbf{X}}\big(\{(x, y) \in \mathbb{R}^2 \colon x + y \leqslant z\}\big) \\ &= \iint_{\{(x,y) \in \mathbb{R}^2 \colon x+y \leqslant z\}} f_{X,Y}(x, y)\, \mathrm{d}A \end{aligned}$$

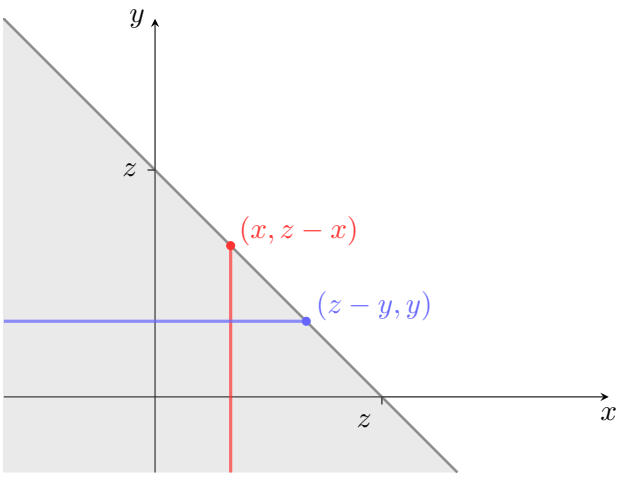


Figure 14.1: The region in the codomain $\mathbb{R}^2$ of $\mathbf{X} \colon \Omega \to \mathbb{R}^2$ over which we integrate the joint density $f_{\mathbf{X}}$ in order to find the probability of the event $\{X + Y \leqslant z\} := \{\omega \in \Omega \colon X(\omega) + Y(\omega) \leqslant z\} \subseteq \Omega$.

Now we integrate the joint density $f_{X,Y}$ over this region. One way to integrate over this region in $\mathbb{R}^2$ is:

- Fix x and consider a slice of the region of constant x. This variable slice is the integral of $f_{X,Y}$ over the possible y values (for each fixed x).

- Then integrate this variable slice over the possible x values.

- Fix y and consider a slice of the region of constant y.
  - Then integrate over the possible x values (for each fixed y).

Therefore, our cumulative distribution function can be computed as:

$$\begin{aligned} F_Z(z) = \mathbb{P}(\{X+Y \leqslant z\}) &= \iint_{\{(x,y)\in\mathbb{R}^2 \,:\, x+y\leqslant z\}} f_{X,Y}(x,y)\,\mathrm{d}A \\ &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,\mathrm{d}y\right)\mathrm{d}x = \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{z-y} f_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \end{aligned}$$

We must now differentiate $F_Z(z)$ with respect to z to find the density of $Z = X+Y$. This can be done with the Leibniz integral rule:

14.3.2 Leibniz's Integral Rule

Theorem 14.3.1 (Differentiation of the Integral Depending on a Parameter) Suppose that $f(x,t)$ and $\frac{\partial f}{\partial t}(x,t)$ are continuous on the rectangle $[a,b]\times(c,d)$. Then $\forall t \in (c,d)$

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b f(x,t)\,\mathrm{d}x = \int_a^b \frac{\partial f}{\partial t}(x,t)\,\mathrm{d}x$$

Proof. Both integrals in the theorem statement

$$F(t) \coloneqq \int_a^b f(x,t)\,\mathrm{d}x \quad \text{and} \quad G(t) = \int_a^b \frac{\partial f}{\partial t}(x,t)\,\mathrm{d}x$$

exist by the continuity of the respective integrands.

The goal is to show that $F'(t) = G(t)$.

For any $t \in (c,d)$, we can find c_1, d_1 s.t. $c_1 < t < d_1$ and $[c_1,d_1] \subset (c,d)$. Since f and $\frac{\partial f}{\partial t}$ are continuous on $[a,b]\times[c_1,d_1]$, they are also uniformly continuous on the same set. Consider the difference quotient

$$\frac{F(t+h)-F(t)}{h}.$$

Now consider the difference

$$\begin{aligned} \left|\frac{F(t+h)-F(t)}{h} - G(t)\right| &\coloneqq \left|\int_a^b \left(\frac{f(x,t+h)-f(x,t)}{h} - \frac{\partial f}{\partial t}(x,t)\right)\mathrm{d}x\right| \\ &\leqslant \int_a^b \left|\frac{f(x,t+h)-f(x,t)}{h} - \frac{\partial f}{\partial t}(x,t)\right|\mathrm{d}x \end{aligned}$$

Let $h \neq 0$ be such that $t+h \in [c_1,d_1]$. By the mean value theorem for f, $\exists \xi \in (t,t+h)$ such that

$$\frac{f(x,t+h)-f(x,t)}{h} = \frac{\partial f}{\partial t}(x,t)\bigg|_{t=\xi}.$$

Therefore our difference is equal to

$$\left|\int_a^b \left(\frac{\partial f}{\partial t}(x,\xi) - \frac{\partial f}{\partial t}(x,t)\right)\mathrm{d}x\right|.$$

Since $\frac{\partial f}{\partial t}$ is uniformly continuous on $[a,b]\times[c_1,d_1]$, $\forall \varepsilon > 0$, $\exists \delta_\varepsilon > 0$:

$$|t-\xi| < \delta_\varepsilon \implies \left|\frac{\partial f}{\partial t}(x,\xi) - \frac{\partial f}{\partial t}(x,t)\right| < \varepsilon.$$

Therefore, for any $h < \delta_\varepsilon$, we have that

$$\left|\frac{F(t+h)-F(t)}{h} - G(t)\right| \leqslant \int_a^b \varepsilon\,\mathrm{d}x = \varepsilon|b-a|$$

which implies that F is differentiable at t with derivative $F' = G$. ∎

Here is the main theorem of this section. The FTC is a particular case of the following theorem where $a(x) = a \in \mathbb{R}$ is constant, $b(x) = x$ and $f(x,t) = f(t)$ doesn't depend on x:

Theorem 14.3.2 (**Leibniz Integral Rule**) Suppose that $f(x,t)$ and its derivative $\frac{\partial f}{\partial t}$ are continuous on $[\alpha,\beta] \times (c,d)$. Suppose further that for all $t \in (c,d)$: $a(t) \in [\alpha,\beta] \ni b(t)$ and both $a(t)$ and $b(t)$ are differentiable. Then for any $(x,t) \in [\alpha,\beta] \times (c,d)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,t)\,\mathrm{d}x = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(x,t)\,\mathrm{d}x + f(b(t),t)\cdot b'(t) - f(a(t),t)\cdot a'(t).$$

Proof. Define for $(t,a,b) \in (c,d)\times[\alpha,\beta]\times[\alpha,\beta]$ where $a \in \mathbb{R} \ni b$:

$$I(t,a,b) := \int_a^b f(x,t)\,\mathrm{d}x$$

By **Theorem 14.3.1**,

$$\frac{\partial I}{\partial t}(t,a,b) = \int_a^b \frac{\partial f}{\partial t}(x,t)\,\mathrm{d}x$$

and by the fundamental theorem of calculus,

$$\frac{\partial I}{\partial b}(t,a,b) = f(b,t) \quad \frac{\partial I}{\partial a}(t,a,b) = -f(a,t).$$

Then the chain rule of differentiation implies that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b f(x,t)\,\mathrm{d}x \\
&= \frac{\mathrm{d}}{\mathrm{d}t} I(t,a(t),b(t)) \\
&= \frac{\partial I}{\partial t}(t,a(t),b(t))\cdot\frac{\mathrm{d}t}{\mathrm{d}t} + \frac{\partial I}{\partial a}(t,a(t),b(t))\cdot\frac{\mathrm{d}a}{\mathrm{d}t} + \frac{\partial I}{\partial b}(t,a(t),b(t))\cdot\frac{\mathrm{d}b}{\mathrm{d}t} \\
&\overset{14.3.1}{=} \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(x,t)\,\mathrm{d}x - f(a(t),t)\cdot a'(t) + f(b(t),t)\cdot b'(t).
\end{aligned}$$

■

Going back to finding the density of Z:

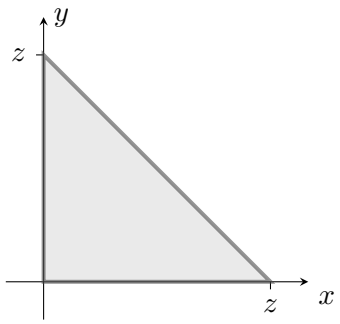
$$\begin{aligned}
f_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z}F_Z(z) \\
&= \frac{\mathrm{d}}{\mathrm{d}z}\int_{-\infty}^{+\infty}\int_{-\infty}^{z-y} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{+\infty}\frac{\mathrm{d}}{\mathrm{d}z}\left(\int_{-\infty}^{z-y} f_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\
&= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{z-y}\frac{\partial}{\partial z}f_{X,Y}(x,y)\,\mathrm{d}x + f_{X,Y}(z-y,y)\cdot\frac{\mathrm{d}}{\mathrm{d}z}(z-y) - f_{X,Y}(-\infty,y)\cdot\frac{\mathrm{d}}{\mathrm{d}z}(-\infty)\right)\mathrm{d}y \\
&= \int_{-\infty}^{+\infty} f_{X,Y}(z-y,y)\,\mathrm{d}y
\end{aligned}$$

Example 14.3.3 If we suppose further that X and Y are independent random variables, the joint density splits and we can recognise the density of Z

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)\,\mathrm{d}y =: f_X(z) \star f_Y(z)$$

as the convolution of the densities $f_X(z)$ and $f_Y(z)$.

Example 14.3.4 As a special case, let X and Y be non-negative random variables and Z their sum. Then the region over which one integrates to find the CDF of Z is the following **bounded** triangle:



since the joint density $f_{X,Y}(x,y)$ is zero for $x<0$ and $y<0$. Therefore,

$$F_Z(z)=\mathbb{P}(\{Z\leqslant z\})=\mathbb{P}(\{0\leqslant Z\leqslant z\})=\mathbb{P}(\{0\leqslant X+Y\leqslant z\})=\int_{y=0}^{y=z}\int_{x=0}^{x=z-y} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

so the bounds of integration are different for different problems.

How will you know what to do? Always do the problem given to you.

Professor Unnikrishna Pillai
One Function of Two Random Variables $Z=X+Y$ (Part 1 of 6)

There are typically some common sense checks e.g. if X and Y are positive, then $Z=X+Y$ is positive. If your work doesn't show this, something's gone awry.

14.3.3 $Z=X-Y$

Draw the line $z=x-y$ i.e. $y=x-z$.

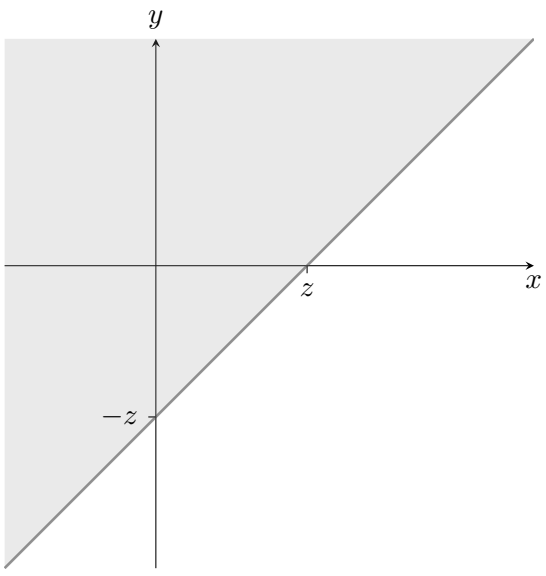


Figure 14.2: The region in $\mathbb{R}^2$ corresponding to the event $\{X-Y\leqslant z\}$.

$$\begin{aligned}
f_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) \\
&= \frac{\mathrm{d}}{\mathrm{d}z} \mathbb{P}(\{Z \leqslant z\}) \\
&= \frac{\mathrm{d}}{\mathrm{d}z} \mathbb{P}(\{X - Y \leqslant z\}) \\
&= \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{\infty} \int_{x=-\infty}^{z+y} f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \\
&= \int_{-\infty}^{\infty} f_{X,Y}(z + y, y)\, \mathrm{d}y \quad \text{by Leibniz's Integral Rule} \\
&= f_X(-z) \star f_Y(y) \quad \text{if } X \text{ and } Y \text{ are independent.}
\end{aligned}$$

14.3.4 $Z = X - Y$ (Non-Negative X, Y)

In the case that X and Y are positive random variables, $Z = X - Y$ may still be negative or positive (unlike the previous section with $Z = X + Y$).

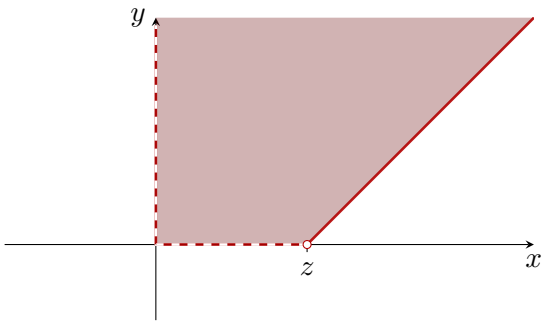


Figure 14.3: Let $z > 0$. The region in $\mathbb{R}^2$ corresponding to the event $\{X - Y \leqslant z\}$ for positive random variables X and Y.

The easier way to integrate over this region is horizontally (over lines of constant Y first with respect to x, and then with respect to y). Doing so vertically would introduce a change of integration bounds over $X = z$. This would be inconvenient.

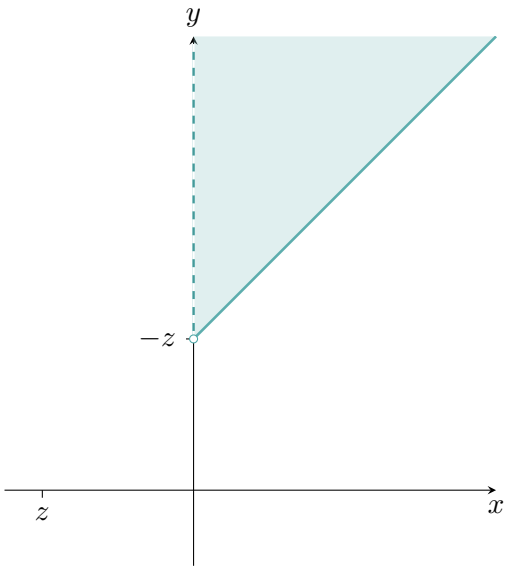


Figure 14.4: Let $z < 0$. The region in $\mathbb{R}^2$ corresponding to the event $\{X - Y \leqslant z\}$ for positive random variables X and Y. Note that $-z > 0$ is the Y-intercept.

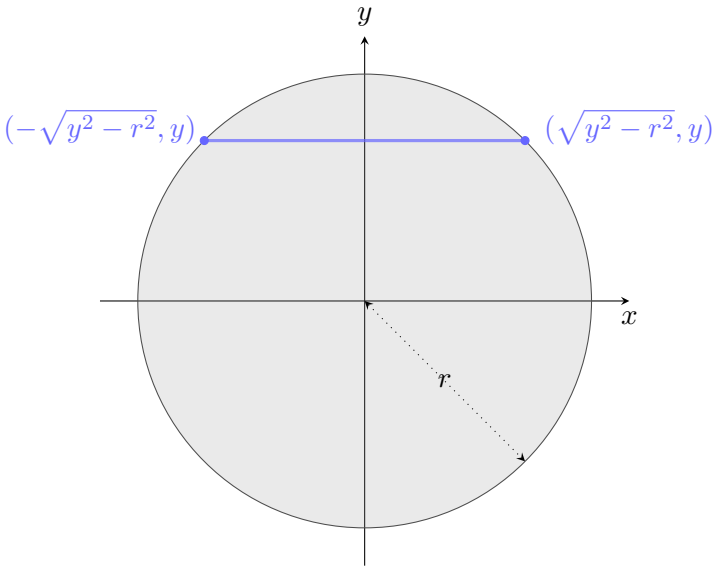
$$\therefore \quad F_Z(z) = \begin{cases} \displaystyle\int_{y=0}^{\infty}\int_{0}^{z+y} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y, & \text{if } z > 0 \\ \displaystyle\int_{y=-z}^{\infty}\int_{0}^{z+y} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y, & \text{if } z \leqslant 0 \end{cases}$$

14.3.5 $R = \sqrt{X^2 + Y^2}$

Let X and Y be independent, zero-mean, normally distributed random variables with equal variance σ^2. The CDF of the "amplitude" $R = \sqrt{X^2 + Y^2}$ is given by

$$\begin{aligned} F_R(r) &= \mathbb{P}(\{R \leqslant r\}) \\ &= \mathbb{P}\Big(\{\sqrt{X^2 + Y^2} \leqslant r\}\Big) \\ &= \mathbb{P}\big(\{X^2 + Y^2 \leqslant r^2\}\big) \\ &= \mathbb{P}\big(\{\omega \in \Omega\colon (X(\omega))^2 + (Y(\omega))^2 \leqslant r^2\}\big) \\ &= \iint_{\{(x,y)\in\mathbb{R}^2\colon x^2+y^2\leqslant r^2\}} f_{X,Y}(x,y)\,\mathrm{d}A \end{aligned}$$

The region being traced out is the area within a circle of radius r and centre $(0,0)$, boundary inclusive. This is where the density of R is defined.



Thus, the CDF of R is given by

$$F_R(r) = \int_{-r}^{r}\int_{-\sqrt{r^2-y^2}}^{+\sqrt{r^2-y^2}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

The density can again be found by differentiating the cumulative distribution function:

$$\begin{aligned}
f_R(r) &= \frac{\mathrm{d}}{\mathrm{d}r}F_R(r) \\
&= 1 \cdot \int_{-\sqrt{r^2-r^2}}^{+\sqrt{r^2-r^2}} f_{X,Y}(x,r)\,\mathrm{d}x - (-1)\int_{-\sqrt{r^2-(-r)^2}}^{+\sqrt{r^2-(-r)^2}} f_{X,Y}(x,-r)\,\mathrm{d}x \\
&\quad + \int_{-r}^{r} \frac{\partial}{\partial r}\left(\int_{-\sqrt{r^2-y^2}}^{+\sqrt{r^2-y^2}} f_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\
&= \int_{-r}^{r} \frac{\partial}{\partial r}\left(\int_{-\sqrt{r^2-y^2}}^{+\sqrt{r^2-y^2}} f_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\
&= \int_{-r}^{r} \Bigg(f_{X,Y}(\sqrt{r^2-y^2},y)\frac{\mathrm{d}}{\mathrm{d}r}\left(\sqrt{r^2-y^2}\right) - f_{X,Y}(-\sqrt{r^2-y^2},y)\frac{\mathrm{d}}{\mathrm{d}r}\left(-\sqrt{r^2-y^2}\right) \\
&\qquad\qquad + \int_{-\sqrt{r^2-y^2}}^{+\sqrt{r^2-y^2}} \frac{\partial}{\partial r} f_{X,Y}(x,y)\,\mathrm{d}x \Bigg)\mathrm{d}y \\
&= \int_{-r}^{r} \left(\frac{r}{\sqrt{r^2-y^2}} f_{X,Y}(\sqrt{r^2-y^2},y) + \frac{r}{\sqrt{r^2-y^2}} f_{X,Y}(-\sqrt{r^2-y^2},y)\right)\mathrm{d}y \\
&= \int_{-r}^{r} \left(\frac{r}{\sqrt{r^2-y^2}} f_X(\sqrt{r^2-y^2}) f_Y(y) + \frac{r}{\sqrt{r^2-y^2}} f_X(-\sqrt{r^2-y^2}) f_Y(y)\right)\mathrm{d}y \quad \text{by indep.} \\
&= \int_{-r}^{r} \frac{r}{\sqrt{r^2-y^2}} \frac{2}{2\pi\sigma^2} \exp\left(\frac{-1}{2\sigma^2}\left(r^2-y^2+y^2\right)\right)\mathrm{d}y \\
&= \frac{r}{\pi\sigma^2}\exp\left(-\tfrac{1}{2\sigma^2}r^2\right) \underbrace{\int_{-r}^{r} \frac{1}{\sqrt{r^2-y^2}}\,\mathrm{d}y}_{\text{even}} \\
&= \frac{2r}{\pi\sigma^2}\exp\left(-\tfrac{1}{2\sigma^2}r^2\right) \int_{0}^{r} \frac{1}{\sqrt{r^2-y^2}}\,\mathrm{d}y \\
&= \frac{r}{\sigma^2}\exp\left(-\tfrac{1}{2\sigma^2}r^2\right) \quad \text{for } r \geqslant 0.
\end{aligned}$$

The **Rayleigh distribution** has density

$$f(x) = \frac{x}{\beta^2}\exp\left(-x^2/(2\beta^2)\right)$$

for $x > 0$ and $\beta > 0$. Comparing the density of R with this, R does indeed have a Rayleigh distribution with $\beta = \sigma$.

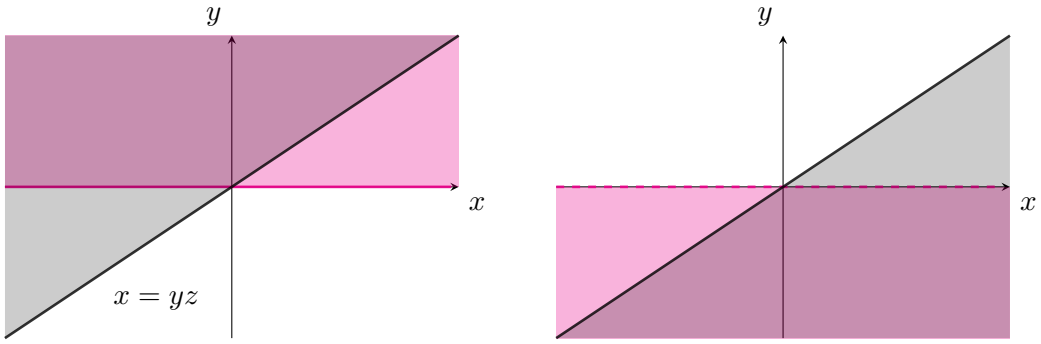
14.3.6 $Z = X/Y$

$$F_Z(z) = \mathbb{P}(\{Z \leqslant z\}) = \mathbb{P}(\{X/Y \leqslant z\})$$

If X and Y were deterministic, we'd have no issues multiplying out by Y but now we have to worry about whether Y is positive or negative.

Define $A = \{Y \geqslant 0\}$ so $A^{\mathsf{c}} = \{Y < 0\}$. We can use this to partition our event $\{X/Y \leqslant z\}$:

$$\begin{aligned}
F_Z(z) &= \mathbb{P}(\{X/Y \leqslant z\}) \\
&= \mathbb{P}(\{X/Y \leqslant z\} \cap (A \sqcup A^{\mathsf{c}})) \\
&= \mathbb{P}(\{X/Y \leqslant z\} \cap A) + \mathbb{P}(\{X/Y \leqslant z\} \cap A^{\mathsf{c}}) \\
&= \mathbb{P}(\{X/Y \leqslant z,\ Y \geqslant 0\}) + \mathbb{P}(\{X/Y \leqslant z,\ Y < 0\}) \\
&= \mathbb{P}(\{X \leqslant Yz,\ Y \geqslant 0\}) + \mathbb{P}(\{X \geqslant Yz,\ Y < 0\})
\end{aligned}$$



Integrating over lines of constant Y first gives:

$$\begin{aligned} f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z} F_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z} \int_{y=0}^{\infty} \int_{x=-\infty}^{yz} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y + \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{0} \int_{x=yz}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &=: I_1 + I_2 \end{aligned}$$

I_1: The integral with respect to y has integral bounds that don't depend on z so the boundary terms in Leibniz's rule $f(x,b(x))b'(x) - f(x,a(x))a'(x)$ vanish. We're left with

$$\begin{aligned} I_1 &= \int_0^{\infty} \frac{\partial}{\partial z} \left(\int_{-\infty}^{yz} f_{X,Y}(x,y)\,\mathrm{d}x \right) \mathrm{d}y \\ &= \int_0^{\infty} \left(f_{X,Y}(yz,y) \frac{\mathrm{d}}{\mathrm{d}z}(yz) - f_{X,Y}(-\infty,y) \frac{\mathrm{d}}{\mathrm{d}z}(-\infty) + \int_{-\infty}^{yz} \frac{\partial}{\partial z} f_{X,Y}(x,y)\,\mathrm{d}x \right) \mathrm{d}y \\ &= \int_0^{\infty} y \cdot f_{X,Y}(yz,y)\,\mathrm{d}y. \end{aligned}$$

I_2: By similar logic, this term becomes

$$I_2 = \int_{-\infty}^{0} (-y) f_{X,Y}(yz,y)\,\mathrm{d}y.$$

In total,

$$\begin{aligned} f_Z(z) &= \int_0^{\infty} y \cdot f_{X,Y}(yz,y)\,\mathrm{d}y + \int_{-\infty}^{0} (-y) f_{X,Y}(yz,y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} |y| \cdot f_{X,Y}(yz,y)\,\mathrm{d}y \end{aligned}$$

Example 14.3.5 Suppose that X and Y are independent and Gaussian with equal mean 0 and equal variance $\sigma^2 = 1$. Then

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} |y| f_X(yz) f_Y(y)\,\mathrm{d}y \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |y| \exp\left(-\frac{(yz)^2}{2} \right) \exp\left(-\frac{(y)^2}{2} \right) \mathrm{d}y \\ &= \frac{1}{2\pi} 2 \int_0^{\infty} y \exp\left(-\frac{y^2(z^2+1)}{2} \right) \mathrm{d}y \\ &= \frac{1}{\pi} \int_0^{\infty} \frac{1}{1+z^2} e^{-u}\,\mathrm{d}u \qquad \text{Let } u = \frac{(1+z^2)y^2}{2} \implies \frac{1}{1+z^2}\,\mathrm{d}u = y\,\mathrm{d}y \\ &= \frac{1}{\pi(1+z^2)} \quad \text{for } z \in \mathbb{R} \end{aligned}$$

This is a **Cauchy density** that has fat tails in comparison to the Gaussian because it only decreases quadratically according to its density as $|z| \to +\infty$:

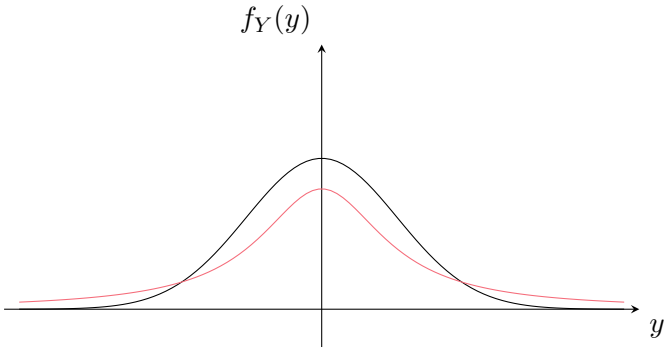


Figure 14.5: Visual comparison of the standard normal density and a Cauchy density.

It turns out that one can relax all the other conditions apart from X and Y being jointly Gaussian for their ratio $Z = X/Y$ to have a Cauchy distribution.

14.3.7 $Z = \max(X, Y)$, $W = \min(X, Y)$

Such random variables naturally occur in the study of order statistics (and reliability analysis). In general, if you have random variables $X_1, \ldots, X_n$ representing the outcomes of n trials of some experiment, one can consider the "best-case analysis" or "worst-case analysis" by ordering each set of observations:

- Teach a class 10 times
- Each class has 30 students
- Take the best and worst score each time
- This will tell you something about the student performance over various realisations (over time)

The simplest case is the max or min of two random variables. These are non-linear functions. We pose the same question as before — what are the densities of Z and W, $f_Z(z)$ and $f_W(w)$ respectively? Despite being non-linear, there's a natural partition that Z and W admit:

$$Z = \max(X, Y) = \begin{cases} X, & \text{if } X \geqslant Y \\ Y, & \text{if } X < Y \end{cases} \qquad W = \min(X, Y) = \begin{cases} Y, & \text{if } X \geqslant Y \\ X, & \text{if } X < Y \end{cases}$$

These are indeed a partition of the outcome space.

Let A denote the set $\{\omega \in \Omega \colon X(\omega) \geqslant Y(\omega)\}$. Then

- $A^{\mathsf{c}} = \{\omega \in \Omega \colon X(\omega) < Y(\omega)\}$
- $A \cup A^{\mathsf{c}} = \Omega$
- $A \cap A^{\mathsf{c}} = \emptyset$.

We'll make use of this partition several times in this section for examples including $\max(X, Y)$, $\min(X, Y)$ and combinations thereof so when we refer to $A \sqcup A^{\mathsf{c}}$, that's what it is.

14.3.8 $Z = \max(X, Y)$

Explicitly, the CDF of $Z = \max(X, Y)$ can be written as

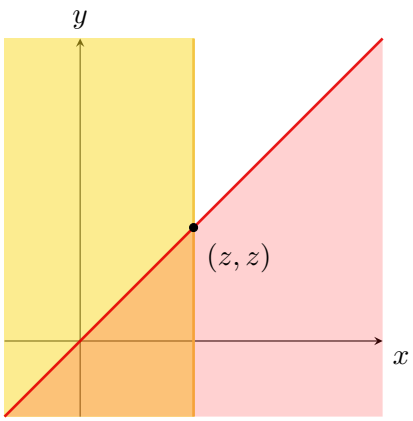
$$\begin{aligned} F_Z(z) &= \mathbb{P}(\{Z \leqslant z\}) \\ &:= \mathbb{P}(\{\omega \in \Omega \colon Z(\omega) \leqslant z\}) \\ &= \mathbb{P}(\{\omega \in \Omega \colon \max(X(\omega), Y(\omega)) \leqslant z\}) \\ &=: \mathbb{P}(B) \\ &= \mathbb{P}(B \cap \Omega) \\ &= \mathbb{P}(B \cap (A \sqcup A^{\mathsf{c}})) \\ &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^{\mathsf{c}}) \end{aligned}$$

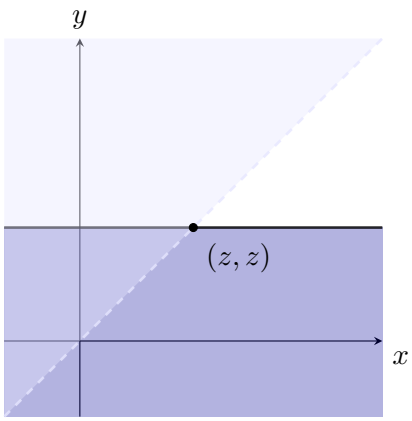
where

$$\begin{aligned} B \cap A &= \{\omega \in \Omega \colon \max(X(\omega), Y(\omega)) \leqslant z \text{ and } X(\omega) \geqslant Y(\omega)\} \\ &= \{\omega \in \Omega \colon X(\omega) \leqslant z \text{ and } X(\omega) \geqslant Y(\omega)\} \end{aligned}$$

$$\begin{aligned} B \cap A^{\mathsf{c}} &= \{\omega \in \Omega \colon \max(X(\omega), Y(\omega)) \leqslant z \text{ and } X(\omega) < Y(\omega)\} \\ &= \{\omega \in \Omega \colon Y(\omega) \leqslant z \text{ and } X(\omega) < Y(\omega)\} \end{aligned}$$

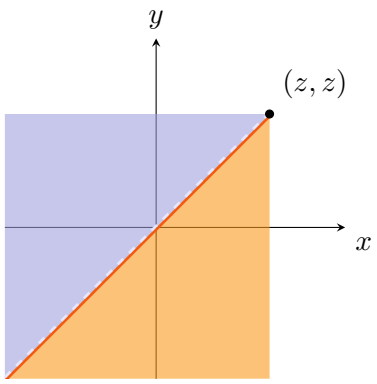
We can visually inspect the regions in the codomain ($\mathbb{R}^2$) of our real random vector $\mathbf{X} = (X, Y)$ corresponding to the events $B \cap A$ and $B \cap A^{\mathsf{c}}$:





Superimposing these diagrams, we can see that the shaded regions are disjoint and their union is

$$\{(x, y) \in \mathbb{R}^2 \colon x \leqslant z,\ y \leqslant z\} = \{(x, y) \in \mathbb{R}^2 \colon (x, y) \leqslant (z, z)\}.$$



Therefore,

$$\begin{aligned} F_Z(z) &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^{\mathsf{c}}) \\ &= \mathbb{P}(\{\omega \in \Omega \colon X(\omega) \leqslant z, Y(\omega) \leqslant z\}) \\ &= F_{X,Y}(z, z) \end{aligned}$$

As a sanity check, if $Z = \max(X, Y) \leqslant z$, then certainly $X \leqslant z$ and $Y \leqslant z$. More compactly, the earlier calculation may be written as:

$$\begin{aligned} F_Z(z) = \mathbb{P}(\{Z \leqslant z\}) &= \mathbb{P}(\{Z \leqslant z\} \cap (\{X \geqslant Y\} \sqcup \{X < Y\})) \\ &= \mathbb{P}(\{Z \leqslant z,\ X \geqslant Y\}) + \mathbb{P}(\{Z \leqslant z,\ X < Y\}) \\ &= \mathbb{P}(\{X \leqslant z,\ X \geqslant Y\}) + \mathbb{P}(\{Y \leqslant z,\ X < Y\}) \end{aligned}$$

This method generalises to $\max(X_1, \ldots, X_n)$.

14.3.9 $W = \min(X, Y)$

Explicitly, the CDF of $W = \min(X, Y)$ can be written as

$$\begin{aligned} F_W(w) &= \mathbb{P}(\{W \leqslant w\}) \\ &:= \mathbb{P}(\{\omega \in \Omega \colon W(\omega) \leqslant w\}) \\ &= \mathbb{P}(\{\omega \in \Omega \colon \min(X(\omega), Y(\omega)) \leqslant w\}) \\ &=: \mathbb{P}(B) \\ &= \mathbb{P}(B \cap \Omega) \\ &= \mathbb{P}(B \cap (A \sqcup A^{\mathsf{c}})) \\ &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^{\mathsf{c}}) \end{aligned}$$

where

$$\begin{aligned} B \cap A &= \{\omega \in \Omega \colon \min(X(\omega), Y(\omega)) \leqslant w \text{ and } X(\omega) \geqslant Y(\omega)\} \\ &= \{\omega \in \Omega \colon Y(\omega) \leqslant w \text{ and } X(\omega) \geqslant Y(\omega)\} \\ B \cap A^{\mathsf{c}} &= \{\omega \in \Omega \colon \min(X(\omega), Y(\omega)) \leqslant w \text{ and } X(\omega) < Y(\omega)\} \\ &= \{\omega \in \Omega \colon X(\omega) \leqslant w \text{ and } X(\omega) < Y(\omega)\} \end{aligned}$$

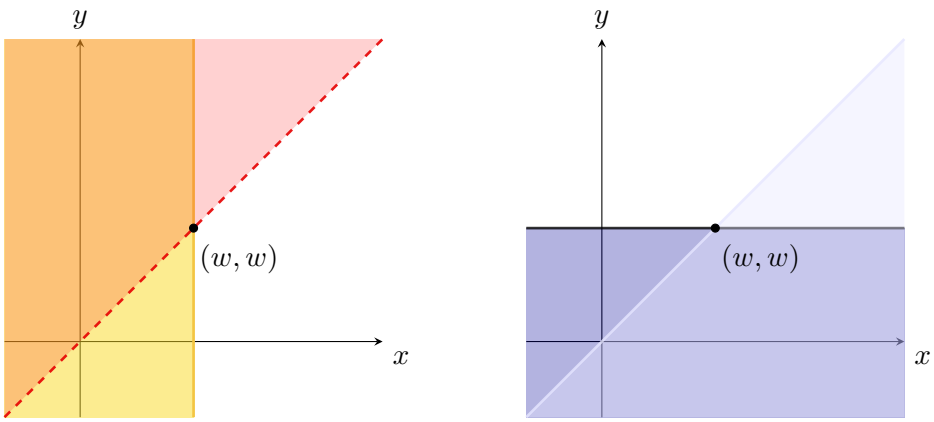


Figure 14.6: The intersections of the shaded regions represent the regions in $\mathbb{R}^2$ corresponding to the subsets $B \cap A^{\mathsf{c}}$ (left) and $B \cap A$ (right) of Ω.

Re-combining these two regions gives:

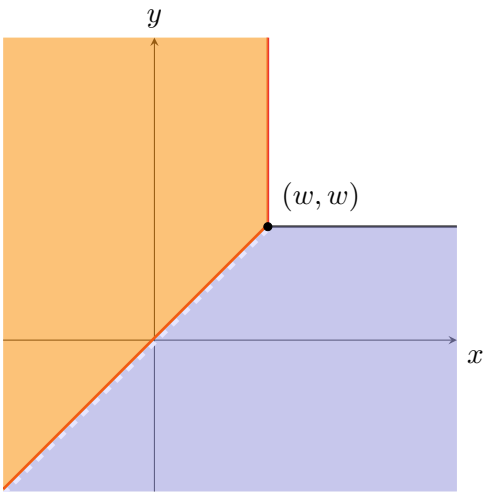


Figure 14.7: Superimposition.

Therefore,

$$\begin{aligned} F_W(w) &= \mathbb{P}(\{X \leqslant w\}) + \mathbb{P}(\{Y \leqslant w\}) - \mathbb{P}(\{X \leqslant w, Y \leqslant w\}) \\ &= F_X(w) + F_Y(w) - F_{X,Y}(w,w) \end{aligned}$$

Example 14.3.6 If X and Y are independent, the density of W can be found by differentiating:

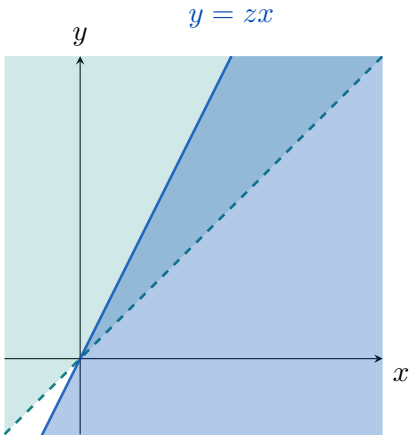
$$\begin{aligned} f_W(w) &= \frac{\mathrm{d}}{\mathrm{d}w} F_W(w) \\ &= \frac{\mathrm{d}}{\mathrm{d}w} \left(F_X(w) + F_Y(w) - F_X(w)F_Y(w)\right) \\ &= f_X(w) + f_Y(w) - \left(f_X(w)F_X(w) + F_X(w)f_Y(w)\right) \\ &= f_X(w)\left(1 - F_Y(w)\right) + f_Y(w)\left(1 - F_X(w)\right) \end{aligned}$$

14.3.10 $Z = \max(X,Y)/\min(X,Y)$

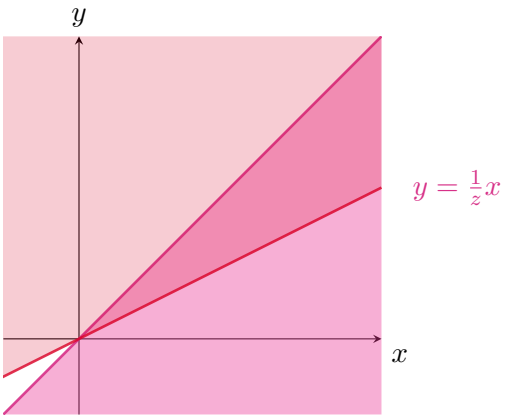
Assume that X and Y are non-negative random variables.

$$\begin{aligned} F_Z(z) = \mathbb{P}(\{Z \leqslant z\}) &= \mathbb{P}\left(\left\{\frac{\max(X,Y)}{\min(X,Y)} \leqslant z\right\}\right) \\ &=: \mathbb{P}(B) \\ &= \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^{\mathsf{c}}) \\ &= \mathbb{P}\left(\left\{\tfrac{X}{Y} \leqslant z,\ X \geqslant Y\right\}\right) + \mathbb{P}\left(\left\{\tfrac{Y}{X} \leqslant z,\ X < Y\right\}\right) \\ &= \mathbb{P}(\{X \leqslant Yz,\ X \geqslant Y\}) + \mathbb{P}(\{Y \leqslant Xz,\ X < Y\}) \end{aligned}$$

Note that $\max(X,Y)/\min(X,Y)$ is always greater than or equal to 1 i.e. $Z \geqslant 1$. The geometric consequences are that:

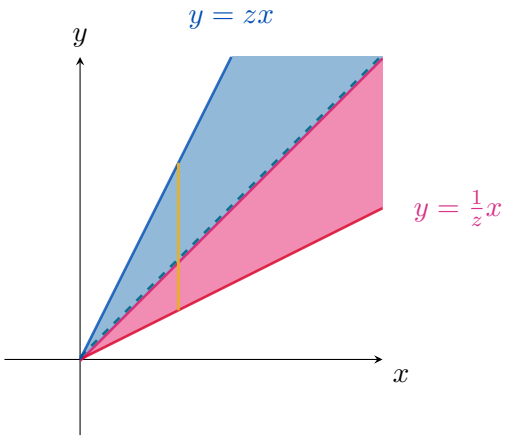


$y = zx$ has a steeper gradient than $y = x$ (dashed) since $Z \geqslant 1$.



$y = \frac{1}{z}x$ has a less steep gradient than $y = x$ because $1/Z \leqslant 1$.

Both regions are disjoint so we may integrate the joint density over their union to calculate $\mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^{\mathsf{c}})$:



Integrating over a slice of constant y first and then with respect to y gives:

$$\begin{aligned} f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z}F_Z(z) &= \frac{\mathrm{d}}{\mathrm{d}z}\int_0^\infty \int_{x=y/z}^{x=yz} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_0^\infty \left(\frac{\partial}{\partial z}\int_{x=y/z}^{x=yz} f_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\ &= \int_0^\infty \left(f_{X,Y}(yz,y)\frac{\mathrm{d}}{\mathrm{d}z}(yz) - f_{X,Y}(y/z,y)\frac{\mathrm{d}}{\mathrm{d}z}(y/z)\right)\mathrm{d}y \\ &= \int_0^\infty \left(y \cdot f_{X,Y}(yz,y) + \frac{y}{z^2}\cdot f_{X,Y}(y/z,y)\right)\mathrm{d}y \end{aligned}$$

14.4 Method of Transformations

An offshoot of the method of distribution functions. Provided that g is either strictly increasing or strictly decreasing, we can find a simple method of writing down the density function of $Y = g(X)$.

Lemma 14.4.1 If $g\colon A \to B$ is strictly monotone, then g^{-1} exists. Also g is strictly increasing (resp. strictly decreasing) if and only if g^{-1} is strictly increasing (resp. decreasing).

Now we compute the CDF of $Y = g(X)$ in terms of the CDF of X.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(\{Y \leqslant y\}) \\ &= \mathbb{P}(\{g(X) \leqslant y\}) \\ &= \mathbb{P}(\{\omega \in \Omega \colon g(X(\omega)) \leqslant y\}) \end{aligned}$$

This is the branching point where g being strictly increasing or strictly decreasing alters the final expression.

- If g is strictly increasing, $g(X(\omega)) \leqslant y \iff g^{-1}(g(X(\omega))) \leqslant g^{-1}(y)$ where $g^{-1}(g(X(\omega))) = X(\omega)$. Therefore, the following sets are equal:

  $$\{\omega \in \Omega \colon g(X(\omega)) \leqslant y\} = \{\omega \in \Omega \colon X(\omega) \leqslant g^{-1}(y)\}$$

  Therefore, the density of Y is given by

  $$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \leqslant g^{-1}(y)\}\big) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y}(g^{-1}(y)) \end{aligned}$$

- If g is strictly decreasing, $g(X(\omega)) \leqslant y \iff g^{-1}(g(X(\omega))) \geqslant g^{-1}(y)$ where $g^{-1}(g(X(\omega))) = X(\omega)$. Therefore, the following sets are equal:

  $$\{\omega \in \Omega \colon g(X(\omega)) \leqslant y\} = \{\omega \in \Omega \colon X(\omega) \geqslant g^{-1}(y)\}$$

  Therefore, the density of Y is given by

  $$\begin{aligned} f_Y(y) &= \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \geqslant g^{-1}(y)\}\big) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} \Big(1 - \mathbb{P}\big(\{\omega \in \Omega \colon X(\omega) \leqslant g^{-1}(y)\}\big)\Big) \\ &= \frac{\mathrm{d}}{\mathrm{d}y} \big(1 - F_X(g^{-1}(y))\big) \\ &= -f_X(g^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y}(g^{-1}(y)) \end{aligned}$$

For g increasing, $\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$ is positive and of course the density f_X is non-negative, so the density f_Y is non-negative.

For g decreasing, f_X is non-negative and $\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$ is negative so $-\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$ is positive.

Therefore, both expressions are valid and we summarise them into a single formula

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

Example 14.4.2 (The Probability Transform) Let X be a continuous random variable whose cumulative distribution function F_X is strictly increasing on the support of X. Then F_X has an inverse function.

Proof. Let $U = F_X(X)$. Then for $u \in [0,1]$:

$$\begin{aligned} F_U(y) &= \mathbb{P}(\{U \leqslant u\}) \\ &= \mathbb{P}(\{F_X(X) \leqslant u\}) \\ &= \mathbb{P}\big(\{F_X^{-1}(F_X(X)) \leqslant F_X^{-1}(u)\}\big) \\ &= \mathbb{P}\big(\{X \leqslant F_X^{-1}(u)\}\big) \\ &= F_X(F_X^{-1}(u)) \\ &= u \end{aligned}$$

i.e. U is a uniformly distributed random variable on $[0,1]$ ■

14.5 Method of Moment-Generating Functions

This method is based on a uniqueness theorem. Two equivalent versions are as follows:

Theorem 14.5.1 (Theorem 6.1 [6]) Let $M_X(t)$ and $M_Y(t)$ denote the moment-generating functions of random variables X and Y respectively. If both MGFs exist, and for all t in some neighbourhood of 0: $M_X(t) = M_Y(t)$, then X and Y have the same probability distribution.

Theorem 14.5.2 (2.3.11 [1]) Let F_X, F_Y be two CDFs all of whose moments exist. If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only $\mathbb{E}(X^r) = \mathbb{E}(Y^r)$ for all non-negative integers r.

If $U = h(Y_1, \ldots, Y_n)$, the goal of this method is to determine the distribution of U by finding its moment-generating function $M_U(t) = \mathbb{E}\big(e^{tU}\big)$. Once found, we compare it with the MGFs of well-known distributions. If $M_U(t)$ is identical to one of these, the above uniqueness theorem can be used to conclude U's probability distribution.

The method of moment-generating functions is very useful for finding the distribution of a sum of independent random variables.

Theorem 14.5.3 Let $Y_1, \ldots, Y_n$ be mutually independent random variables with respective moment generating functions $M_{Y_1}(t), \ldots, M_{Y_n}(t)$. Define $U = \sum_{i=1}^n Y_i$ Then

$$M_U(t) = \prod_{i=1}^n M_{Y_i}(t).$$

Proof.

$$\begin{aligned} M_U(t) \coloneqq \mathbb{E}(\exp(tU)) = \mathbb{E}\left(\exp\left(t\sum_{i=1}^n Y_i\right)\right) &= \mathbb{E}\left(\prod_{i=1}^n \exp(tY_i)\right) \\ &= \prod_{i=1}^n \mathbb{E}(\exp(tY_i)) \quad \text{by mutual independence} \\ &\eqqcolon \prod_{i=1}^n M_{Y_i}(t) \end{aligned}$$

■

The method of moment-generating functions can be used to establish some useful results for the distribution of a function of normally distributed random variables. These will be used later in **15.1.1**.

Example 14.5.4 (**Example 6.10 [6]**) Suppose that Y is normally distributed with mean μ and variance σ^2. Show that $Z = (Y - \mu)/\sigma$ has a standard normal distribution (mean 0 and variance 1).

Solution.

$$M_Z(t) = \mathbb{E}\left(e^{tZ}\right) = \mathbb{E}\left(e^{(t/\sigma)(Y-\mu)}\right) = M_{Y-\mu}\left(\tfrac{t}{\sigma}\right)$$

Now note that

$$\begin{aligned}
M_{Y-\mu}(t) &= \mathbb{E}\left(e^{t(Y-\mu)}\right) \\
&= \int_{-\infty}^{\infty} e^{t(y-\mu)} f_Y(y)\,\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{t(y-\mu)} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}y \\
&\overset{u=y-\mu}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{tu} \exp\left(\frac{-u^2}{2\sigma^2}\right) \mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(\frac{-1}{2\sigma^2}\left((u-(t\sigma^2))^2 - (t\sigma^2)^2\right)\right) \mathrm{d}u \\
&= \exp\left(\tfrac{\sigma^2 t^2}{2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(\tfrac{-1}{2\sigma^2}\left((u-(t\sigma^2))^2\right)\right) \mathrm{d}u \\
&= \exp\left(\tfrac{\sigma^2}{2} t^2\right).
\end{aligned}$$

This term is the integral of the probability density function of a normally distributed random variable with mean $t\sigma^2$ and variance σ^2 over its support. Therefore, it is equal to 1. Finally, the moment generating function of Z is given by

$$M_Z(t) = M_{Y-\mu}(\tfrac{t}{\sigma}) = \exp\left(\tfrac{\sigma^2}{2}\left(\tfrac{t}{\sigma}\right)^2\right) = \exp\left(\tfrac{1}{2}t^2\right).$$

Comparing this to the MGF of a standard normal random variable and appealing to the uniqueness theorem concludes the proof. ■

Example 14.5.5 (**Example 6.11 [6]**) Let $Z \sim \mathcal{N}(0,1)$. Use the method of moment-generating functions to find the probability distribution of Z^2.

Solution.

$$\begin{aligned}
M_{Z^2}(t) = \mathbb{E}\left(e^{tZ^2}\right) &= \int_{-\infty}^{\infty} e^{tz^2} f_Z(z)\,\mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(tz^2 - \tfrac{z^2}{2}\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-z^2\left(\tfrac{1}{2} - t\right)\right) \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{1}{2}-t}} \underbrace{\int_{-\infty}^{\infty} e^{-u^2}\,\mathrm{d}u}_{=\sqrt{\pi}} \qquad \begin{array}{l}\text{by the substitution } u = z\sqrt{(1/2)-t}, \\ \text{assuming } (1/2)-t \geqslant 0\end{array} \\
&= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{2}}{\sqrt{1-2t}} \sqrt{\pi} \\
&= \frac{1}{\sqrt{1-2t}}
\end{aligned}$$

Once again comparing with known MGFs, $M_{Z^2}(t)$ is identical to the MGF of a random variable with a $\chi^2_{\nu=1}$ distribution

$$(1-2t)^{-\nu/2}\Big|_{\nu=1.}$$

Equivalently, this is the MGF of a random variable with Gamma($\alpha = \frac{1}{2}, \beta = 2$) distribution

$$(1-\beta t)^{-\alpha}\Big|_{\alpha=\frac{1}{2},\ \beta=2.}$$

Therefore, the density of Z^2 is equal to

$$f_Z(z) = \frac{z^{-1/2}e^{-z/2}}{\Gamma(\frac{1}{2})2^{1/2}}\mathbb{1}_{[0,\infty)}(z).$$

■

14.6 Multivariable/variate Transformations Using Jacobians

Deferring this until later when I encounter a use-case. I did indeed find a use-case later on in 15.1.1.

14.7 Order Statistics

Many functions of random variables of interest depend on the relative magnitudes of the observed values e.g. we may be interested in the fastest time in a race, the highest test score etc. Thus, we often order observed random variables according to their magnitudes. The resulting ordered variables are called **order statistics**.

Formally, let $Y_1, \ldots, Y_n$ denote independent random variables. Suppose further that they are all absolutely continuous with cumulative distribution function $F(y)$ and density $f(y)$. We denote the ordered random variables by $Y_{(1)}, \ldots, Y_{(n)}$ where $Y_{(1)} \leqslant Y_{(2)} \leqslant \ldots \leqslant Y_{(n)}$. Using this notation,

$$Y_{(1)} = \min(Y_1, \ldots, Y_n) \qquad Y_{(n)} = \max(Y_1, \ldots, Y_n)$$

are the minimum and maximum of the random variables respectively. The density functions of $Y_{(1)}$ and $Y_{(n)}$ can be found using the method of distributions — this was done earlier in **14.3.8**. Another method to find the densities is as follows:

Since $Y_{(n)}$ is the maximum of the Y_i, the event $\{Y_{(n)} \leqslant y\}$ occurs iff the events $\{Y_i \leqslant y\}$ occur for every $i = 1, \ldots, n$. That is:

$$\begin{aligned} F_{Y_{(n)}} = \mathbb{P}\big(\{Y_{(n)} \leqslant y\}\big) &= \mathbb{P}(\{Y_1 \leqslant y, \ldots, Y_n \leqslant y\}) \\ &= \mathbb{P}(\{Y_1 \leqslant y\}) \cdot \ldots \cdot \mathbb{P}(\{Y_n \leqslant y\}) \quad \text{by independence} \\ &= (F(y))^n \quad \text{all } Y_i \text{ are identically distributed} \end{aligned}$$

Let $g_{(n)}(y)$ denote the density of $Y_{(n)}$.

$$\therefore g_{(n)}(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}\big(\{Y_{(n)} \leqslant y\}\big) = n(F(y))^{n-1}f(y).$$

The density of $Y_{(1)}$ can be found similarly:

The event that $\{Y_{(1)} \geqslant y\}$ occurs iff $\{Y_i \geqslant y\}$ occurs for $i = 1, \ldots, n$. Therefore,

$$\begin{aligned} F_{Y_{(1)}} = \mathbb{P}\big(\{Y_{(1)} \leqslant y\}\big) &= 1 - \mathbb{P}\big(\{Y_{(1)} \geqslant y\}\big) \\ &= 1 - \mathbb{P}(\{Y_1 \geqslant y, \ldots, Y_n \geqslant y\}) \\ &= 1 - \mathbb{P}(\{Y_1 \geqslant y\}) \cdot \ldots \cdot \mathbb{P}(\{Y_n \geqslant y\}) \quad \text{by independence} \\ &= 1 - \prod_{i=1}^{n} \mathbb{P}(\{Y_i \geqslant y\}) \\ &= 1 - \prod_{i=1}^{n} (1 - \mathbb{P}(\{Y_i \geqslant y\})) \\ &= 1 - \prod_{i=1}^{n} (1 - F(y)) \qquad \text{all } Y_i \text{ are identically distributed} \\ &= 1 - (1 - F(y))^n \end{aligned}$$

$$\therefore \; g_{(1)}(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}\big(\{Y_{(1)} \leqslant y\}\big) = -n(1 - F(y))^{n-1}\frac{\mathrm{d}}{\mathrm{d}y}(1 - F(y)) = n(1 - F(y))^{n-1}f(y).$$

Example 14.7.1 Consider the case $n = 2$. Find the joint density of $Y_{(1)}$ and $Y_{(2)}$.

Consider the distribution function of $\mathbf{Y} = (Y_{(1)}, Y_{(2)})$

$$F_{\mathbf{Y}}(y_1, y_2) = \mathbb{P}\big(\{Y_{(1)} \leqslant y_1, Y_{(2)} \leqslant y_2\}\big)$$

The event

$$\begin{aligned}\{\mathbf{Y} \leqslant (y_1, y_2)\} &= \{(Y_{(1)}, Y_{(2)}) \leqslant (y_1, y_2)\} \\ &= \{Y_{(1)} \leqslant y_1\} \cap \{Y_{(2)} \leqslant y_2\} \\ &= \{\min(Y_1, Y_2) \leqslant y_1\} \cap \{\max(Y_1, Y_2) \leqslant y_2\}\end{aligned}$$

It's always true that $Y_{(1)} \leqslant Y_{(2)}$ since the minimum of two numbers can never exceed their maximum. This means that if $y_1 > y_2$, we can write the event $\{Y_{(1)} \leqslant y_1\}$ as the disjoint union $\{Y_{(1)} \leqslant y_2\} \sqcup \{y_2 < Y_{(1)} \leqslant y_1\}$. Therefore,

$$\begin{aligned}\{\mathbf{Y} \leqslant (y_1, y_2)\} &= \{Y_{(1)} \leqslant y_1\} \cap \{Y_{(2)} \leqslant y_2\} \\ &= \big(\{Y_{(1)} \leqslant y_2\} \sqcup \{y_2 < Y_{(1)} \leqslant y_1\}\big) \cap \{Y_{(2)} \leqslant y_2\} \\ &= \big(\{Y_{(1)} \leqslant y_2\} \cap \{Y_{(2)} \leqslant y_2\}\big) \sqcup \big(\{y_2 < Y_{(1)} \leqslant y_1\} \cap \{Y_{(2)} \leqslant y_2\}\big) \\ &= \big(\{Y_{(1)} \leqslant y_2\} \cap \{Y_{(2)} \leqslant y_2\}\big) \sqcup \underbrace{\{Y_{(2)} \leqslant y_2 < Y_{(1)} \leqslant y_1\}}_{= \emptyset \text{ because } Y_{(1)} \leqslant Y_{(2)}} \\ &= \{Y_{(1)} \leqslant y_2\} \cap \{Y_{(2)} \leqslant y_2\}\end{aligned}$$

Thus, we can conclude that

$$\{\mathbf{Y} \leqslant (y_1, y_2)\} = \begin{cases}\{Y_{(1)} \leqslant y_1\} \cap \{Y_{(2)} \leqslant y_2\}, & y_1 \leqslant y_2 \\ \{Y_{(1)} \leqslant y_2\} \cap \{Y_{(2)} \leqslant y_2\}, & y_1 > y_2.\end{cases}$$

By definition $\mathbf{Y} = (Y_{(1)}, Y_{(2)}) = \begin{cases}\{(Y_1, Y_2), & \text{if } Y_1 \leqslant Y_2 \\ \{(Y_2, Y_1), & \text{if } Y_1 \leqslant Y_2.\end{cases}$

- Case 1: $y_1 \leqslant y_2$

$$\begin{aligned}F_{\mathbf{Y}}(y_1, y_2) = \mathbb{P}(\{\mathbf{Y} \leqslant (y_1, y_2)\}) &= \mathbb{P}\big(\{Y_{(1)} \leqslant y_1, Y_{(2)} \leqslant y_2)\}\big) \\ &= \mathbb{P}(\{\min(Y_1, Y_2) \leqslant y_1, \max(Y_1, Y_2) \leqslant y_2\}) \\ &= \mathbb{P}\Big(\underbrace{\{Y_1 \leqslant y_1, Y_2 \leqslant y_2\}}_{\text{the case } Y_1 \leqslant Y_2} \cup \underbrace{\{Y_2 \leqslant y_1, Y_1 \leqslant y_2\}}_{\text{the case } Y_1 > Y_2}\Big) \\ &=: \mathbb{P}(A \cap B)\end{aligned}$$

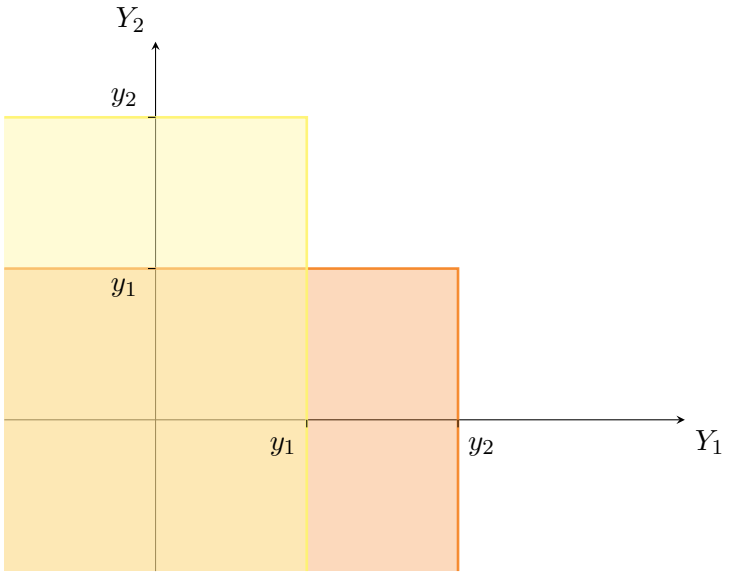


Figure 14.8: The region of interest $A \cap B$ is doubly counted in $\mathbb{P}(A \cup B)$ so we must calculate $\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

$$
\begin{aligned}
&\mathbb{P}(A \cap B) \\
&= \mathbb{P}(\{Y_1 \leqslant y_1, Y_2 \leqslant y_2\}) + \mathbb{P}(\{Y_2 \leqslant y_1, Y_1 \leqslant y_2\}) - \mathbb{P}(\{Y_1 \leqslant y_1, Y_1 \leqslant y_2, Y_2 \leqslant y_1, Y_2 \leqslant y_2\}) \\
&= \mathbb{P}(\{Y_1 \leqslant y_1, Y_2 \leqslant y_2\}) + \mathbb{P}(\{Y_2 \leqslant y_1, Y_1 \leqslant y_2\}) - \mathbb{P}(\{Y_1 \leqslant y_1, Y_2 \leqslant y_1\}) \\
&= \mathbb{P}(\{Y_1 \leqslant y_1\})\mathbb{P}(\{Y_2 \leqslant y_2\}) + \mathbb{P}(\{Y_2 \leqslant y_1\})\mathbb{P}(\{Y_1 \leqslant y_2\}) - \mathbb{P}(\{Y_1 \leqslant y_1\})\mathbb{P}(\{Y_2 \leqslant y_1\}) \\
&\qquad \text{by independence} \\
&= F(y_1)F(y_2) + F(y_1)F(y_2) - (F(y_1))^2 \quad \text{by identical distribution} \\
&= 2F(y_1)F(y_2) - (F(y_1))^2
\end{aligned}
$$

- Case 2: $y_1 > y_2$

  This case is simpler. The joint CDF of $\mathbf{Y}$ is given by

$$
\begin{aligned}
F_{\mathbf{Y}}(y) &= \mathbb{P}\big(\{Y_{(1)} \leqslant y_1, Y_{(2)} \leqslant y_2)\}\big) \\
&= \mathbb{P}(\{\min(Y_1, Y_2) \leqslant y_1, \max(Y_1, Y_2) \leqslant y_2\}) \\
&= \mathbb{P}(\{Y_1 \leqslant y_2, Y_2 \leqslant y_2\} \cup \{Y_2 \leqslant y_2, Y_1 \leqslant y_2\}) \\
&= \mathbb{P}(\{Y_1 \leqslant y_2, Y_2 \leqslant y_2\}) \\
&= \mathbb{P}(\{Y_1 \leqslant y_2\})\mathbb{P}(\{Y_2 \leqslant y_2\}) \quad \text{by independence} \\
&= F(y_2)F(y_2) \quad \text{by identical distribution} \\
&= (F(y_2))^2
\end{aligned}
$$

In summary:

$$
F_{\mathbf{Y}}(y_1, y_2) = F_{(Y_{(1)}, Y_{(2)})}(y_1, y_2) = \begin{cases} 2F(y_1)F(y_2) - (F(y_2))^2, & \text{if } y_1 \leqslant y_2 \\ (F(y_2))^2, & \text{if } y_1 > y_2 \end{cases}
$$

and their joint density is denoted by $g_{(1)(2)}$ and is obtained by partial differentiation

$$
g_{(1)(2)}(y_1, y_2) = \frac{\partial^2}{\partial y_1 \partial y_2} F_{(Y_{(1)}, Y_{(2)})}(y_1, y_2) = \begin{cases} 2f(y_1)f(y_2), & \text{if } y_1 \leqslant y_2 \\ 0, & \text{if } y_1 > y_2. \end{cases}
$$

In general, the joint density of $Y_{(1)}, \ldots, Y_{(n)}$ is found to be

$$g_{(1)\ldots(n)}(y_1, \ldots, y_n) = \begin{cases} n! \prod_{i=1}^{n} f(y_i), & \text{if } y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n \\ 0, & \text{otherwise.} \end{cases}$$

CHAPTER 15

Sampling Distributions

Now that we have a few methods for determining the distribution of a function T of a collection of random variables $Y_1, \ldots, Y_n$, we can turn our attention back to the sampling distribution of an estimator. Recall that an estimator is a composition $T \circ \mathbf{X}$ where T is a statistic and, for the purposes of idealising the sampling process, $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample.[1]

Properties of the Sample Mean

Lemma 15.0.1 Let $x_1, \ldots, x_n$ be real numbers and let $\overline{x} = (x_1 + \ldots + x_n)/n$.

(a) $\min_a \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2$

(b) $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - n(\overline{x})^2$

Proof.

(a) Let $f(a) \coloneqq \sum_{i=1}^{n} (x_i - a)^2$. Differentiating with respect to a shows that the minimum of f is attained at $a = \overline{x}$.

(b)

$$(n-1)s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - 2\overline{x} \underbrace{\left(\sum_{i=1}^{n} x_i \right)}_{n\overline{x}} + n(\overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2 \right) - n(\overline{x})^2$$

■

Useful Results For Sampling Distributions

Lemma 15.0.2 Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathbb{E}(g(X_1))$ and $\mathrm{Var}(g(X_1))$ exist. Then

(a) $\mathbb{E}\left(\sum_{i=1}^{n} g(X_i) \right) = n\mathbb{E}(g(X_1))$

(b) $\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i) \right) = n\mathrm{Var}(g(X_1))$

Proof.

(i) The first expression follows from the linearity of expectation.

[1] Recall that a random sample is a collection of i.i.d. random variables defined on the same probability space.

(ii)

$$
\begin{aligned}
&\operatorname{Var}\left(\sum_{i=1}^{n} g(X_i)\right) \\
&= \mathbb{E}\left(\left(\sum_{i=1}^{n} g(X_i) - \mathbb{E}\left(\sum_{i=1}^{n} g(X_i)\right)\right)^2\right) \\
&= \mathbb{E}\left(\left(\sum_{i=1}^{n} g(X_i) - \sum_{i=1}^{n} \mathbb{E}(g(X_i))\right)^2\right) \\
&= \mathbb{E}\left(\left(\sum_{i=1}^{n} \left(g(X_i) - \mathbb{E}(g(X_i))\right)\right)^2\right) \\
&\overset{7.1}{=} \mathbb{E}\left(\left(\sum_{i=1}^{n} \left(g(X_i) - \mathbb{E}(g(X_i))\right)^2\right) + 2\sum_{\substack{i=1 \\ i>j}}^{n} \left(g(X_i) - \mathbb{E}(g(X_i))\right)\left(g(X_j) - \mathbb{E}(g(X_j))\right)\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\Big(\left(g(X_i) - \mathbb{E}(g(X_i))\right)^2\Big) + 2\sum_{\substack{i=1 \\ i>j}}^{n} \operatorname{Cov}(g(X_i), g(X_j)) \\
&= \sum_{i=1}^{n} \operatorname{Var}(g(X_i)) \text{ by } \textbf{Corollary 7.2.2} \\
&= n\operatorname{Var}(g(X_1))
\end{aligned}
$$

■

Theorem 15.0.3 (Theorem 5.2.6 [1]) Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

(a) $\mathbb{E}(\overline{X}) = \mu$

(b) $\operatorname{Var}(\overline{X}) = \sigma^2/n$

(c) $\mathbb{E}(S^2) = \sigma^2$

Proof.

(a) Trivial by linearity: $\mathbb{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \mu$

(b) Note that

$$\operatorname{Var}(\overline{X}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \operatorname{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right)$$

which is of the form $\operatorname{Var}(\sum_{i=1}^{n} g(X_i))$ where $g(x) = x/n$. By **Lemma 15.0.2 (b)**, we conclude that

$$\operatorname{Var}(\overline{X}) = n\operatorname{Var}(g(X_1)) = n\operatorname{Var}\left(\frac{X_1}{n}\right) = \frac{n}{n^2}\operatorname{Var}(X_1) = \frac{\sigma^2}{n}.$$

(c) By **Lemma 15.0.1 (b)**

$$
\begin{aligned}
\mathbb{E}(S^2) &:= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\left(\sum_{i=1}^{n}(X_i)^2\right) - n(\overline{X})^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i)^2\right) - \frac{n}{n-1}\mathbb{E}\left((\overline{X})^2\right)
\end{aligned}
$$

- For each i, $\mathbb{E}(X_i) < \infty$ and $\text{Var}(X_i) < \infty$ by assumption.
- Parts (a) and (b) tell us that $\mathbb{E}\big(\overline{X}\big)$ and $\text{Var}\big(\overline{X}\big)$ are finite.

In both cases, the general expression $\text{Var}(Y) = \mathbb{E}\big(Y^2\big) - (\mathbb{E}(Y))^2$ holds and tells us that if we replace Y with any of the X_i or $\overline{X}$, then $\mathbb{E}\big(Y^2\big) < \infty$.

- For the first term, the hypotheses of **Lemma 15.0.2 (a)** are satisfied with $g(x) = x^2$.
- For the second term, we can simply substitute in $\mathbb{E}(\overline{X}^2) = \text{Var}\big(\overline{X}\big) + (\mathbb{E}\big(\overline{X}\big))^2$:

Finally, we conclude that

$$\begin{aligned}
\mathbb{E}\big(S^2\big) &= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} g(X_i)\right) - \frac{n}{n-1}\mathbb{E}\big((\overline{X})^2\big) \\
&= \frac{1}{n-1} n\mathbb{E}(g(X_1)) - \frac{n}{n-1}\left(\text{Var}\big(\overline{X}\big) + (\mathbb{E}\big(\overline{X}\big))^2\right) \qquad \text{by Lemma 15.0.2 (a)} \\
&= \frac{n}{n-1}\left(\text{Var}(X_1) + (\mathbb{E}(X_1))^2\right) - \frac{n}{n-1}\left(\tfrac{\sigma^2}{n} + \mu^2\right) \\
&= \frac{n}{n-1}\left(\sigma^2 + \mu^2 - \tfrac{\sigma^2}{n} - \mu^2\right) = \sigma^2
\end{aligned}$$

■

Definition 15.0.4 An **unbiased estimator** is an estimator whose expectation equals the estimand.

Example 15.0.5 Relationships (a) $\mathbb{E}\big(\overline{X}\big) = \mu$, and (c) $\mathbb{E}\big(S^2\big) = \sigma^2$ demonstrate that $\overline{X}$ and S^2 are unbiased estimators of μ and σ^2 respectively.

Sampling Distribution of $\overline{X}$

Lemma 15.0.6 (Exercise 5.5 [1]) Let $X_1, \ldots, X_n$ be i.i.d. with density function $f_X(x)$, and let $\overline{X}$ denote the sample mean. Show that

$$f_{\bar{X}}(x) = n f_{X_1 + \ldots + X_n}(nx)$$

is the density function of $\overline{X}$.

Proof.

$$\begin{aligned}
F_{\bar{X}}(x) &= \mathbb{P}\big(\overline{X} \leqslant x\big) \\
&= \mathbb{P}\big(\tfrac{1}{n}(X_1 + \ldots + X_n) \leqslant x\big) \\
&= \mathbb{P}(X_1 + \ldots + X_n \leqslant nx) \\
&= F_{X_1 + \ldots + X_n}(nx)
\end{aligned}$$

Upon differentiating with respect to x,

$$f_{\bar{X}}(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_{\bar{X}}(x) = n f_{X_1 + \ldots + X_n}(nx).$$

■

Lemma 15.0.7 (Theorem 5.2.7 [1]) Let $X_1, \ldots, X_n$ be a random sample from a population with MGF $M_X(t)$. Then the MGF of the sample mean is

$$M_{\overline{X}}(t) = \left(M_X(\tfrac{t}{n})\right)^n.$$

Proof.

$$\begin{aligned}
M_{\overline{X}}(t) &\coloneqq \mathbb{E}\big(\exp(t\overline{X})\big) \\
&= \mathbb{E}\big(\exp\left(\tfrac{t}{n}(X_1 + \ldots + X_n)\right)\big) \coloneqq M_{X_1+\ldots+X_n}\left(\tfrac{t}{n}\right) \\
&= \mathbb{E}\left(\prod_{i=1}^{n} \exp\left(\tfrac{t}{n}X_i\right)\right) \\
&= \prod_{i=1}^{n} \mathbb{E}\big(\exp\left(\tfrac{t}{n}X_i\right)\big) \quad \text{by independence} \\
&= \prod_{i=1}^{n} \mathbb{E}\big(\exp\left(\tfrac{t}{n}X_1\right)\big) \quad \text{by identical distribution} \\
&=: \prod_{i=1}^{n} M_{X_1}\left(\tfrac{t}{n}\right) \\
&= \left(M_{X_1}\left(\tfrac{t}{n}\right)\right)^n
\end{aligned}$$

■

The above two results transform any statement about the density of $X_1+\ldots+X_n$ to a statement about the density of $\overline{X}$.

If the MGF lemma isn't applicable (either because the MGF of $\overline{X}$ is unrecognisable or because the population MGF doesn't exist) then the method of transformations may prove useful for finding the density of $X_1 + \ldots + X_n$.

15.1 Sampling From A Normally Distributed Population

Many real-world phenomena have relative frequency distributions that can be modelled adequately by a normal distribution. The sample mean of a normal random sample is normal:

Theorem 15.1.1 (Theorem 7.1 [7]) Let $X_1, \ldots, X_n$ be a random sample from a normal population i.e. a population whose associated distribution is $\mathcal{N}(\mu, \sigma^2)$. Then $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Proof. Recall that the MGF of $X_i \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$M_{X_i}(t) = \exp\left(t\mu + \tfrac{t^2}{2}\sigma^2\right).$$

It follows that

$$M_{\overline{X}}(t) \overset{15.0.7}{=} \left(M_{X_1}\left(\tfrac{t}{n}\right)\right)^n = \exp\left(\sum_{i=1}^{n}\left(\tfrac{t}{n}\mu + \tfrac{(t/n)^2}{2}\sigma^2\right)\right) = \exp\left(t\mu + \tfrac{t^2}{2}\left(\tfrac{\sigma^2}{n}\right)\right)$$

i.e. $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. ■

Example 15.1.2 A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It's been observed that the amount of fill dispensed by the machine is normally distributed with $\sigma = 1$ ounce. A sample of $n = 9$ bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting), and the ounces of fill are measured for each. Find the probability that the sample mean will be within 0.3 ounces of the true mean μ for the chosen machine setting.

Solution: 'All bottled with the same machine setting' is an assumption of identical distribution. Each trial of filling a bottle is independent. Therefore, $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, 9$ constitute a

random sample of size $n = 9$. From **Theorem 15.1.1**, $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/9)$. Our task is to find

$$\begin{aligned}\mathbb{P}\big(|\overline{X} - \mu| \leqslant 0.3\big) &= \mathbb{P}\big(-0.3 \leqslant \overline{X} - \mu \leqslant 0.3\big) \\ &= \mathbb{P}\left(\frac{-0.3}{\dots} \leqslant \frac{\overline{X} - \mu}{\dots} \leqslant \frac{0.3}{\dots}\right)\end{aligned}$$

This is written in terms of the standard normal random variable so we can read off the probabilities from a statistical table.

$$\begin{aligned}\dots &= \mathbb{P}\left(\frac{-0.3}{\sigma/\sqrt{n}} \leqslant \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant \frac{0.3}{\sigma/\sqrt{n}}\right) \\ &= \mathbb{P}(-0.9 \leqslant Z \leqslant 0.9) \\ &= 1 - 2\mathbb{P}(Z \leqslant -0.9) \quad \text{or equally } 1 - 2\mathbb{P}(Z \geqslant 0.9) \\ &\approx \text{ read off table}\end{aligned}$$

Example 15.1.3 Continued: How many observations should be included in the sample if we wish for $\overline{X}$ to be within 0.3 ounces of μ with probability 0.95?

Solution: By similar logic,

$$\begin{aligned}0.95 = \mathbb{P}\big(|\overline{X} - \mu| \leqslant 0.3\big) = \mathbb{P}\left(\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} \leqslant \frac{0.3\sqrt{n}}{\sigma}\right) &= \mathbb{P}\big(|Z| \leqslant 0.3\sqrt{n}\big) \\ &= 1 - 2\mathbb{P}\big(Z > 0.3\sqrt{n}\big)\end{aligned}$$

Rearranging gives us that $\mathbb{P}(Z > 0.3\sqrt{n}) = \frac{1-0.95}{2} = 0.025$. Now we look at the standard normal table for a probability of 0.025 and read off that $0.3\sqrt{n} = 1.96 \implies \sqrt{n} = 1.96/0.3 \implies n \approx 42.68\overline{4}$. Take $n = 43$.

> Apparently, sums of squares of the observations in a random sample from a normal population are important because they follow a chi-squared distribution and that's important for hypothesis testing. Also, the chi-squared distribution is supposedly related to the sample variance S^2 of a normal random sample. For now, it's poorly motivated in the texts I'm reading.

Theorem 15.1.4 (Theorem 6.4 [7]) Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ be a random sample. Define

$$Z_i := \frac{X_i - \mu}{\sigma}.$$

Then the Z_i are mutually independent and

$$\sum_{i=1}^{n} Z_i^2 \sim \chi_n^2.$$

Proof. The argument is as follows:

- X_i independent $\overset{6.5.6}{\implies}$ Z_i independent (as each Z_i is a function of only X_i).
- We've already demonstrated[2] that $X_i \sim \mathcal{N}(\mu, \sigma^2) \implies Z_i \sim \mathcal{N}(0, 1)$ in **Example 14.5.4**.

[2] Alternatively, we can notice that the X_i are i.i.d. and belong to the same location-scale family with distributions in $\mathcal{N}(\mu, \sigma^2)$ so Z is in standard measure as per **Definition 10.10.4**.

- Define $V := \sum_{i=1}^{n} Z_i^2$ and argue via moment-generating functions:

$$\begin{aligned} M_V(t) := \mathbb{E}(\exp(tV)) &= \mathbb{E}\left(\prod_{i=1}^{n} \exp(tZ_i^2)\right) \\ &= \prod_{i=1}^{n} \mathbb{E}\big(\exp(tZ_i^2)\big) \text{ by independence of the } Z_i \\ &=: \prod_{i=1}^{n} M_{Z_i^2}(t) \\ &= \left(M_{Z_i^2}(t)\right)^n \text{ by identical distribution of the } Z_i \\ &= \left(\frac{1}{\sqrt{1-2t}}\right)^n \text{ by } \textbf{example 14.5.5} \end{aligned}$$

■

Theorem 15.1.4 (just above) links to **(c)** in **Theorem 15.1.5** below. Namely, the former uses the population mean in the definition of Z_i, but the latter's part (c) uses the sample mean (the consequence of which is a reduction by 1 of the d.f. in the resulting distribution).

Theorem 15.1.4	**Theorem 15.1.5**
$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i-\mu}{\sigma}\right)^2$	$\frac{(n-1)}{\sigma^2} S^2 = (n-1) \sum_{i=1}^{n} \left(\frac{X_i-\overline{X}}{\sigma}\right)^2$
$\sim \chi_n^2$	$\sim \chi_{n-1}^2$

Theorem 15.1.5 (Theorem 5.3.1 [1]) Let $X_1, \ldots, X_n$ be a random sample from a normal population (whose associated distribution follows a normal density $\mathcal{N}(\mu, \sigma^2)$.) Then:

(a) **Theorem 15.1.1**

(b) $\overline{X}$ and S^2 are independent random variables

(c) $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

Proof. Let's prove (b):

We can assume, without loss of generality, that $\mu = 0$ and $\sigma = 1$ and discuss sums of squares of independent, standard normal variables. This is because the section on location-scale families says there's a correspondence between any location-scale distributed random variable X and the "standard" location-scale distributed random variable Z. This correspondence is defined by

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

We'll use this to demonstrate that $\overline{X}$ and S^2 are independent random vectors. First, note that:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \\ &= \frac{1}{n-1} \left((X_1 - \overline{X})^2 + \sum_{i=2}^{n} (X_i - \overline{X})^2 \right) \\ &= \frac{1}{n-1} \left(\left(\sum_{i=2}^{n} (X_i - \overline{X}) \right)^2 + \sum_{i=2}^{n} (X_i - \overline{X})^2 \right) \end{aligned}$$

That last line comes from noticing that

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0 \implies (X_1 - \overline{X}) = -\sum_{i=2}^{n}(X_i - \overline{X})$$

and then squaring this relation.

This means that we've written S^2 as a function of only $(X_2 - \overline{X}, \ldots, X_n - \overline{X})$. The next step is to demonstrate independence from $\overline{X}$. Now we can refer back to **Theorem 13.2.3** and write the *joint density* of the random sample $X_1, \ldots, X_n$:

$$\begin{aligned} f(x_1, \ldots, x_n) &= \frac{1}{\sqrt{2\pi(1)^2}} \exp\left(\frac{-(x_1 - 0)^2}{2(1)^2}\right) \cdot \ldots \cdot \frac{1}{\sqrt{2\pi(1)^2}} \exp\left(\frac{-(x_n - 0)^2}{2(1)^2}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \end{aligned}$$

Now if we perform the linear transformation

$$\begin{cases} x_1 \longmapsto \overline{x} =: y_1 \\ x_2 \longmapsto x_2 - \overline{x} =: y_2 \\ \vdots \\ x_n \longmapsto x_n - \overline{x} =: y_n, \end{cases}$$

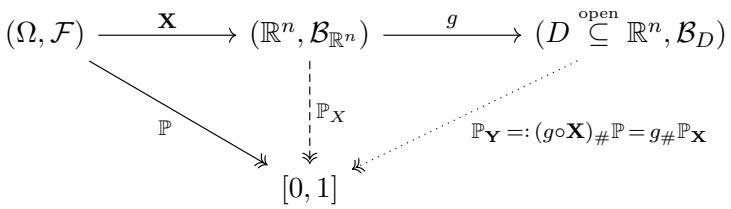
then the density f transforms according to the following section: ◧

15.1.1 Multivariable-Multivariate Transformations

Consider an absolutely continuous random vector $\mathbf{X}\colon \Omega \to \mathbb{R}^n$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We can define a new random vector $\mathbf{Y} = g \circ \mathbf{X}$ on the same probability space by transforming $\mathbf{X}$ with a suitable measurable[3] function g. Let $g\colon C \to D$ also be injective, where C is an open set s.t. $\operatorname{supp}(\mathbf{X}) \subseteq C$.

For the purpose of what follows, for each $\boldsymbol{x}^* \in \operatorname{supp}(\mathbf{X})$, suppose that the Jacobian matrix $J_g(\boldsymbol{x}^*)$ of g at $\boldsymbol{x}^*$ is invertible (i.e. $\det J_g(\boldsymbol{x}^*) \neq 0$) and continuous at and near $\boldsymbol{x}^*$. As a consequence of these conditions, g is locally invertible[4] near $\boldsymbol{x}^*$ and its local inverse has Jacobian $J_{g^{-1}}(g(\boldsymbol{x}^*)) = (J_f(\boldsymbol{x}^*))^{-1}$. This will come in handy shortly.

We can speak of the probability distribution of $\mathbf{Y}$:



It is defined for all $B \in \mathcal{B}_D$ by

$$\mathbb{P}_{\mathbf{Y}}(B) = \mathbb{P}_{g\circ\mathbf{X}}(B) = (g_{\#}\mathbb{P}_{\mathbf{X}})(B) := \mathbb{P}_{\mathbf{X}}\big(g^{-1}(B)\big) = \int_{g^{-1}(B)} f_{\mathbf{X}}(\boldsymbol{x})\, \mathrm{d}\lambda_{\mathbb{R}^n}(\boldsymbol{x}) \quad \text{since } \mathbb{P}_{\mathbf{X}} \ll \lambda_{\mathbb{R}^n}$$

At this point, I want a way to transform this integral into an integral over the corresponding transformed region which is a subset of $g(\operatorname{supp}(\mathbf{X})) \subseteq \operatorname{supp}(\mathbf{Y})$. The reason for the inclusion instead of defining $g(\operatorname{supp}(\mathbf{X}))$ to be the support of $\mathbf{Y}$ is that we can't necessarily guarantee that the former is a closed set — given some standard **regularity assumptions**, the support is always closed.

[3] Since g is measurable, so too is the composition $\mathbf{Y} = g \circ \mathbf{X}$ of measurable functions.

[4] This is the higher-dimensional analogue of strict monotonicity in the single-variable case.

Theorem 15.1.6 (Jacobian Change of Variables) Let B be an open subset of $\mathbb{R}^n$, $h\colon B \to \mathbb{R}^n$ be an injective, $\mathcal{C}^1$ function whose Jacobian is non-zero for every $\boldsymbol{y} \in B$. Then for any real-valued, compactly supported, continuous function f with support contained in $h(B)$:

$$\int_{h(B)} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_B f(h(\boldsymbol{y}))|\det(Dh)(\boldsymbol{y})|\,\mathrm{d}\boldsymbol{y}$$

where the transformation is $(x_1,\ldots,x_n) = h(y_1,\ldots,y_n)$.

In the context of our integral, $h = g^{-1}\colon B \to g^{-1}(B)$, $f = f_{\mathbf{X}}$ and we have that:

$$\begin{aligned}
\int_{g^{-1}(B)} f_{\mathbf{X}}(\boldsymbol{x})\,\mathrm{d}\lambda_{\mathbb{R}^n}(\boldsymbol{x}) &= \int_{h(B)} f_{\mathbf{X}}(\boldsymbol{x})\,\mathrm{d}\lambda_{\mathbb{R}^n}(\boldsymbol{x}) \\
&= \int_B f_{\mathbf{X}}(h(\boldsymbol{y}))|\det J_h(\boldsymbol{y})|\,\mathrm{d}\boldsymbol{y} \\
&= \int_B f_{\mathbf{X}}(g^{-1}(\boldsymbol{y}))|\det J_{g^{-1}}(\boldsymbol{y})|\,\mathrm{d}\boldsymbol{y} \\
&= \int_B \underbrace{f_{\mathbf{X}}(g^{-1}(\boldsymbol{y}))\frac{1}{|\det J_g(\boldsymbol{y})|}}_{f_{\mathbf{Y}}(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y}
\end{aligned}$$

Resuming Proof of (b): The transformation $g(x_1,\ldots,x_n) = (\overline{x}, x_2-\overline{x},\ldots,x_n-\overline{x})$ has Jacobian

$$J_g(x_1,\ldots,x_n) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \cdots & \frac{\partial g_n}{\partial x_n} \end{bmatrix}$$

where $g_i(x_1,\ldots,x_n)$ is the i^{th} component of $g(\boldsymbol{x})$. If $i > 1$:

$$\frac{\partial g_i}{\partial x_j} = \frac{\partial}{\partial x_j}(x_i - \overline{x}) = \frac{\partial}{\partial x_j}\left(x_i - \sum_{k=1}^n x_k\right) = \delta_{ij} - \frac{1}{n}\sum_{k=1}^n \delta_{kj} = \delta_{ij} - \frac{1}{n} = \begin{cases} 1-\frac{1}{n}, & \text{if } j = i \\ 0-\frac{1}{n}, & \text{if } j \neq i \end{cases}$$

Else, $i = 1$ and:

$$\frac{\partial g_1}{\partial x_j} = \frac{\partial}{\partial x_j}(\overline{x}) = \frac{1}{n}.$$

Therefore:

$$J_g(x_1,\ldots,x_n) = \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ -1/n & 1-(1/n) & \cdots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \cdots & 1-(1/n) \end{bmatrix}$$

(I've highlighted in red where $i = 1$, in green where $i > 1$ and $i = j$, and in blue where $i > 1$ and $i \neq j$.)

If we denote by $\boldsymbol{a}_i$, the i^{th} row of the Jacobian matrix, we can perform the following elementary row operations $\boldsymbol{a}_i \longmapsto \boldsymbol{a}_i + \boldsymbol{a}_1$ for $i > 1$ (on $J_g(\boldsymbol{x})$ while the determinant remains unchanged. The resulting determinant is given by

$$\det J_g(\boldsymbol{x}) = \det \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \frac{1}{n}$$

where we've expanded out the determinant about the $(1,1)$ entry.

Now the inverse of g, defined by $g^{-1}(y_1, \ldots, y_n) = (x_1, \ldots, x_n)$ can be found by expressing each x_i in terms of the y_i i.e. $x_i = g^{-1}(\boldsymbol{y})$.

We know that

$$y_1 = g_1(\boldsymbol{x}) = \overline{x} = \frac{x_1 + \ldots + x_n}{n} \iff ny_1 = x_1 + x_2 + \ldots + x_n$$

Now we can substitute in the remaining $n-1$ expressions for g, namely that for $i > 1$: $y_i = x_i - \overline{x}$ i.e. $x_i = \underbrace{y_i + y_1}_{= g^{-1}(\boldsymbol{y})}$, into our expression for y_1:

$$ny_1 = x_1 + \sum_{i=2}^{n}(y_i + y_1) \iff x_1 = y_1 - \left(\sum_{i=2}^{n} y_i\right)$$

Thus, the inverse transformation is defined by

$$g^{-1}(y_1, \ldots, y_n) = \left(y_1 - \sum_{i=2}^{n} y_i, y_2 + y_1, \ldots, y_n + y_1\right).$$

Finally, we can conclude that the density transforms according to the following equation:

$$\begin{aligned}
f_{\mathbf{Y}}(\boldsymbol{y}) &= f_{\mathbf{X}}(g^{-1}(\boldsymbol{y}))\frac{1}{|\det J_g(\boldsymbol{y})|} \\
&= f_{\mathbf{X}}\left(y_1 - \sum_{i=2}^{n} y_i, y_2 + y_1, \ldots, y_n + y_1\right)\frac{1}{|1/n|} \\
&= \frac{1}{1/n} \cdot \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\left(\left(y_1 - \sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n}(y_i + y_1)^2\right)\right) \\
&= \ldots \\
&= \frac{n}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}n(y_1)^2\right) \exp\left(-\frac{1}{2}\left(\sum_{i=2}^{n}(y_i^2) + \left(\sum_{i=2}^{n} y_i\right)^2\right)\right)
\end{aligned}$$

The above calculation shows that the joint density of $Y_1 = \overline{X}$ and $(Y_2, \ldots, Y_n) = (X_2 - \overline{X}, \ldots, X_n - \overline{X})$ factors according to **Theorem 6.5.5** so they are mutually independent.

Then we can let $U_1 = \mathrm{id}(Y_1)$ and $U_2 = S^2(Y_2, \ldots, Y_n)$ in the statement of **Theorem 6.5.6**, so that $\overline{X}$ and S^2 are mutually independent.

Now for part (c). We must show that $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Rest of proof goes here.

■

Example 15.1.7 (Example 7.5 [7]) Sometimes it's useful to specify an interval of values that will include S^2 with a high probability.

From the prior example about the bottling machine, ounces of fill are assumed to be distributed according to $\mathcal{N}(\mu, \sigma^2 = 1)$. Suppose that we select a random sample of 10 bottles and measure the amount of fill in each. If these 10 observations are used to calculate S^2, it might be useful to specify b_1 and b_2 s.t.

$$\mathbb{P}\big(b_1 \leqslant S^2 \leqslant b_2\big) = 0.9$$

Solution: There's a bit of a theme here. Inequality manipulation to read off a table!

$$\begin{aligned}\mathbb{P}\big(b_1 \leqslant S^2 \leqslant b_2\big) &= \mathbb{P}\Big(\tfrac{(n-1)b_1}{\sigma^2} \leqslant \tfrac{(n-1)}{\sigma^2}S^2 \leqslant \tfrac{(n-1)b_2}{\sigma^2}\Big) \\ &= \mathbb{P}\big((n-1)b_1 \leqslant (n-1)S^2 \leqslant (n-1)b_2\big) \quad \text{since } \sigma^2 = 1\end{aligned}$$

Since $n = 10$, note that $9S^2 \sim \chi^2_9$ and its density looks like the following plot:

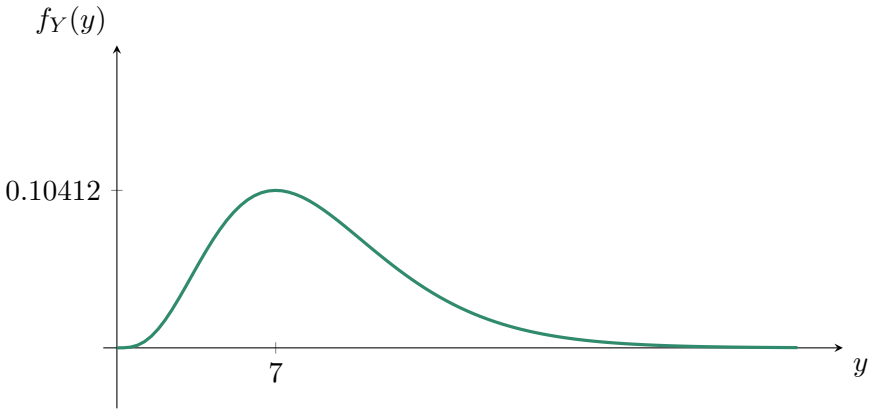


Figure 15.1: A plot for the density of $9S^2 \sim \Gamma(\alpha = 9/2, \beta = 2)$.

Now we can read off values from a statistical table for the interval endpoints that satisfy a χ^2 distribution with 9 degrees of freedom containing 90% of the area under the curve. Such an interval is not unique. Wackely et. al choose values a_1 and a_2 that cut off areas of 0.05 in the lower and upper tails, respectively. Reading off values from a table gives

$$3.325 = a_1 = \frac{(n-1)b_1}{\sigma^2} = 9b_1 \iff b_1 = \frac{3.325}{9} = 0.369$$

and

$$16.919 = a_2 = \frac{(n-1)b_2}{\sigma^2} = 9b_2 \iff b_2 = \frac{16.919}{9} = 1.880.$$

Thus, one such interval that contains S^2 with probability 0.90 is $(0.369, 1.880)$.

15.2 The t-distribution

If σ^2 is known, then we can refer back to **Theorem 15.1.1** which tells us that $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Then the quantity

$$\frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has a standard normal distribution and we can use this to estimate μ.

However, in most practical cases the variance σ^2 is unknown and we resort to the obvious — estimate σ with the sample standard deviation $S = \sqrt{S^2}$, leaving us with the quantity

$$\frac{\overline{X} - \mu}{(S/\sqrt{n})}$$

as the object of our investigation to estimate μ. What's the distribution of this quantity? We can write it in terms of things we've already seen:

$$\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\sigma}}{\frac{S/\sqrt{n}}{\sigma}} = \frac{\frac{\overline{X}-\mu}{\sigma}}{\frac{\sqrt{S^2}}{\sigma}\cdot\frac{1}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{(\sigma/\sqrt{n})}}{\sqrt{\frac{S^2}{\sigma^2}}}$$

- The numerator has a $\mathcal{N}(0,1)$ distribution.
- The denominator is a scalar multiple of a χ^2_{n-1} distributed random variable that is independent of the numerator by parts (c) and (b) of **Theorem 15.1.5**, respectively.

This reduces the problem to finding the distribution of $\frac{Z}{\sqrt{W/\nu}}$ where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_\nu$ are independent random variables.

15.2.1 The Density of Student's t-distribution

Definition 15.2.1 Let $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_\nu$ be independent. Then

$$T = \frac{Z}{\sqrt{W/\nu}}$$

is said to have a **t-distribution with ν degrees of freedom**, denoted $T \sim t_\nu$.

The density of the t_ν distribution is not explicitly given in Wackerly et al. — rather being guided through some exercises. The fundamental property of the Gamma function will be used often. For $\Re(z) > 0$:

$$\Gamma(z+1) = z\Gamma(z) \tag{Γ}$$

Exercise 5 (4.89 [6], 4.111 [7]) Suppose that $Y \sim \text{Gamma}(\alpha, \beta)$, $\alpha, \beta > 0$.

(a) If a is any positive or negative value s.t. $a + \alpha > 0$, show that

$$\mathbb{E}(Y^a) = \frac{\beta^a \Gamma(a+\alpha)}{\Gamma(\alpha)}.$$

(b) Why did the answer to (a) require that $a + \alpha > 0$?

(c) Show that, with $a = 1$, the result in (a) gives $\mathbb{E}(Y) = \alpha\beta$.

(d) Use the result in (a) to give an expression for $\mathbb{E}\left(\sqrt{Y}\right)$. What do you need to assume about α?

(e) Use the result in (a) to give an expression for $\mathbb{E}(1/Y)$, $\mathbb{E}\left(1/\sqrt{Y}\right)$, and $\mathbb{E}\left(1/Y^2\right)$. What do you need to assume about α in each case?

Proof. Note that $\alpha > 0$ presupposes all the following manipulations involving $Y \sim \text{Gamma}(\alpha, \beta)$.

(a)

$$\begin{aligned}
\mathbb{E}(Y^a) &= \int_{\mathbb{R}} y^a f_Y(y)\,dy \\
&= \int_{\mathbb{R}} y^a \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha\Gamma(\alpha)}\mathbb{1}_{[0,+\infty)}(y)\,dy \\
&= \frac{\beta^a\Gamma(a+\alpha)}{\Gamma(\alpha)}\int_{\mathbb{R}} \frac{y^{(a+\alpha)-1}e^{-y/\beta}}{\beta^{a+\alpha}\Gamma(a+\alpha)}\mathbb{1}_{[0,+\infty)}(y)\,dy \\
&= \frac{\beta^a\Gamma(a+\alpha)}{\Gamma(\alpha)}.
\end{aligned}$$

The last integral evaluates to 1 as it's the integral of the density of a $\text{Gamma}(a+\alpha, \beta)$ distributed random variable over its support.

(b) We require $(\alpha > 0) \wedge (a+\alpha > 0)$ so that $\Gamma(t)$ is well-defined for $t \in \{\alpha, \alpha+a\}$.

(c)
$$\mathbb{E}\big(Y^1\big) = \frac{\beta^1\Gamma(\alpha+1)}{\Gamma(\alpha)} \overset{(\Gamma)}{=} \frac{\beta\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = \alpha\beta$$
is valid for $\alpha > 0$.

(d)
$$\mathbb{E}\Big(\sqrt{Y}\Big) = \frac{\beta^{1/2}\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)}$$
is valid for $(\alpha > 0) \wedge (\alpha + \frac{1}{2} > 0) \iff (\alpha > 0) \wedge (\alpha > -1/2) \iff \alpha > 0$.

(e)
- $\mathbb{E}\big(Y^{-1}\big) = \dfrac{\Gamma(\alpha-1)}{\beta\Gamma(\alpha)} \overset{(\Gamma)}{=} \dfrac{1}{\beta(\alpha-1)}$, where $(\alpha > 0) \wedge (\alpha > 1)$.
- $\mathbb{E}\Big(Y^{-1/2}\Big) = \dfrac{\Gamma(\alpha-(1/2))}{\beta^{1/2}\Gamma(\alpha)}$, where $(\alpha > 0) \wedge (\alpha > 1/2)$ i.e. $\alpha > 1/2$.
- $\mathbb{E}\big(Y^{-2}\big) = \dfrac{\Gamma(\alpha-2)}{\beta^2\Gamma(\alpha)} \overset{(\Gamma)}{=} \dfrac{1}{\beta^2(\alpha-1)(\alpha-2)}$, where $(\alpha > 0) \wedge (\alpha > 2)$ i.e. $\alpha > 2$.

■

Exercise 6 (4.90 [6], 4.112 [7]) Suppose that $Y \sim \chi^2_\nu$. Use the results from 4.89 in your answers to the following. These results will be useful when we study the t and F distributions.

(a) Give an expression for $\mathbb{E}(Y^a)$ if $\nu > -2a$.

(b) Why did your answer in (a) require that $\nu > -2a$?

(c) Use the result in (a) to give an expression for $\mathbb{E}\Big(\sqrt{Y}\Big)$. What do you need to assume about ν?

(d) Use the result in (a) to give an expression for $\mathbb{E}(1/Y)$, $\mathbb{E}\Big(1/\sqrt{Y}\Big)$, and $\mathbb{E}\big(1/Y^2\big)$. What do you need to assume about ν in each case?

Proof. $Y \sim \chi^2_\nu$ is equivalent to $Y \sim \text{Gamma}\left(\alpha = \frac{\nu}{2}, \beta = 2\right)$.

(a) Using the last exercise's part (a), we have that
$$\mathbb{E}(Y^a) = \frac{2^a\Gamma(a+\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})}$$

(b) where $(\alpha > 0) \wedge (a + \alpha > 0) \iff (\nu/2 > 0) \wedge (a + \frac{\nu}{2} > 0) \iff (\nu > 0) \wedge (\nu > -2a)$.

(c)
$$\mathbb{E}\Big(\sqrt{Y}\Big) = \mathbb{E}\Big(Y^{1/2}\Big) = \frac{2^{1/2}\,\Gamma(\frac{\nu}{2}+\frac{1}{2})}{\Gamma(\frac{\nu}{2})}$$
is valid for $(\nu > 0) \wedge (\nu > -2(1/2) \iff \nu > 0$.

(d) (i) $\mathbb{E}\big(Y^{-1}\big) = \dfrac{\Gamma(\frac{\nu}{2}-1)}{2\Gamma(\frac{\nu}{2})}$, where $(\nu > 0) \wedge (\nu > -2(-1))$ i.e. $\nu > 2$.

(ii) $\mathbb{E}\Big(Y^{-1/2}\Big) = \dfrac{\Gamma(\frac{\nu}{2}-\frac{1}{2})}{2^{1/2}\Gamma(\frac{\nu}{2})}$, where $(\nu > 0) \wedge (\nu > -2(-1/2))$ i.e. $\nu > 1$.

(iii)
$$\mathbb{E}\big(Y^{-2}\big) = \frac{\Gamma(\frac{\nu}{2}-2)}{2^2\Gamma(\frac{\nu}{2})} \overset{(\Gamma)}{=} \frac{1}{4\big(\frac{\nu}{2}-1\big)\big(\frac{\nu}{2}-2\big)} = \frac{1}{(\nu-2)(\nu-4)}$$
where $(\nu > 0) \wedge (\nu > -2(-2))$ i.e. $\nu > 4$.

■

Exercise 7 (7.72 [6], 7.98 [7]) Suppose that T is defined as the ratio $\frac{Z}{\sqrt{W/\nu}}$ where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_\nu$, with Z and W independent.

(a) If W is fixed at w, then T is given by Z/c where $c = \sqrt{w/\nu}$. Use this idea to find the conditional density of T for a fixed $W = w$.

(b) Find the joint density of T and W using $f(t,w) = f(t \mid w)f(w)$.

> What follows is the proof one would see in an undergraduate-level textbook — many suggestive symbols that behave sensibly without rigorous definition. A rigorous proof can be found later in Example 18.3.17.

Proof. Suppose that $Y = (Y_1, Y_2)$. Later, it will be demonstrated that the conditional density of Y_1 given $Y_2 = y_2$, denoted by $f_{Y_1|Y_2}(y_1|y_2)$, is defined as the integrand of

$$F(y_1 \mid y_2) = \int_{-\infty}^{y_1} \frac{f_Y(t_1, y_2)}{f_{Y_2}(y_2)}\,\mathrm{d}t_1$$

for any y_2 s.t. $f_{Y_2}(y_2) > 0$. Recall that $f_{Y_2}(y_2)$ is a density, not the probability of the event $\{Y_2 = y_2\}$.

(a) The independence of Z and W ensures that conditioning T (which is a function of Z and W) on the event $\{W = w\}$ doesn't affect the distribution of Z (in T) i.e. independence allows us to write[5] the incredibly suggestive abomination that is

$$(T \mid \{W = w\}) = \sqrt{\frac{\nu}{w}}Z.$$

What do those symbols mean? Absolutely nothing... but something? Thus, $(T \mid \{W = w\}) = Z/c$ where $c = \sqrt{w/\nu}$ and $Z \sim \mathcal{N}(0,1)$. Now note that by **Example 11.0.4**, $(T \mid \{W = w\}) \sim \mathcal{N}(0, 1/c^2)$ where $1/c^2 = \nu/w$.

(b) Note that $W \sim \chi^2_\nu$ and $(T \mid \{W = w\}) \sim \mathcal{N}(0, \nu/w)$. The joint density of T and W is given by

$$\begin{aligned} f_{T,W}(t,w) &= f_{T|W=w}(t \mid w) \cdot f_W(w) \\ &= \frac{1}{\sqrt{2\pi(\sqrt{\nu/w})^2}} \exp\left(\frac{-(t-0)^2}{2(\sqrt{\nu/w})^2}\right) \cdot \frac{w^{(\nu/2)-1}e^{-w/2}}{2^{\nu/2}\Gamma(\nu/2)} \mathbb{1}_{[0,+\infty)}(w) \\ &= \frac{\sqrt{w}}{\sqrt{2\pi\nu}} \exp\left(\frac{-t^2 w}{2\nu}\right) \cdot \frac{w^{(\nu/2)-1}e^{-w/2}}{2^{\nu/2}\Gamma(\nu/2)} \mathbb{1}_{[0,+\infty)}(w) \end{aligned}$$

The marginal distribution of T is found by marginalising the joint density:

$$\begin{aligned} f_T(t) &= \int_{\mathbb{R}} f_{T,W}(t,w)\,\mathrm{d}w \\ &= \int_{\mathbb{R}} f_{T|W=w}(t \mid w) f_W(w)\,\mathrm{d}w \\ &= \int_{\mathbb{R}} \frac{\sqrt{w}}{\sqrt{2\pi\nu}} \exp\left(\frac{-t^2 w}{2\nu}\right) \cdot \frac{w^{(\nu/2)-1}e^{-w/2}}{2^{\nu/2}\Gamma(\nu/2)} \mathbb{1}_{[0,+\infty)}(w)\,\mathrm{d}w \\ &= \frac{1}{\sqrt{\pi\nu}\ 2^{(\nu+1)/2}\Gamma(\nu/2)} \int_0^\infty w^{\left(\frac{\nu}{2}+\frac{1}{2}\right)-1} \exp\left(-w \cdot \frac{1}{2}\left(\frac{t^2}{\nu}+1\right)\right)\mathrm{d}w \end{aligned}$$

Now compare this to the density of a Gamma(α, β) distribution

$$\frac{y^{\alpha-1}\exp(-y \cdot (1/\beta))}{\beta^\alpha \Gamma(\alpha)}$$

[5] My reasoning is shaky right now but I superficially understand conditioning T on $\{W = w\}$ as the concept of restricting the probability space to events where $W(\omega) = w$ (i.e considering the subset generated by $\{W = w\}$, of the σ-algebra $\sigma(W)$. The independence of W and Z tells us by definition that their respectively generated σ-algebras $\sigma(W)$ and $\sigma(Z)$ are mutually $\mathbb{P}$-independent i.e.

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

for all $A \in \sigma(Z)$, $B \in \sigma(W)$. This independence allows us to treat Z as though it were still standard-normally distributed under the act of conditioning T on $\{W = w\}$.

so that the location and scale parameters are

$$\alpha = \frac{\nu}{2} + \frac{1}{2} \quad \text{and} \quad \frac{1}{\beta} = \frac{1}{2}\left(\frac{t^2}{\nu} + 1\right)$$
$$\iff \beta = 2\left(\frac{t^2}{\nu} + 1\right)^{-1}.$$

We now multiply and divide the integrand by $\beta^\alpha \Gamma(\alpha)$ in order to rewrite it as the integral of a Gamma(α, β) distributed random variable over its support:

$$\begin{aligned} f_T(t) &= \frac{\beta^\alpha \Gamma(\alpha)}{\sqrt{\pi\nu}\, 2^{(\nu+1)/2}\, \Gamma(\nu/2)} \overbrace{\int_0^\infty \frac{w^{\alpha-1} \exp(-w/\beta)}{\beta^\alpha \Gamma(\alpha)} \, \mathrm{d}w}^{=1} \\ &= \frac{\Gamma(\frac{\nu}{2} + \frac{1}{2})}{\sqrt{\pi\nu}\, 2^{(\nu+1)/2} \left(\frac{1}{2}\left(\frac{t^2}{2} + 1\right)\right)^{\frac{\nu}{2}+\frac{1}{2}} \Gamma(\frac{\nu}{2})} \\ &= \frac{\Gamma(\frac{\nu}{2} + \frac{1}{2})}{\sqrt{\pi\nu}\left(\frac{t^2}{2} + 1\right)^{\frac{\nu}{2}+\frac{1}{2}} \Gamma(\frac{\nu}{2})} \end{aligned}$$

■

15.2.2 Properties of the t_ν-distribution

Corollary 15.2.2 The t-distribution is symmetric about the origin.

Exercise 8 (*7.14 [6], *7.30 [7]) Let $T = \dfrac{Z}{\sqrt{W/\nu}}$ where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_\nu$ are independent.

(a) Give $\mathbb{E}(Z)$ and $\mathbb{E}(Z^2)$.

(b) According to the result derived in 6, if Y has a χ^2_ν distribution, then

$$\mathbb{E}(Y^a) = \frac{2^a \Gamma(a + \frac{\nu}{2})}{\Gamma(\frac{\nu}{2})}, \quad \text{if } \nu > -2a.$$

Use this result, the result from (a), and the structure of T to show the following:

(i) $\mathbb{E}(T) = 0$ if $\nu > 1$.

(ii) $\mathrm{Var}(T) = \nu/(\nu - 2)$ if $\nu > 2$.

Proof.

(a) $\mathbb{E}(Z) = 0$ is clear. $\mathrm{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 \implies \mathbb{E}(Z^2) = 1 + 0^2 = 1$.

(b) (i) The expectation $\mathbb{E}(T)$ can be calculated.

$$\begin{aligned} \mathbb{E}(T) &= \mathbb{E}\left(\frac{Z}{\sqrt{W/\nu}}\right) \\ &= \mathbb{E}(Z \cdot g(W)) \quad \text{where } g(W) = \left(\sqrt{W/\nu}\right)^{-1} \text{ is a function of only } W \\ &= \mathbb{E}(Z) \cdot \mathbb{E}(g(W)) \quad \text{by independence of } Z \text{ and } W \\ &\overset{(a)}{=} 0 \cdot \mathbb{E}(g(W)) \\ &= 0 \end{aligned}$$

(ii) The variance of T is given by

$$\begin{aligned}
\operatorname{Var}(T) &:= \mathbb{E}\big(T^2\big) - (\mathbb{E}(T))^2 \\
&\overset{(a)}{=} \mathbb{E}\big(T^2\big) \\
&= \mathbb{E}\Big(\tfrac{Z^2}{W/\nu}\Big) \\
&= \nu\,\mathbb{E}\big(Z^2\big)\mathbb{E}\big(W^{-1}\big) \quad \text{by independence of } Z \text{ and } W \\
&\overset{(a)}{=} \nu\,\mathbb{E}\big(W^{-1}\big) \\
&= \nu\, 2^a\, \frac{\Gamma(a+\frac{\nu}{2})}{\Gamma(\frac{\nu}{2})} \quad \text{where } a = -1 \text{ and } \nu > 2 \\
&= \frac{\nu}{2}\,\frac{\Gamma(\frac{\nu}{2}-1)}{\Gamma(\frac{\nu}{2})} \\
&\overset{(\Gamma)}{=} \frac{\nu}{2}\,\frac{\Gamma(\frac{\nu}{2}-1)}{(\frac{\nu}{2}-1)\,\Gamma(\frac{\nu}{2}-1)} \\
&= \frac{\nu}{2}\,\frac{1}{(\frac{\nu}{2}-1)} \\
&= \frac{\nu}{\nu-2}
\end{aligned}$$

■

In summary, if $T \sim t_\nu$, then T satisfies the following:

- The density of t_ν is symmetric.
- For $\nu > 1$, the expected value of T is 0.
- for $\nu > 2$, the variance of T is $(\nu/2)\frac{\Gamma((\nu/2)-1)}{\Gamma(\nu/2)}$.
    - If ν is an even number greater than 2 (i.e. $\nu \in 2\mathbb{N}_{>1}$), the variance is $\nu/(\nu-2) > 1$.

So we see that the expected value is the same as that of a standard normal distribution but for $\nu > 2$, it has variance greater than $1 = \operatorname{Var}(Z)$ where $Z \sim \mathcal{N}(0,1)$. This means that the t-density has fatter tails than the standard normal density.

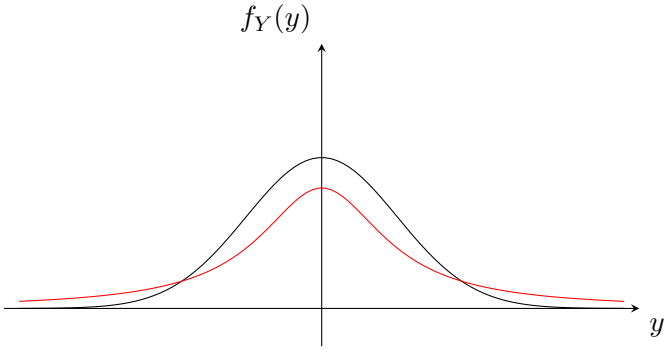


Figure 15.2: The densities of $Z \sim \mathcal{N}(0,1)$ (in black) and $T \sim t_1$ (in red).

15.3 The F-Distribution

Another important derived distribution is Snedcor's F. This distribution arises when comparing the variances of two normal populations based on information contained in two independent samples from each respective population. More precisely, let X and Y be random variables that represent the act of drawing a single value from the populations Ω_1 and Ω_2 with respective normal distributions

$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ respectively. Let X and Y be independent and consider a random sample from each population:

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2).$$

We can estimate σ_X^2 by calculating the sample variance S_X^2 of the random sample $X_1, \ldots, X_n$. Similarly we can estimate σ_Y^2 with S_Y^2 using $Y_1, \ldots, Y_m$. The quantity of interest is σ_X^2/σ_Y^2. It would seem reasonable that the ratio of sample variances

$$\frac{S_X^2}{S_Y^2}$$

could be used to make some inference about the relative magnitudes of the population variances.

From **Theorem 15.1.5** (c), we know that the scaled sample variance $\frac{(n-1)S^2}{\sigma^2}$ follows a χ^2_{n-1} distribution. We can capitalise on this, dividing through by the ratio of population variances σ_X^2/σ_Y^2 to obtain the quantity:

$$\frac{\left(\frac{S_X^2}{\sigma_X^2}\right)}{\left(\frac{S_Y^2}{\sigma_Y^2}\right)}$$

Then we can multiply and divide both the numerator and denominator by $(n-1)$ and $(m-1)$ respectively to get

$$\frac{\frac{(n-1)S_X^2}{\sigma_X^2} \Big/ (n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2} \Big/ (m-1)}$$

Thus:

- The numerator is the quotient of a χ^2_{n-1} distributed random variable by $(n-1)$.
- The denominator is the quotient of a χ^2_{m-1} distributed random variable by $(m-1)$.
- The numerator and denominator are independent.

This reduces the problem to finding the distribution of the following random variable:

Definition 15.3.1 Let W_1 and W_2 be independent χ^2-distributed random variables with ν_1 and ν_2 degrees of freedom respectively. Then the random variable

$$F = \frac{\left(\frac{W_1}{\nu_1}\right)}{\left(\frac{W_2}{\nu_2}\right)}$$

has an $F_{n-1,m-1}$ **distribution with** $(n-1)$ **numerator degrees of freedom, and** $(m-1)$ **denominator degrees of freedom**.

15.3.1 The Density of the F-distribution

Not explicitly given in Wackerly but it is guided through some exercises:

Exercise 9 (*7.73 [6], *7.99 [7]) Let F be defined as before.

(a) If W_2 is fixed at w_2, then $F = W_1/c$ where $c = w_2\nu_1/\nu_2$. Find the conditional density of F given fixed $W_2 = w_2$.

(b) Find the joint density of F and W_2.

(c) Integrate over w_2 to show that the probability density function of F, say $g(x)$, is given by

$$g(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} x^{(\nu_1/2)-1}\left(1+\frac{\nu_1 x}{\nu_2}\right)^{-(\nu_1+\nu_2)/2} \mathbb{1}_{(0,\infty)}(x).$$

Proof.

(a) We know that given $\{W_2 = w_2\}$, $F = \frac{1}{c}W_1$ where $W_1 \sim \chi^2_{\nu_1}$. I'll use the method of moment-generating functions to determine the distribution of F given $\{W_2 = w_2\}$. The MGF of $F = \frac{1}{c}W_1$ is

$$M_F(t) = \mathbb{E}(\exp(tF)) = \mathbb{E}(\exp(tW_1/c)) = M_{W_1}(t/c) = M_{W_1}(t\nu_2/w_2\nu_1)$$

Since a $\chi^2_{\nu_1}$ distributed random variable is also a Gamma$(\nu_1/2, 2)$ distributed random variable, the rightmost MGF is

$$\begin{aligned} M_{W_1}(t\nu_2/w_2\nu_1) &\overset{11.0.5}{=} \frac{1}{(1-\beta(t\nu_2/w_2\nu_1))^\alpha}\Big|_{\alpha=\nu_1/2,\,\beta=2} \\ &= \frac{1}{(1-(2\nu_2/w_2\nu_1)\,t)^{\nu_1/2}} \end{aligned}$$

which is the MGF of a Gamma-distributed random variable with shape parameter $\alpha = \nu_1/2$ and scale parameter $\beta = 2\nu_2/w_2\nu_1$.

$$\therefore\ F \mid \{W_2 = w_2\} \sim \text{Gamma}\left(\frac{\nu_1}{2}, \frac{2\nu_2}{w_2\nu_1}\right).$$

(b) The joint density of F and W_2 splits according to the familiar formula:

$$\begin{aligned} &f_{F,W_2}(x, w_2) \\ &= f_{F|W_2=w_2}(x \mid w_2) \cdot f_{W_2}(w_2) \\ &= \frac{x^{(\nu_1/2)-1}\exp\left(-x\Big/\frac{2\nu_2}{w_2\nu_1}\right)}{\left(\frac{2\nu_2}{w_2\nu_1}\right)^{\nu_1/2}\Gamma\left(\frac{\nu_1}{2}\right)}\mathbb{1}_{(0,\infty)}(x) \cdot \frac{w_2^{(\nu_2/2)-1}\exp(-\frac{w_2}{2})}{2^{\nu_2/2}\Gamma\left(\frac{\nu_2}{2}\right)}\mathbb{1}_{(0,\infty)}(w_2) \\ &= \underbrace{\frac{x^{(\nu_1/2)-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}\mathbb{1}_{(0,\infty)}(x)}_{=\xi(x,\nu_1,\nu_2)} \cdot \frac{w_2^{\frac{\nu_1+\nu_2}{2}-1}\exp\left(-\frac{w_2}{2}\left(1+\frac{x\nu_1}{\nu_2}\right)\right)}{2^{(\nu_1+\nu_2)/2}}\mathbb{1}_{(0,\infty)}(w_2) \end{aligned}$$

(c) Integrate over the joint density to marginalise out W_2 and obtain the probability distribution of F:

$$\begin{aligned} g(x) &= \int_{\mathbb{R}} f_{F,W_2}(x, w_2)\,\mathrm{d}w_2 \\ &= \xi(x,\nu_1,\nu_2)\int_{\mathbb{R}} \frac{w_2^{\frac{\nu_1+\nu_2}{2}-1}\exp\left(-\frac{w_2}{2}\left(1+\frac{x\nu_1}{\nu_2}\right)\right)}{2^{(\nu_1+\nu_2)/2}}\mathbb{1}_{(0,\infty)}(w_2)\,\mathrm{d}w_2 \end{aligned}$$

Now consider the substitution $u = w_2\left(1+\frac{\nu_1 x}{\nu_2}\right) \implies \mathrm{d}w_2 = \left(1+\frac{\nu_1 x}{\nu_2}\right)^{-1}\mathrm{d}u$. When $w_2 = 0$,

$u = 0$. As $w_2 \to +\infty$, $u \to +\infty$ since $x \geqslant 0$, $\nu_1, \nu_2 > 0$. Then:

$$\begin{aligned} g(x) &= \xi(x,\nu_1,\nu_2) \int_0^\infty \frac{\left(u\left(1+\frac{\nu_1 x}{\nu_2}\right)^{-1}\right)^{\frac{\nu_1+\nu_2}{2}-1} \exp(-u/2)}{2^{(\nu_1+\nu_2)/2}} \left(1+\tfrac{\nu_1 x}{\nu_2}\right)^{-1} \mathrm{d}u \\ &= \xi(x,\nu_1,\nu_2)\left(1+\tfrac{\nu_1 x}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}} \int_0^\infty \frac{u^{\frac{\nu_1+\nu_2}{2}-1}\exp(-u/2)}{2^{(\nu_1+\nu_2)/2}}\,\mathrm{d}u \\ &= \xi(x,\nu_1,\nu_2)\left(1+\tfrac{\nu_1 x}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}} \Gamma\left(\tfrac{\nu_1+\nu_2}{2}\right) \underbrace{\int_0^\infty \frac{u^{\frac{\nu_1+\nu_2}{2}-1}\exp(-u/2)}{2^{(\nu_1+\nu_2)/2}\,\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}\,\mathrm{d}u}_{=1} \\ &= \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{(\nu_1/2)-1}\left(1+\tfrac{\nu_1 x}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}} \mathbb{1}_{(0,\infty)}(x) \end{aligned}$$

■

15.3.2 Properties of the F_{ν_1,ν_2}-distribution

Exercise 10 (*7.16 [6], *7.34 [7]) Suppose that W_1 and W_2 are independent χ^2-distributed random variables with ν_1 and ν_2 degrees of freedom, respectively. Let $F = \frac{W_1/\nu_1}{W_2/\nu_2}$. Then F has an F_{ν_1,ν_2}-distribution. Show that:

(a) $\mathbb{E}(F) = \nu_2/(\nu_2-2)$ if $\nu_2 > 2$.

(b) $\mathrm{Var}(F) = \dfrac{2(\nu_2)^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ if $\nu_2 > 4$.

Proof. From 5 and 6, given $Y \sim \chi^2_\nu$ we have that

$$\mathbb{E}(Y^a) = \frac{2^a\Gamma\left(\frac{\nu}{2}+a\right)}{\Gamma\left(\frac{\nu}{2}\right)}, \qquad (\heartsuit)$$

where $\nu > -2a$.

(a)

$$\mathbb{E}(F) \overset{7.1.3}{=} \frac{\nu_2}{\nu_1}\mathbb{E}(W_1)\mathbb{E}\left(W_2^{-1}\right) \overset{(\heartsuit)}{=} \frac{\nu_2}{\nu_1}\frac{2^1\Gamma\left(\frac{\nu_1}{2}+1\right)}{\Gamma\left(\frac{\nu_1}{2}\right)} \cdot \frac{2^{-1}\Gamma\left(\frac{\nu_2}{2}-1\right)}{\Gamma\left(\frac{\nu_2}{2}\right)},$$

where $((\nu_1 > -2) \wedge (\nu_1 > 0))$ and $((\nu_2 > 2) \wedge (\nu_2 > 0))$ i.e. $\nu_1 > 0$ and $\nu_2 > 2$. Using (Γ), we can simplify this expression to

$$\mathbb{E}(F) = \frac{\nu_2}{\nu_1}\frac{\frac{\nu_1}{2}\Gamma\left(\frac{\nu_1}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)} \cdot \frac{\Gamma\left(\frac{\nu_2}{2}-1\right)}{\left(\frac{\nu_2}{2}-1\right)\Gamma\left(\frac{\nu_2}{2}-1\right)} = \frac{\nu_2}{\nu_1}\frac{\frac{\nu_1}{2}}{\left(\frac{\nu_2}{2}-1\right)} = \frac{\nu_2}{\nu_2-2}.$$

(b) Since $\mathrm{Var}(F) = \mathbb{E}\left(F^2\right) - (\mathbb{E}(F))^2$ and $\mathbb{E}(F)$ is known from (a), we focus on the first term:

$$\begin{aligned} \mathbb{E}\left(F^2\right) &\overset{7.1.3}{=} \left(\frac{\nu_2}{\nu_1}\right)^2 \mathbb{E}\left(W_1^2\right)\mathbb{E}\left(W_2^{-2}\right) \\ &\overset{(\heartsuit)}{=} \left(\frac{\nu_2}{\nu_1}\right)^2 \frac{2^2\Gamma\left(\frac{\nu_1}{2}+2\right)}{\Gamma\left(\frac{\nu_1}{2}\right)} \frac{2^{-2}\Gamma\left(\frac{\nu_2}{2}-2\right)}{\Gamma\left(\frac{\nu_2}{2}\right)} \qquad \text{for } \nu_1 > 0 \text{ and } \nu_2 > 4 \\ &= \ldots \\ &= \frac{\nu_2^2(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)} \end{aligned}$$

The claim follows from subtraction.

■

At this point, I began reading Chapter 6 (Principle of Data Reduction) from [1], but I was quickly halted by mention of the conditional distribution of a random variable given another random variable. I was already uncomfortable with how I handled the calculation for the density of the t-distribution in **Exercise 7**, so I added a link to a later chapter offering a rigorous treatment.

The following 4 chapters are dedicated to measurability (thereby defining a statistic), conditional expectation, and conditional probability, culminating in some results that allow one to rigorously speak about conditioning. After this, I'll return to estimation, and data reduction.

CHAPTER 16

More Measurability

We've previously defined a *statistic* (with a view towards estimation) as a "suitably" measurable function T s.t. its composition $T \circ \mathbf{X}$, where $\mathbf{X}$ is a random sample, is measurable. This isn't precise since we haven't mentioned the associated σ-algebras and the subsequent nature of the respective functions' measurability.

This chapter is dedicated to answering the more nuanced question on what "suitably" means i.e. how much can we relax the requirements on T and still maintain that its composition $T \circ X$ with a random element[1] X is still measurable.

Let X be a random element (variable/vector) i.e. $X \in \text{Meas}_{\mathcal{F},\mathcal{E}}(\Omega\,;E)$ where $(E,\mathcal{E})$ is taken to be a Borel space. For the purposes of what follows, it's ok to assume that $(E,\mathcal{E})$ is either $(\mathbb{R},\mathcal{B}_{\mathbb{R}})$ or $(\mathbb{R}^n,\mathcal{B}_{\mathbb{R}^n})$ — both are Borel spaces.

16.1 Theorem A.42

A powerful (and often un-named) theorem characterises when a measurable function is a function of another measurable function (by way of inclusion of their generated σ-algebras):

Theorem 16.1.1 (A.42 [5, p. 587]) Let $(S_1,\mathcal{A}_1)$, $(S_2,\mathcal{A}_2)$, and $(S_3,\mathcal{A}_3)$ be measurable spaces. Suppose further that $\mathcal{A}_3$ contains all singletons from S_3. Let $f \in \text{Meas}_{\mathcal{A}_1,\mathcal{A}_2}(S_1\,;S_2)$ and denote by $\mathcal{A}_*$ the **image σ-field** of f i.e.

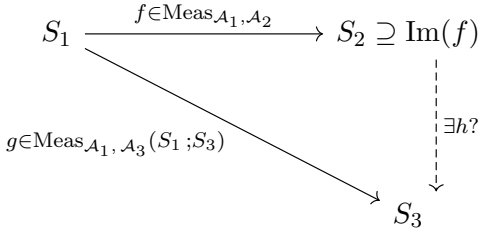
$$\mathcal{A}_* = \{\text{Im}(f) \cap A\colon A \in \mathcal{A}_2\}.$$

Let $g \in \text{Meas}_{\mathcal{A}_1,\mathcal{A}_3}(S_1\,;S_3)$. Then,

$$g \in \text{Meas}_{\sigma(f),\mathcal{A}_3}(S_1\,;S_3) \iff \exists h \in \text{Meas}_{\mathcal{A}_*,\mathcal{A}_3}(\text{Im}(f)\,;S_3) \text{ s.t. } g = h \circ f.$$

Remarks 16.1.2

- Pictorially:



- The condition that $g \in \text{Meas}_{\sigma(f),\mathcal{A}_3}(S_1\,;S_3)$ means that for every $B \in \mathcal{A}_3$, $g^{-1}(B) \in \sigma(f)$ and we know that $\sigma(g) = \{g^{-1}(B)\colon B \in \mathcal{A}_3\}$ so the $\sigma(f)$-measurability of g means that $\sigma(g) \subseteq \sigma(f)$.
- The moral of this representation theorem is that $\sigma(f)$ contains precisely the full/perfect information needed to determine f so any function that's measurable with respect to $\sigma(f)$ can be written purely in terms of f. In the sense of probability, f reveals at least as much information about a random outcome $\omega \in \Omega$ as g does (in the case where f,g are random variables) i.e. knowing f determines g.

[1] Note that we're working in the more general scenario where X is simply a random element, and not a random sample in particular.

- The assumption that $\mathcal{A}_3$ contains all singletons (which in our case will always be satisfied by $(\mathbb{R}, \mathcal{B}_\mathbb{R})$) is to rule out some trivialities like the following example:[2]

  Example Let $S_1 = S_2 = S_3 = \mathbb{R}$, and $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{B}_\mathbb{R}$. Let $\mathcal{A}_3 = \{\emptyset, \mathbb{R}\}$ the trivial σ-field. Every function $g\colon S_1 \to S_3$ is $\sigma(f)$-measurable since

  $$\begin{aligned}\sigma(g) &= \{g^{-1}(A)\colon A \in \mathcal{A}_3\} \\ &= \{g^{-1}(A)\colon A \in \{\emptyset, \mathbb{R}\}\} \\ &= \{\emptyset, \mathbb{R}\}\end{aligned}$$

  and the trivial σ-algebra on S_1 is a subset of every σ-algebra on S_1, and in particular a subset of $\sigma(f)$. However, if we let $g = \mathrm{id}_\mathbb{R}$, then a choice of $f(s) = s^2$ means that g is not a function of f.

Proof.

$\Longleftarrow$ Suppose that $h \in \mathrm{Meas}_{\mathcal{A}_*, \mathcal{A}_3}(\mathrm{Im}(f)\,; S_3)$ is such that $g = h \circ f$. We wish to show that $g \in \mathrm{Meas}_{\sigma(f), \mathcal{A}_3}(S_1\,; S_3)$. Let $A \in \mathcal{A}_3$ and consider its pre-image under g:

$$\begin{aligned}g^{-1}(A) &= (h \circ f)^{-1}(A) \\ &= f^{-1}(\underbrace{h^{-1}(A)}_{\in \mathcal{A}_*}) \\ &= f^{-1}(B \cap \mathrm{Im}(f)) \quad \text{for some } B \in \mathcal{A}_2 \\ &= f^{-1}(B) \in \sigma(f)\end{aligned}$$

$\Longrightarrow$ Assume that $g \in \mathrm{Meas}_{\sigma(f), \mathcal{A}_3}(S_1\,; S_3)$. Since $\mathcal{A}_3$ contains all singletons, the measurability of g tells us that

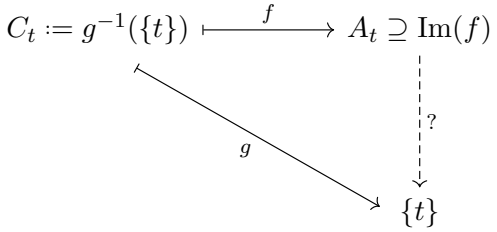
$$C_t \coloneqq g^{-1}(\{t\}) \in \sigma(f).$$

By the definition of $\sigma(f)$, $\exists A_t \in \mathcal{A}_2$ s.t. $C_t = f^{-1}(A_t)$.

$$\therefore\ C_t \coloneqq g^{-1}(\{t\}) = f^{-1}(A_t)$$

Constructing h:

Ultimately, we wish to construct a $h \in \mathrm{Meas}_{\mathcal{A}_*, \mathcal{A}_3}(\mathrm{Im}(f)\,; S_3)$ s.t. $g = h \circ f$. The current situation is



Define $h(s)$ as the unique $t \in S_3$ s.t. $s \in A_t$. (This definition leverages the measurability of f and g through pre-images so it's a natural way to define h.) Note that h is well-defined because the A_t form a partition of $\mathrm{Im}(f)$. More precisely, the $A_t \cap \mathrm{Im}(f)$ form the partition. This is because

- The $A_t \cap \mathrm{Im}(f)$ are certainly pairwise disjoint because if we suppose, for $t \neq t'$ that:

  $$\exists s \in (A_t \cap \mathrm{Im}(f)) \cap (A_{t'} \cap \mathrm{Im}(f)) = A_t \cap A_{t'} \cap \mathrm{Im}(f),$$

[2] and perhaps for some reasons I don't understand yet.

then $\exists a \in S_1$ s.t. $f(a) = s$. Since $s \in A_t$ and $s \in A_{t'}$, we have that

$$\begin{aligned} & a = f^{-1}(A_t) = g^{-1}(\{t\}) \text{ and } a = f^{-1}(A_{t'}) = g^{-1}(\{t'\}) \\ \implies & g(a) = t \text{ and } g(a) = t' \\ \implies & t = t' \end{aligned}$$

which contradicts the assumption that $t \neq t'$.

- Their union covers $\mathrm{Im}(f)$:
  Let $s \in \mathrm{Im}(f)$. Then $\exists a \in S_1$ s.t. $f(a) = s$. Let $t \coloneqq g(a) \in S_3$ i.e. $a \in g^{-1}(\{t\}) = f^{-1}(A_t)$ so $f(a) \in A_t$ i.e. $s \in A_t$. Thus, we've show that for any $s \in \mathrm{Im}(f)$, $s \in A_t$ i.e. $\mathrm{Im}(f) \subseteq \bigcup_{t \in S_3} A_t$ from which it follows that

$$\mathrm{Im}(f) = \left(\bigcup A_t\right) \cap \mathrm{Im}(f) = \bigcup A_t \cap \mathrm{Im}(f).$$

If $t \neq t'$, then $A_t \cap A_{t'} \cap \mathrm{Im}(f) = \emptyset$ so h is well-defined.

Then for any $a \in S_1$, let $t \coloneqq g(a)$. Since $a \in g^{-1}(\{t\})$, then $s \coloneqq f(a) \in A_t$. We've defined $h(s)$ as the unique $t' \in S_3$ s.t. $s \in A_{t'}$. Therefore, $t = t'$ i.e.

$$g(a) = t = h(s) = h(f(a))$$

Therefore, $g = h \circ f$.

Measurability of h:

Let $A \in \mathcal{A}_3$. Then we wish to show that $h^{-1}(A) \in \mathcal{A}_*$ i.e. that there exists some $B \in \mathcal{A}_2$ s.t. $B \cap \mathrm{Im}(f) = h^{-1}(A)$.

Since $g \in \mathrm{Meas}_{\sigma(f)}$, then $g^{-1}(A) \in \sigma(f)$ i.e. $\exists B \in \mathcal{A}_2$ s.t. $g^{-1}(A) = f^{-1}(B)$. We'll show that $h^{-1}(A) = B \cap \mathrm{Im}(f)$ for that particular B:

- $\subseteq$ Let $s \in h^{-1}(A)$ i.e. $h(s) \in A$. Let $t = h(s)$. Then $s = f(x)$ for some $x \in C_t \subseteq g^{-1}(A) = f^{-1}(B)$. This implies that $s = f(x) \in B$ i.e. $s \in B \cap \mathrm{Im}(f)$.
- $\supseteq$ Let $s \in B \cap \mathrm{Im}(f)$. Then $s = f(x)$ for some $x \in f^{-1}(B) = g^{-1}(A)$. This implies that $h(s) = h(f(x)) = g(x) \in A$ i.e. $s \in h^{-1}(A)$.

Therefore, $h^{-1}(A) = B \cap \mathrm{Im}(f)$ which concludes the proof.

■

To re-iterate:

- Given a random variable X, a function Y is $\sigma(X)$-measurable iff it can be written as an appropriately measurable function of X.
- Theorem A.42 really gets at testing to see how small one can make the codomain of f and its corresponding image σ-algebra $\mathcal{A}_*$ when discussing the existence of h and its measurability so that g factors.
- A stronger[3] version of Theorem A.42 is typically stated where $h \in \mathrm{Meas}_{\mathcal{A}_2, \mathcal{A}_3}(S_2 ; S_3)$ instead of the weaker $h \in \mathrm{Meas}_{\mathcal{A}_*, \mathcal{A}_3}(\mathrm{Im}(f) ; S_3)$. With this stronger assumption, the theorem statement is called the Doob-Dynkin Representation Theorem.
  - The Doob-Dynkin *Lemma* is a logically equivalent formulation that flips the perspective: Instead of asking when g factors, we instead begin with f and then ask when composing with h yields a measurable function.

[3] This is stronger because if $h \in \mathrm{Meas}_{\mathcal{A}_2, \mathcal{A}_3}(S_2 ; S_3)$, then $h|_{\mathcal{A}_*} \in \mathrm{Meas}_{\mathcal{A}_*, \mathcal{A}_3}(\mathrm{Im}(f) ; S_3)$.

16.2 Application: Defining a Statistic

In accordance with Theorem A.42:

- $(S_1, \mathcal{A}_1) = (\Omega, \mathcal{F})$
- $(S_2, \mathcal{A}_2) = (E, \mathcal{E})$ is a Borel space
- $(S_3, \mathcal{A}_3) = (S, \mathcal{S})$ is a space for which $\mathcal{S}$ contains all singletons
- $f = X \in \mathrm{Meas}_{\mathcal{F},\mathcal{E}}(\Omega\,; E)$

Let $g \in \mathrm{Meas}_{\mathcal{F},\mathcal{S}}(\Omega\,; S)$. Then the theorem states that

$$g \in \mathrm{Meas}_{\sigma(X),\mathcal{S}}(\Omega\,; S) \iff \exists T \in \mathrm{Meas}_{\mathcal{A}_*,\mathcal{S}}(\mathrm{Im}(X)\,; S) \text{ s.t. } g = T \circ X.$$

The right-hand side of the equivalence presents a (suitably) measurable function of a random element X which inspires the following:

Definition 16.2.1 A **statistic** is a function $T \in \mathrm{Meas}_{\mathcal{A}_*,\mathcal{S}}(\mathrm{Im}(X)\,; S)$ where $\mathcal{A}_*$ is the trace σ-algebra of $\mathcal{E}$ on $X(\Omega)$, and $\mathcal{S}$ contains all singletons.

- Typically, we take the stronger condition that $T \in \mathrm{Meas}_{\mathcal{E},\mathcal{S}}(E\,; S)$ i.e. that T is a map from the entirety of E that is measurable with respect to the entirety of $\mathcal{E}$ on E.

Historically, statisticians considered a statistic to merely be a function of observed data $\mathbf{x}$. What a fantastic time to be alive! Kolmogorov came along and ~~ruined everything~~ formalised statistics and probability. Now we think of σ-algebras as information and inclusion of σ-algebras generated by random variables as function composition.

CHAPTER 17

Conditional Expectation

Conditional expectation is an object (random variable) that captures the idea of the best approximation of a random variable X given partial information with respect to the full/complete information that characterises X.

17.1 Prequel to Abstract Conditional Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

- Given $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, recall the naïve conditional probability measure on $\mathcal{F}$ given B defined in **Definition 2.9.1**. We may also denote $\mathbb{P}(\cdot \mid B)$ by $\mathbb{P}_B$.
  - The idea of conditioning with respect to an event B with $\mathbb{P}(B) > 0$ is tantamount to defining the new probability space $(\Omega, \mathcal{F}, \mathbb{P}_B)$.
- In similar spirit to $\mathbb{E}(X)$, we may use this conditional probability measure given B to define, if it exists, the "average" value that a random variable X takes given the information that the event B occurs with $\mathbb{P}(B) > 0$. This is denoted by $\mathbb{E}_{\mathbb{P}_B}(X)$ and defined by the Lebesgue integral of $X \colon \Omega \to \mathbb{R}$ with respect to the measure $\mathbb{P}(\cdot \mid B)$

$$\mathbb{E}_{\mathbb{P}_B}(X) \coloneqq \int_\Omega X \, \mathrm{d}\mathbb{P}_B.$$

If the integral exists, one calls $\mathbb{E}_{\mathbb{P}_B}(X)$ the **conditional expectation of X given B**.

Lemma 17.1.1 $L^1(\Omega, \mathcal{F}, \mathbb{P}) = L^1(\Omega, \mathcal{F}, \mathbb{P}_B)$ and[1] $\mathbb{E}_{\mathbb{P}_B}(X) = \dfrac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}$.

Proof.

1. If $X = \mathbb{1}_A$ for some $A \in \mathcal{F}$:
$$\int_\Omega X \, \mathrm{d}\mathbb{P}_B = \int_\Omega \mathbb{1}_A \, \mathrm{d}\mathbb{P}_B = \mathbb{P}_B(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{E}(\mathbb{1}_{A\cap B})}{\mathbb{P}(B)} = \frac{\mathbb{E}(\mathbb{1}_A \mathbb{1}_B)}{\mathbb{P}(B)} = \frac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}.$$
2. Let X be a simple, measurable function with standard representation
$$X = \sum_{j=1}^n a_j \mathbb{1}_{A_j}$$
where $\{A_j\}_{j=1}^n$ are pairwise disjoint and $a_j = X^{-1}(A_j)$. Then
$$\begin{aligned}
\int_\Omega X \, \mathrm{d}\mathbb{P}_B &= \int_\Omega \sum_{j=1}^n a_j \mathbb{1}_{A_j} \, \mathrm{d}\mathbb{P}_B \\
&= \sum_{j=1}^n a_j \int_\Omega \mathbb{1}_{A_j} \, \mathrm{d}\mathbb{P}_B \quad \text{by linearity} \\
&= \sum_{j=1}^n a_j \frac{\mathbb{E}(\mathbb{1}_{A_j}\mathbb{1}_B)}{\mathbb{P}(B)} \quad \text{by Step 1} \\
&= \frac{1}{\mathbb{P}(B)} \mathbb{E}\Big(\underbrace{\Big(\sum_{j=1}^n a_j \mathbb{1}_{A_j} \Big)}_{=X} \mathbb{1}_B \Big) \quad \text{by linearity}
\end{aligned}$$

[1] The numerator of this expression can be interpreted as the mean value of X knowing that the event B occurs.

4. Then for any $X : \Omega \to \mathbb{R}$ that is $\mathcal{F}$-measurable, we may decompose $X = X^+ - X^-$ and can find sequences of simple functions $\{X_n^\pm\}_{n\in\mathbb{N}} \uparrow X^\pm$. Note that $|X| = X^+ + X^-$. Then
$$|X| = X^+ + X^- \geqslant X_n^+ + X_n^- = |X_n^+ - X_n^-| =: |X_n|$$
i.e. $|X| \in L^1$ is our dominating function and since we have pointwise convergence, the DCT applies.
$$\therefore \underbrace{\int_\Omega (X_n^+ - X_n^-)\,\mathrm{d}\mathbb{P}_B}^{n\to\infty} \longrightarrow \int_\Omega X\,\mathrm{d}\mathbb{P}_B.$$
Since X_n^+ and X_n^- are simple functions, we can use Step 2 to re-write the limand as:
$$\begin{aligned}
&\int_\Omega (X_n^+ - X_n^-)\,\mathrm{d}\mathbb{P}_B \\
&= \int_\Omega X_n^+\,\mathrm{d}\mathbb{P}_B - \int_\Omega X_n^-\,\mathrm{d}\mathbb{P}_B \\
&= \frac{1}{\mathbb{P}(B)}\left(\int_\Omega X_n^+ \mathbb{1}_B\,\mathrm{d}\mathbb{P} - \int_\Omega X_n^- \mathbb{1}_B\,\mathrm{d}\mathbb{P}\right) \quad \text{by Step 2} \\
&\overset{n\to\infty}{\longrightarrow} \frac{1}{\mathbb{P}(B)}\left(\int_\Omega X^+ \mathbb{1}_B\,\mathrm{d}\mathbb{P} - \int_\Omega X^- \mathbb{1}_B\,\mathrm{d}\mathbb{P}\right) \quad \text{by the MCT} \\
&= \frac{1}{\mathbb{P}(B)}\int_\Omega (X^+ - X^-)\mathbb{1}_B\,\mathrm{d}\mathbb{P} \\
&= \frac{1}{\mathbb{P}(B)}\int_\Omega X\mathbb{1}_B\,\mathrm{d}\mathbb{P} \\
&= \frac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}
\end{aligned}$$
and by the uniqueness of limits in $\mathbb{R}$, these two terms are equal.

■

- An astute observation is that $\mathbb{E}_{\mathbb{P}_{B^\mathsf{c}}}(X)$ may be defined similarly, and so we may form a two-valued step function on Ω:
$$\begin{aligned}
\mathbb{E}[X \mid \{B, B^\mathsf{c}\}](\omega) &:= \begin{cases} \mathbb{E}_{\mathbb{P}_B}(X) & \text{if } \omega \in B \\ \mathbb{E}_{\mathbb{P}_{B^\mathsf{c}}}(X) & \text{if } \omega \in B^\mathsf{c}. \end{cases} \\
&= \mathbb{E}_{\mathbb{P}_B}(X)\mathbb{1}_B + \mathbb{E}_{\mathbb{P}_{B^\mathsf{c}}}(X)\mathbb{1}_{B^\mathsf{c}}
\end{aligned}$$
If either B or B^c is $\mathbb{P}$-null, then the corresponding value e.g. $\mathbb{E}_{\mathbb{P}_{B^\mathsf{c}}}(X)$ is undetermined.

- A natural follow-up of that observation is to ask the question

> "Let's say we know the outcome of an experiment $\omega \in \Omega$ lies in any one of a collection of sets $\mathcal{P} = \{B_n\}_{n\in\mathbb{N}}$ partitioning the outcome space Ω. What is the expected value of an *integrable* random variable $X : \Omega \to \mathbb{R}$ at such an ω?"

The immediate solution is to generalise our two-valued function and construct a map that gives the "local average" of X given coarse[2] information $\mathcal{P}$ where an outcome of the corresponding experiment lies in Ω. This comes in the form of a step function that takes on the numerical value $\mathbb{E}_{\mathbb{P}_{B_n}}(X)$ on each B_n. Denote this function by $\mathbb{E}[X \mid \mathcal{G}]$ where $\mathcal{G} = \sigma(\mathcal{P})$, call

[2] I'm guessing this'll be refined later.

it the **conditional expectation of X relative to the partition $\mathcal{P}$**, and define it as the function

$$\begin{aligned}\omega \longmapsto \mathbb{E}[X \,|\, \sigma(\mathcal{P})](\omega) &:= \sum_{n\in\mathbb{N}} \mathbb{E}_{\mathbb{P}_{B_n}}(X)\mathbb{1}_{B_n}(\omega) \\ &= \sum_{n\in\mathbb{N}} \frac{1}{\mathbb{P}(B_n)} \left(\int_{B_n} X \,\mathrm{d}\mathbb{P}\right) \mathbb{1}_{B_n}(\omega).\end{aligned}$$

- This simple function is clearly $\sigma(\mathcal{P})$-measurable so $\mathbb{E}[X \,|\, \mathcal{P}]$ is a random variable.
- The same consideration as before holds. Namely, if some of the B_n are $\mathbb{P}$-null then their respective corresponding values $\mathbb{E}_{\mathbb{P}_{B_n}}(X)$ are undetermined.

For the eagle-eyed reader, you may notice that I'm using slightly different parentheses. For any function pertaining to probability/expectation:

- I will use regular parentheses to denote functions that output a number e.g. $\mathbb{P}(A)$ and $\mathbb{E}(X)$.
- I will use square parentheses to denote random variables e.g. $\mathbb{E}[X \,|\, \mathcal{P}]$ and, as we shall see in the next section, $\mathbb{P}[A \,|\, \mathcal{G}]$.

The distinction is important for me in that it helps me to keep track of which objects are being considered.

- In the spirit of local averages, one can investigate how $\mathbb{E}[X \,|\, \mathcal{P}]$ averages over some $A \in \sigma(\mathcal{P}) =: \mathcal{G}$, and how this compares to X.

  Any $A \in \mathcal{G}$ is of the form $A = \bigsqcup_{n\in J} B_n$ where $J \subseteq \mathbb{N}$. Then:

$$\begin{aligned}\int_A \mathbb{E}[X \,|\, \mathcal{P}] \,\mathrm{d}\mathbb{P} &= \sum_{n\in J} \frac{1}{\mathbb{P}(B_n)} \left(\int_{B_n} X \,\mathrm{d}\mathbb{P}\right) \underbrace{\mathbb{P}(B_n \cap A)}_{=\mathbb{P}(B_n)} \\ &= \sum_{n\in J} \int_{B_n} X \,\mathrm{d}\mathbb{P} \\ &= \int_A X \,\mathrm{d}\mathbb{P} \quad \text{by linearity.}\end{aligned}$$

  This equality can be re-written as

$$\mathbb{E}(\mathbb{1}_A \mathbb{E}[X \,|\, \mathcal{P}]) = \mathbb{E}(\mathbb{1}_A X) \quad \text{for all } A \in \sigma(\mathcal{P}) =: \mathcal{G}$$

  so over every event "within the purview" of the partition (i.e. within $\sigma(\mathcal{P})$), the conditional expectation relative to the partition averages out to the same value as X.

Le Gall follows a slightly different path (with a different emphasis) when building up to an abstract definition of conditional expectation.

- We can define the conditional expectation of $X \in L^1$ given a discrete random variable Y:

  Consider $Y : \Omega \to E$ where E is countable and equipped with the discrete σ-algebra 2^E. Let $E' = \{y \in E : \mathbb{P}(\{Y = y\}) > 0\}$. For $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, we may consider for every $y \in E'$:

$$\mathbb{E}(X \,|\, \{Y = y\}) = \frac{\mathbb{E}\big(X\mathbb{1}_{\{Y=y\}}\big)}{\mathbb{P}(\{Y = y\})}$$

as a special case of $\mathbb{E}_{\mathbb{P}_B}(X)$ with $B = \{Y = y\}$ s.t. $\mathbb{P}(B) > 0$. Now we can use this to define the conditional expectation of X given Y-discrete as the real random variable

$$\mathbb{E}[X \,|\, Y] = \varphi(Y)$$

where $\varphi\colon E \to \mathbb{R}$ is defined by[3]

$$\varphi(y) = \begin{cases} \mathbb{E}(X \,|\, \{Y = y\}) & \text{if } y \in E' \\ 0 & \text{if } y \in E \setminus E'. \end{cases}$$

Note that by the Doob-Dynkin Theorem, $\mathbb{E}(X \,|\, Y)$ is a $\sigma(Y)$-measurable function.

Proposition 17.1.2 (11.2 [3]) Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. We have that $\mathbb{E}(|\mathbb{E}[X \,|\, Y]|) \leqslant \mathbb{E}(X)$ and thus $\mathbb{E}(X \,|\, Y) \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, for every bounded $\sigma(Y)$-measurable real random variable W:

$$\mathbb{E}(WX) = \mathbb{E}(W\mathbb{E}[X \,|\, Y]).$$

Proof. From the definition of $\mathbb{E}[X \,|\, Y]$, we have that

$$\begin{aligned} \mathbb{E}(|\mathbb{E}[X \,|\, Y]|) &= \sum_{y \in E'} \mathbb{P}(\{Y = y\}) \frac{|\mathbb{E}\big(X \mathbb{1}_{\{Y=y\}}\big)|}{\mathbb{P}(\{Y = y\})} \\ &\leqslant \sum_{y \in E} \mathbb{E}\big(|X| \mathbb{1}_{\{Y=y\}}\big) \\ &= \mathbb{E}\left(|X| \sum_{y \in E} \mathbb{1}_{\{Y=y\}}\right) \\ &= \mathbb{E}(|X| \mathbb{1}_\Omega) \quad \text{since } \bigsqcup_{y \in E} \{Y = y\} = Y^{-1}(E) = \Omega \\ &= \mathbb{E}(|X|) \end{aligned}$$

If W is $\sigma(Y)$-measurable (and bounded), then we can find a (bounded) function $\psi\colon E \to \mathbb{R}$

[3]The choice of value on $E \setminus E'$ is irrelevant because it only influences $\mathbb{E}(X \,|\, Y)$ on a set of probability zero since

$$\mathbb{P}\big(\{Y \in E \setminus E'\}\big) = \sum_{y \in E \setminus E'} \mathbb{P}(\{Y = y\}) = 0.$$

s.t. $W = \psi(Y)$ by Doob-Dynkin. It follows that

$$\begin{aligned}
&\mathbb{E}(\psi(Y)\mathbb{E}[X \,|\, Y]) \\
&= \sum_{y\in E} \psi(y)\varphi(y)\mathbb{P}(\{Y=y\}) \\
&= \sum_{y\in E'} \psi(y)\varphi(y)\mathbb{P}(\{Y=y\}) \quad \text{since } \forall y \in E\setminus E' \colon \mathbb{P}(\{Y=y\}) = 0 \\
&= \sum_{y\in E'} \psi(y)\frac{\mathbb{E}\big(X\mathbb{1}_{\{Y=y\}}\big)}{\mathbb{P}(\{Y=y\})}\mathbb{P}(\{Y=y\}) \\
&= \sum_{y\in E'} \psi(y)\mathbb{E}\big(X\mathbb{1}_{\{Y=y\}}\big) \\
&= \sum_{y\in E'} \mathbb{E}\big(\psi(Y)X\mathbb{1}_{\{Y=y\}}\big) \qquad \text{y becomes Y inside of the expectation because $\mathbb{1}_{\{Y=y\}}$ is present} \\
&= \sum_{y\in E'} \mathbb{E}\big(\psi(Y)X\mathbb{1}_{\{Y=y\}}\big) + \sum_{y\in E\setminus E'} \mathbb{E}\big(\psi(Y)X\mathbb{1}_{\{Y=y\}}\big) \qquad \text{this term is zero because if } \mathbb{P}(A)=0 \text{ then for any } X\in L^1\colon \mathbb{E}(X\mathbb{1}_A)=0 \\
&\overset{\leftrightarrow}{=} \mathbb{E}\big(\psi(Y)X\mathbb{1}_{Y^{-1}(E')}\big) + \mathbb{E}\big(\psi(Y)X\mathbb{1}_{Y^{-1}(E\setminus E')}\big) \\
&= \mathbb{E}\big(\psi(Y)X(\mathbb{1}_{Y^{-1}(E')} + \mathbb{1}_{Y^{-1}(E\setminus E')})\big) \\
&= \mathbb{E}(\psi(Y)X\mathbb{1}_{\Omega}) \\
&= \mathbb{E}(\psi(Y)X)
\end{aligned}$$

where the expression indicated by $\leftrightarrow$ has an an exchange of integral and sum by Fubini's theorem. ■

We already know that, for Y-discrete, $\mathbb{E}[X \,|\, Y]$ is $\sigma(Y)$-measurable. The above proposition establishes that it is also integrable, and averages to the same value when integrated against bounded $\sigma(Y)$ measurable functions i.e. when it comes to the information that $\sigma(Y)$ provides about X, the random variable $\mathbb{E}[X \,|\, Y]$ agrees with X on average.

Corollary 17.1.3 If Y' is another discrete random variable such that $\sigma(Y) = \sigma(Y')$, then the following equality holds $\mathbb{P}$-a.s.

$$\mathbb{E}[X \,|\, Y] = \mathbb{E}\big[X \,\big|\, Y'\big].$$

Proof. Apply the previous proposition with $W = \mathbb{1}_{\{\mathbb{E}[X\,|\,Y]>\mathbb{E}[X\,|\,Y']\}}$ which is measurable[4] with respect to $\sigma(Y) = \sigma(Y')$. This gives us two equalities

$$\begin{cases} \mathbb{E}(WX) = \mathbb{E}(W\mathbb{E}[X \,|\, Y]) \\ \mathbb{E}(WX) = \mathbb{E}(W\mathbb{E}[X \,|\, Y']) \end{cases}$$

from which it follows that

$$0 = \mathbb{E}\big(W(\mathbb{E}[X \,|\, Y] - \mathbb{E}\big[X \,\big|\, Y'\big])\big) = \mathbb{E}\big(\mathbb{1}_{\{\mathbb{E}[X\,|\,Y]>\mathbb{E}[X\,|\,Y']\}}(\mathbb{E}[X \,|\, Y] - \mathbb{E}\big[X \,\big|\, Y'\big])\big).$$

This is possible if $\mathbb{1}_{\{\mathbb{E}[X\,|\,Y]>\mathbb{E}[X\,|\,Y']\}} = 0$ almost surely so we may conclude that $\mathbb{E}[X \,|\, Y] \leqslant \mathbb{E}[X \,|\, Y']$ almost surely. By interchanging the roles of Y and Y', we obtain the reverse inequality. Hence, $\mathbb{E}[X \,|\, Y] = \mathbb{E}[X \,|\, Y']$ almost surely. ■

This corollary tells us that the information encoded in[5] the σ-algebra $\sigma(Y)$ is what matters, not the functional form of the random variable Y (or Y') that generates it.

[4] Note that $\{\mathbb{E}[X \,|\, Y] > \mathbb{E}[X \,|\, Y']\} = (\mathbb{E}[X \,|\, Y'] - \mathbb{E}[X \,|\, Y])^{-1}((-\infty, 0))$ and the difference of two $\sigma(Y) = \sigma(Y')$-measurable functions is $\sigma(Y)$-measurable, so the pre-image of $(-\infty, 0) \in \mathcal{B}_{\mathbb{R}}$ is in $\sigma(Y)$, and $\mathbb{1}_{\{\mathbb{E}[X\,|\,Y]>\mathbb{E}[X\,|\,Y']\}}$ is $\sigma(Y)$-measurable.

[5] i.e. everything knowable from Y.

17.2 Abstract Conditional Expectation ($X \in L^1$)

Why is it important that we established the idea of conditional expectation depends not on the functional form of the random variable, but the σ-algebra that it generates? Two reasons, the first of which is staring us in the face:

- The definition of $\mathbb{E}[X \mid Y]$ above is for discrete Y; it breaks down when the events $\{Y = y\}$ no longer have positive probability. We cannot divide by zero.
- The formalisation of probability via measure theory shifts the perspective from conditioning on events like $\{Y = y\}$ to conditioning on (sub-)σ-algebras — the mathematical structure that "information" naturally presents itself as i.e. σ-algebras can be considered to be an information variable.
    - In particular, conditioning on a sub-σ-algebra allows one to accommodate for "conditioning on absolutely continuous random variables — the scenario where simple events $\{Y = y\}$ have zero probability.

  Note that one may condition on ***any*** sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, and in this framework a sub-σ-algebra represents **partial information**.

The averaging property in Proposition 11.2 [3] uses bounded, $\sigma(Y)$-measurable functions

$$\mathbb{E}(WX) = \mathbb{E}(W\mathbb{E}[X \mid Y])$$

but an equivalent formulation can be written in terms of indicator functions $\mathbb{1}_B$ where $B \in \sigma(Y)$. This is because any bounded, $\sigma(Y)$-measurable function W may be approximated by a sequence of simple functions which themselves are finite linear combinations of such indicators $\mathbb{1}_B$.

Definition 17.2.1 Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra. Then *the* **conditional expectation of X given $\mathcal{G}$** is **any** random variable s.t.

1. Z is $\mathcal{G}$-measurable.
2. For all $A \in \mathcal{G}$:
$$\mathbb{E}(\mathbb{1}_A Z) = \mathbb{E}(\mathbb{1}_A X).$$

Remarks 17.2.2

- Think of $\mathcal{F}$ as the total information needed to know/characterise X.
- Condition 1 says that $\forall B \in \mathcal{B}_{\mathbb{R}}$, $Z^{-1}(B) \in \mathcal{G}$ i.e. knowing $\mathcal{G}$ is enough to fully determine Z.
- Condition 2 is the averaging property which tells us that for events knowable from $\mathcal{G}$, on average X and Z have the same value.
    - Combining these two intuitive properties, we have that given partial information $\mathcal{G}$ about X, then Z is "the" best possible $\mathcal{G}$-measurable approximation of X.
- All mentions of "the" conditional expectation are statements that are true $\mathbb{P}$-almost surely i.e. if Z_1 and Z_2 satisfy 1 and 2, then
$$\mathbb{P}(\{\omega \in \Omega : Z_1(\omega) \neq Z_2(\omega)\}) = 0.$$
  Thus, we call **any** such Z a **version** of $\mathbb{E}[X \mid \mathcal{G}]$. This uniqueness comes from the Radon-Nikodým Theorem:

Does such a Z exist? Is it unique?

Theorem 17.2.3 $\exists! Z$ $\mathbb{P}$-a.s. satisfying 1 and 2.

Proof.

1. Assume that $X \geqslant 0$.

   Define $\nu \colon \mathcal{G} \to [0, +\infty]$ for all $A \in \mathcal{G}$ by

   $$\begin{aligned}\nu(A) &:= \int_A X \, \mathrm{d}\mathbb{P} \\ &= \int_\Omega X \mathbb{1}_A \, \mathrm{d}\mathbb{P} =: \mathbb{E}(X\mathbb{1}_A).\end{aligned}$$

   We've already shown that ν is indeed a non-negative measure (via the MCT). By construction, for every $A \in \mathcal{G}$: $\mathbb{P}(A) = 0 \implies \nu(A) = 0$ i.e. $\nu \ll \mathbb{P}$. If both measures are σ-finite, we're in a situation where we can apply the Radon-Nikodym Theorem. Note that both measures are in fact finite (which implies σ-finiteness):

   - $\mathbb{P}\big|_{\mathcal{G}}(\Omega) = \mathbb{P}(\Omega) = 1$
   - $\nu(\Omega) := \displaystyle\int_\Omega X \, \mathrm{d}\mathbb{P} =: \mathbb{E}(X) < \infty$ since $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.

   Thus, by the Radon-Nikodym Theorem there exists a $\mathbb{P}$-a.s.-unique non-negative measurable function $\phi \colon \Omega \to [0, +\infty]$ s.t. for every $A \in \mathcal{G}$:

   $$\nu(A) = \int_A \phi \, \mathrm{d}\mathbb{P},$$

   the left-hand side of which is $\mathbb{E}(\mathbb{1}_A X)$ and the right-hand side is $\mathbb{E}(\mathbb{1}_A \phi)$. Thus, the Radon-Nikodym dervative ϕ (also denoted $\frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}}$) of ν with respect to $\mathbb{P}$ is a function Z s.t. 1 and 2 are satisfied in the case of $X \geqslant 0$.

2. In the general case $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$, we may write $X = X^+ - X^-$. The first part of this proof for $X \geqslant 0$ applies to the positive and negative parts respectively i.e. there exist $\mathbb{P}$-a.s. unique, non-negative, $\mathcal{G}$-measurable functions Z^+ and Z^- satisfying 1 and 2. It's clear that for all $A \in \mathcal{G}$:

   $$\begin{aligned}\mathbb{E}(\mathbb{1}_A X) &:= \mathbb{E}\big(\mathbb{1}_A (X^+ - X^-)\big) \\ &= \mathbb{E}\big(\mathbb{1}_A X^+\big) - \mathbb{E}\big(\mathbb{1}_A X^-\big) \quad \text{by linearity} \\ &= \mathbb{E}\big(\mathbb{1}_A Z^+\big) - \mathbb{E}\big(\mathbb{1}_A Z^-\big) \quad \text{by 2} \\ &= \mathbb{E}\big(\mathbb{1}_A (Z^+ - Z^-)\big) \quad \text{by linearity} \\ &=: \mathbb{E}(\mathbb{1}_A Z)\end{aligned}$$

   and $Z = Z^+ - Z^-$ is $\mathcal{G}$-measurable (as the difference of two $\mathcal{G}$-measurable functions).

Thus concludes the proof that the conditional expectation of a random variable $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ given a sub-σ-algebra $\mathcal{G}$ exists and is $\mathbb{P}$-a.s. unique. ■

Definition 17.2.4 Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and Y be a random variable. The **conditional expectation $\mathbb{E}[X \,|\, Y]$ of X given Y** is defined to be $\mathbb{E}[X \,|\, \sigma(Y)]$, where $\sigma(Y)$ is the smallest σ-algebra that makes Y measurable.

17.3 Properties of Conditional Expectation ($X \in L^1$)

Since the conditional expectation is a random variable itself, we may compute its expectation. Recall the averaging property that $\mathbb{E}[X \,|\, \mathcal{G}]$ must satisfy:

$$\mathbb{E}(\mathbb{1}_A X) = \mathbb{E}(\mathbb{1}_A \mathbb{E}[X \,|\, \mathcal{G}]).$$

Since $\mathcal{G}$ is a σ-algebra in its own right, $\Omega \in \mathcal{G}$ so take $A = \Omega \in \mathcal{G}$ to obtain

$$\underbrace{\mathbb{E}(\mathbb{1}_\Omega X)}_{\mathbb{E}(X)} = \underbrace{\mathbb{E}(\mathbb{1}_\Omega \mathbb{E}[X \,|\, \mathcal{G}])}_{\mathbb{E}(\mathbb{E}[X \,|\, \mathcal{G}])}.$$

What we've shown is an important result that has its own name if we consider $\mathcal{G}$ to be a σ-algebra generated by some random variable Y i.e. $\mathcal{G} = \sigma(Y)$.

Theorem 17.3.1 (The **Tower Law** or **Law of Total Expectation**) Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and Y any random variable defined on the same space. Then

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}[X \,|\, Y]),$$

where $\mathbb{E}[X \,|\, Y] \coloneqq \mathbb{E}[X \,|\, \sigma(Y)]$.

The conditional expectation inherits some properties from the unconditional expectation:

- The function $X \longmapsto \mathbb{E}[X \,|\, \mathcal{G}]$ is linear and non-decreasing.
- Monotone convergence i.e. if $X_n \uparrow X$, then

$$\lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}] = \mathbb{E}[X \,|\, \mathcal{G}].$$

- Dominated convergence i.e. if $X_n \to X$ and $\forall n \colon |X_n| \leqslant Y \in \mathcal{L}^1$, then

$$\lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}] = \mathbb{E}[X \,|\, \mathcal{G}].$$

- Perfect information i.e. if $X \in \mathcal{G}$, then $\mathbb{E}[X \,|\, \mathcal{G}] = X$.

  Proof. Since $X \in \mathcal{G}$, all that remains to show is that X satisfies the averaging property. For every $A \in \mathcal{G}$, $\mathbb{E}(\mathbb{1}_A X) = \mathbb{E}(\mathbb{1}_A Z)$ is trivially satisfied when $Z = X$. ■

  Intuition: X being $\mathcal{G}$-measurable means that knowing $\mathcal{G}$ is enough information to determine which values X takes because $\forall B \in \mathcal{B}_\mathbb{R}$, $X^{-1}(A) \in \mathcal{G}$. On the other hand, if $\mathcal{G}$ doesn't give you perfect information about X, then $\exists B \in \mathcal{B}_\mathbb{R}$ s.t. $X^{-1}(B) \notin \mathcal{G}$ i.e. there is some event related to X for which $\mathcal{G}$ cannot encode/tell us about i.e. $X \notin \mathcal{G}$.

- No information relevant to understanding/guessing X i.e. If X and $\mathcal{G}$ are independent, then $\mathbb{E}[X \,|\, \mathcal{G}] = \mathbb{E}(X)$.

  Intuition: The best possible approximation is the "worst" thing we can imagine — a constant function equal to the unconditional expectation. With no relevant information about X, all we can do is average X and return that number as our best guess.

  Proof.

  1. Let $A \in \mathcal{B}_\mathbb{R}$. Then denote the constant function with value $\mathbb{E}(X)$ by f.

$$f^{-1}(A) = \begin{cases} \emptyset & \text{if } \mathbb{E}(X) \notin A \\ \Omega & \text{if } \mathbb{E}(X) \in A \end{cases}$$

     Both $\emptyset \in \mathcal{G} \ni \Omega$ since $\mathcal{G}$ is a σ-algebra. Thus, all constant functions are $\mathcal{G}$-measurable. In particular, $f \equiv \mathbb{E}(X)$ is.

2. Suppose that X and $\mathcal{G}$ are independent. Let $A \in \mathcal{G}$. Then $\mathbb{1}_A$ is a random variable independent of X. It can be shown by Fubini's theorem that if X, Y are independent random variables, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Thus,

$$\begin{aligned}\mathbb{E}(X\mathbb{1}_A) &= \mathbb{E}(\mathbb{1}_A)\mathbb{E}(X) \quad \text{by Fubini} \\ &= \mathbb{E}(\mathbb{1}_A\mathbb{E}(X)) \quad \text{by linearity}\end{aligned}$$

i.e. condition 2 of the definition of the conditional expectation is satisfied by $Z = \mathbb{E}(X)$.

■

- No information (at all) is the case where $\mathcal{G} = \{\emptyset, \Omega\}$. The conclusion is the same. Namely, $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}(X)$.
- Assume that $X \in \mathcal{G}$ (i.e. that we have perfect information for X. Then:
  - $\mathbb{E}[X + Y \mid \mathcal{G}] = X + \mathbb{E}[Y \mid \mathcal{G}]$
  - $\mathbb{E}[XY \mid \mathcal{G}] = X\mathbb{E}[Y \mid \mathcal{G}]$

Proof.

- Observe that

$$\begin{aligned}\mathbb{E}[X + Y \mid \mathcal{G}] &= \mathbb{E}[X \mid \mathcal{G}] + \mathbb{E}[Y \mid \mathcal{G}] \quad \text{by linearity} \\ &= X + \mathbb{E}[Y \mid \mathcal{G}] \quad \text{by perfect information.}\end{aligned}$$

- Restate the definition of the conditional expectation of XY given a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ i.e. any random variable Z s.t.
  1. $Z \in \mathcal{G}$
  2. $\forall A \in \mathcal{G}\colon \mathbb{E}(\mathbb{1}_A XY) = \mathbb{E}(\mathbb{1}_A Z)$

  Let $Z = X\mathbb{E}[Y \mid \mathcal{G}]$. We wish to verify this Z satisfies 1 and 2:
  1. By definition, $\mathbb{E}[Y \mid \mathcal{G}]$ is $\mathcal{G}$-measurable, and by assumption $X \in \mathcal{G}$. Their product is therefore also $\mathcal{G}$-measurable.
  2. We wish to prove that $\forall A \in \mathcal{G}$:

$$\mathbb{E}(\mathbb{1}_A XY) = \mathbb{E}(\mathbb{1}_A \underbrace{X\mathbb{E}[Y \mid \mathcal{G}]}_{Z}).$$

Proof Sketch. Follow the steps of the construction of the Lebesgue integral:

2.1) Prove first for $X = \mathbb{1}_B$
2.2) Linearity implies the case for simple functions
2.3) Conditional expectation satisfies the MCT so it follows that the averaging property holds for non-negative X
2.4) Linearity $\implies$ true for any random variable $X = X^+ - X^-$.

■

Proof of 2.1. The statement of the theorem assumes $X \in \mathcal{G}$. In particular, $X = \mathbb{1}_B \in \mathcal{G} \iff B \in \mathcal{G}$. Then

$$\begin{aligned}\mathbb{E}(\mathbb{1}_A X\mathbb{E}[Y \mid \mathcal{G}]) &\coloneqq \mathbb{E}(\mathbb{1}_A\mathbb{1}_B\mathbb{E}[Y \mid \mathcal{G}]) \\ &= \mathbb{E}(\mathbb{1}_{A\cap B}\mathbb{E}[Y \mid \mathcal{G}]) \\ &= \mathbb{E}(\mathbb{1}_{A\cap B}Y) \quad \text{by property 2 of } \mathbb{E}[Y \mid \mathcal{G}] \\ &= \mathbb{E}(\mathbb{1}_A\mathbb{1}_B Y) \\ &=: \mathbb{E}(\mathbb{1}_A XY)\end{aligned}$$

■

■

In practice, we often don't have enough information (perfect information) to calculate $\mathbb{E}(X)$ directly (or the direct computation is difficult). To make the computation easier, it helps to introduce another random variable Y that's simpler than X and gives partial information "$\mathcal{G}$" about X and note the **Theorem 17.3.1** tells us that

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}[X \mid Y]).$$

The following propositions and lemmata will recover some important formulae discussed in undergraduate probability.

As a preliminary factoid, we consider what measurable functions look like when the σ-algebra in question is generated by a partition $\{B_n\}$ of the outcome space Ω:

Lemma 17.3.2 Let $Z\colon \Omega \to \mathbb{R}$ be $\mathcal{G}$-measurable, where $\mathcal{G} = \sigma(\{B_n\})$. This is equivalent to Z being constant on each B_n. In this case, Z may be written as a sum

$$Z = \sum a_i \mathbb{1}_{B_i}.$$

Proof. Assume that Z is constant on each of the B_n and can therefore be written as a sum $Z = \sum_n a_n \mathbb{1}_{B_n}$. We wish to show that $Z \in \mathcal{G}$. Let $A \in \mathcal{B}_{\mathbb{R}}$ and consider $Z^{-1}(A)$. Fix some n_0. By definition, $\forall \omega \in B_{n_0}\colon Z(\omega) = a_{n_0}$ i.e. $Z^{-1}(\{a_{n_0}\}) = B_{n_0}$. B_{n_0} is disjoint from the other cells. Therefore, we may write

$$Z^{-1}(A) = \Big(\bigsqcup_{n\colon a_n \in A} B_n \Big)$$

which is an element of $\mathcal{G}$ by closure under countable unions.

For the reverse implication, assume that Z is $\mathcal{G}$-measurable i.e. $\forall A \in \mathcal{B}_{\mathbb{R}}$, $Z^{-1}(A) \in \mathcal{G}$. Since $\mathcal{G} = \sigma(\{B_n\})$, $Z^{-1}(A)$ may be written as a disjoint union of cells. In particular, $Z^{-1}(A) \cap B_n$ must be either B_n or $\emptyset$. Fix a particular n_0 and suppose that Z is non-constant on B_{n_0} i.e. $\exists \omega_1, \omega_2 \in B_{n_0}$ s.t. $\omega_1 \neq \omega_2$ and WLOG[6] $Z(\omega_1) < Z(\omega_2)$. There exists some a s.t. $Z(\omega_1) < a < Z(\omega_2)$. Consider $(-a, \infty] \in \mathcal{B}_{\mathbb{R}}$. By measurability of Z

$$Z^{-1}(A) = \bigsqcup_{n \in J} B_n.$$

Since B_{n_0} is a cell in the partition of Ω (that generates $\mathcal{G}$):

$$Z^{-1}(A) \cap B_{n_0} = \begin{cases} B_{n_0} & \text{if } B_{n_0} \subseteq Z^{-1}(A) \\ \emptyset & \text{if } B_{n_0} \cap Z^{-1}(A) = \emptyset \end{cases}$$

Let's examine what happens to ω_1 and ω_2:

- Since $Z(\omega_1) < a$, $\omega_1 \in Z^{-1}((-\infty, a]) = Z^{-1}(A)$. It was given that $\omega_1 \in B_{n_0}$.

$$\therefore\ \omega_1 \in Z^{-1}(A) \cap B_{n_0} \implies Z^{-1}(A) \cap B_{n_0} \neq \emptyset.$$

- On the other hand, $Z(\omega_2) > a$ so $\omega_2 \notin Z^{-1}(A)$. Also, $\omega_2 \in B_{n_0}$ is given.

$$\therefore\ \omega_2 \notin Z^{-1}(A) \cap B_{n_0} \implies Z^{-1}(A) \cap B_{n_0} \underset{\neq}{\subseteq} B_{n_0}.$$

Thus, we've found a non-trivial proper subset of B_{n_0} that is an element of $\mathcal{G}$. The existence of such a set $Z^{-1}(A) \cap B_{n_0}$ contradicts the fact that $\mathcal{G} = \sigma(\{B_n\})$ from which we deduced that the intersection of any set in $\mathcal{G}$ with any atom is either empty or the atom itself.

Therefore, we're forced to conclude our assumption that Z is non-constant on B_{n_0} is false. Since n_0 was arbitrary, we conclude that Z is constant on each atom. ■

[6] Without loss of generality.

The above lemma will be used in the following proposition for its representation of a $\mathcal{G}$-measurable function (where $\mathcal{G} = \sigma(\{B_n\})$) as a sum of indicators:

Proposition 17.3.3 Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that $\mathcal{G} = \sigma(\{B_n\})$ is a sub-σ-algebra of $\mathcal{F}$. Then

$$\mathbb{E}[X \,|\, \mathcal{G}] = \sum_n \mathbb{E}(X \,|\, B_n)\mathbb{1}_{B_n}.$$

Proof. $Z = \mathbb{E}[X \,|\, \mathcal{G}]$ is well-defined and exists $\mathbb{P}$-a.s. uniquely by the Radon-Nikodym Theorem. Property 1 of conditional expectation tells us that $Z \in \mathcal{G}$ and by our prior lemma, it has the representation

$$\mathbb{E}[X \,|\, \mathcal{G}] = Z = \sum_{n=1}^{\infty} a_n \mathbb{1}_{B_n}.$$

The averaging property of conditional expectation tells us that Z satisfies, for every $A \in \mathcal{G}$:

$$\mathbb{E}(\mathbb{1}_A X) = \mathbb{E}(\mathbb{1}_A Z).$$

In particular, for any $B_n \in \mathcal{G} = \sigma(\{B_n\})$:

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{B_n} X) &= \mathbb{E}(\mathbb{1}_{B_n} Z) \\
&= \mathbb{E}\left(\sum_{i=1}^{\infty} a_i \mathbb{1}_{B_i} \mathbb{1}_{B_n}\right) \\
&= \mathbb{E}(a_n \mathbb{1}_{B_n}) \\
&= a_n \mathbb{E}(\mathbb{1}_{B_n}) \quad \text{by linearity} \\
&= a_n \mathbb{P}(B_n)
\end{aligned}$$

i.e. $a_n = \dfrac{\mathbb{E}(X\mathbb{1}_{B_n})}{\mathbb{P}(B_n)}$. We conclude by writing

$$\mathbb{E}[X \,|\, \mathcal{G}] = Z = \sum_{n=1}^{\infty} a_n \mathbb{1}_{B_n} = \sum_{n=1}^{\infty} \frac{\mathbb{E}(X\mathbb{1}_{B_n})}{\mathbb{P}(B_n)} \mathbb{1}_{B_n},$$

where we recover the coefficient of $\mathbb{1}_{B_n}$ as the **conditional expectation of the random variable X given an event B_n.** ■

Example In the case that $\mathcal{G} = \sigma(B) = \{\emptyset, B, B^{\mathsf{c}}, \Omega\}$, the proposition we just proved tells us that

$$\mathbb{E}[X \,|\, \sigma(B)] = a_1 \mathbb{1}_B + a_2 \mathbb{1}_{B^{\mathsf{c}}},$$

where $a_1 = \underbrace{\dfrac{\mathbb{E}(X\mathbb{1}_B)}{\mathbb{P}(B)}}_{=: \mathbb{E}(X \,|\, B)}$ and $a_2 = \underbrace{\dfrac{\mathbb{E}(X\mathbb{1}_{B^{\mathsf{c}}})}{\mathbb{P}(B^{\mathsf{c}})}}_{=: \mathbb{E}(X \,|\, B^{\mathsf{c}})}$, and that $\mathbb{E}[X \,|\, \sigma(B)]$ is a $\sigma(B)$-measurable function.

Corollary 17.3.4 For all events $A \in \mathcal{F}$, the probability of the event A can be computed by conditioning on a partition of Ω i.e.

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \,|\, B_n)\mathbb{P}(B_n).$$

Proof. Take $X = \mathbb{1}_A$ for any $A \in \mathcal{F}$. Then,

$$\begin{aligned}
\mathbb{P}(A) &:= \mathbb{E}(\mathbb{1}_A) \\
&= \mathbb{E}(\mathbb{E}[\mathbb{1}_A \,|\, \mathcal{G}]) \\
&= \mathbb{E}\left(\sum_{n=1}^{\infty} \frac{\mathbb{E}(X\mathbb{1}_{B_n})}{\mathbb{P}(B_n)} \mathbb{1}_{B_n}\right) \quad \text{where } X = \mathbb{1}_A \text{ in the proposition} \\
&= \mathbb{E}\left(\sum_{n=1}^{\infty} \frac{\mathbb{E}(\mathbb{1}_{A\cap B_n})}{\mathbb{P}(B_n)} \mathbb{1}_{B_n}\right) \\
&= \mathbb{E}\left(\sum_{n=1}^{\infty} \frac{\mathbb{P}(A \cap B_n)}{\mathbb{P}(B_n)} \mathbb{1}_{B_n}\right) \\
&=: \mathbb{E}\left(\sum_{n=1}^{\infty} \mathbb{P}(A \,|\, B_n)\, \mathbb{1}_{B_n}\right) \\
&= \sum_{n=1}^{\infty} \mathbb{E}\left(\frac{\mathbb{P}(A \cap B_n)}{\mathbb{P}(B_n)} \mathbb{1}_{B_n}\right) \quad \text{by the MCT since } \mathbb{P}(A \,|\, B_n) \geqslant 0 \\
&= \sum_{n=1}^{\infty} \mathbb{P}(A \,|\, B_n)\mathbb{E}(\mathbb{1}_{B_n}) \quad \text{by linearity} \\
&= \sum_{n=1}^{\infty} \mathbb{P}(A \,|\, B_n)\mathbb{P}(\mathbb{1}_{B_n})
\end{aligned}$$

■

Alternative. By the σ-additivity of $\mathbb{P}$:

$$\begin{aligned}
\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}\left(A \cap \left(\bigsqcup_{n=1}^{\infty} B_n\right)\right) = \mathbb{P}\left(\bigsqcup_{n=1}^{\infty} (A \cap B_n)\right) &= \sum_{n=1}^{\infty} \mathbb{P}(A \cap B_n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(A \,|\, B_n)\mathbb{P}(B_n),
\end{aligned}$$

where a similar argument lets us re-write the summand where the line breaks as the desired product. ■

What's interesting about this corollary is that the abstract definition of conditional probability arrives at the same formula using another approach. The result of the corollary can then be used to deduce the well-known Bayes Formula which is used a lot by statisticians:

$$\begin{aligned}
\mathbb{P}(B_i \,|\, A) &= \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(A \,|\, B_i)\mathbb{P}(B_i)}{\displaystyle\sum_{n=1}^{\infty} \mathbb{P}(A \,|\, B_n)\mathbb{P}(B_n)}
\end{aligned}$$

Corollary 17.3.5 (Conditioning on a Discrete Random Variable) Let $\mathcal{G} = \sigma(Y)$ where Y is a discrete random variable. Then

$$\mathbb{E}(X) = \sum_{y} \mathbb{E}(X \,|\, Y = y)\mathbb{P}(Y = y).$$

Proof. Since Y is discrete, its range is at most countably infinite and $\sigma(Y)$ is generated by all the events $Y^{-1}(\{y\})$ where y is in the range of Y. The disjoint union of the $\{Y^{-1}(\{y\})\}_{y\in\text{range}(Y)}$ is certainly equal to Ω. Thus, the collection certainly constitutes a partition of Ω. Thus,

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(\mathbb{E}[X \mid \sigma(Y)]) \\
&= \mathbb{E}\left(\sum_y \mathbb{E}\big(X \mid Y^{-1}(\{y\})\big)\mathbb{1}_{Y^{-1}(\{y\})}\right) \\
\text{“=”}\ &\mathbb{E}\left(\sum_y \mathbb{E}(X \mid Y = y)\mathbb{1}_{\{Y=y\}}\right) \\
&= \sum_y \mathbb{E}(\mathbb{E}(X \mid \{Y = y\}))\mathbb{E}\big(\mathbb{1}_{\{Y=y\}}\big) \\
&= \sum_y \mathbb{E}(\mathbb{E}(X \mid \{Y = y\}))\mathbb{P}(\{Y = y\})
\end{aligned}$$

■

17.3.1 An Illustrative Example of Conditional Expectation

This example is from Todd Kemp's Lecture 32.1.

Consider the probability space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ where $\mathcal{F}_1 \otimes \mathcal{F}_2$ is the product σ-algebra, and $\mathbb{P}_1 \otimes \mathbb{P}_2$ is the unique product measure defined on the product σ-algebra.

Let $\mathcal{G}$ be the collection $\{A \times \Omega_2 \colon A \in \mathcal{F}_1\}$. This collection injects $\mathcal{F}_1$ into the product σ-algebra in a natural way, and is a sub-σ-algebra.[7]

If we take $X \in L^1(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$, can we identify the conditional expectation Z of X given $\mathcal{G}$?

1. Such a Z would be $\mathcal{G}$-measurable. Since all singletons are Borel-measurable, it suffices to consider $Z^{-1}(\{t\})$ and show it's in an element of $\mathcal{G}$. If Z is $\mathcal{G}$-measurable, then it's certainly constant in the second variable for if $(\omega_1, \omega_2) \longmapsto t$ for some $t \in \mathbb{R}$, then certainly any other (ω_1, ω_2') also maps into t because $Z^{-1}(\{t\}) = \{\omega_1\} \times \Omega_2$. Thus, Z is a projection and so $Z(\omega_1, \omega_2) = Z'(\omega_1)$ for some measurable function Z'.
2. Such a Z would also satisfy the averaging property i.e. $\forall A \in \mathcal{F}_1$:

$$\mathbb{E}(X\mathbb{1}_{A\times\Omega_2}) = \mathbb{E}(Z\mathbb{1}_{A\times\Omega_2})$$

Since Z would be in L^1, we may apply Fubini's theorem in what follows.

$$\begin{aligned}
\mathbb{E}(Z\mathbb{1}_{A\times\Omega_2}) &= \iint_{\Omega_1\times\Omega_2} Z\mathbb{1}_{A\times\Omega_2}\,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) \\
&= \int_{\Omega_2}\left(\int_A Z(\omega_1, \omega_2)\,\mathrm{d}\mathbb{P}_1\right)\mathrm{d}\mathbb{P}_2 \quad \text{by Fubini's Theorem} \\
&= \int_{\Omega_2}\left(\int_A Z'(\omega_1)\,\mathrm{d}\mathbb{P}_1\right)\mathrm{d}\mathbb{P}_2 \\
&= \int_A Z'(\omega_1)\,\mathrm{d}\mathbb{P}_1 \quad \text{since } Z' \text{ doesn't depend on } \omega_2 \text{ and } \mathbb{P}_2(\Omega_2) = 1
\end{aligned}$$

Since $Z' \in L^1$ Now we choose to write the LHS as a double integral in the reverse order:

$$\begin{aligned}
\mathbb{E}(X\mathbb{1}_{A\times\Omega_2}) &= \iint_{\Omega_1\times\Omega_2} X\mathbb{1}_{A\times\Omega_2}\,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2) \\
&= \int_A\left(\int_{\Omega_2} X(\omega_1, \omega_2)\,\mathrm{d}\mathbb{P}_2\right)\mathrm{d}\mathbb{P}_1 \quad \text{by Fubini's Theorem}
\end{aligned}$$

This identifies Z as the L^1 function satisfying

$$Z(\omega_1, \omega_2) = Z'(\omega_1) = \int_{\Omega_2} X(\omega_1, \omega_2)\,\mathrm{d}\mathbb{P}_2.$$

[7] $\emptyset \times \Omega_1 = \emptyset$ is how one shows $\emptyset \in \mathcal{G}$.

17.4 The Conditional Expectation ($X \geqslant 0$)

This section simply quotes results that are proven in [3] 11.2.2.

Convention: Allow X to attain $+\infty$.

The conditional expectation of a non-negative random variable may also be defined.

Theorem 17.4.1 Let $X\colon \Omega \to [0, +\infty]$ be a random variable. Then there exists a $\mathcal{G}$-measurable random variable with values in $[0, +\infty]$, denoted by $\mathbb{E}[X \,|\, \mathcal{G}]$, and is such that for any non-negative $\mathcal{G}$-measurable random variable Z:

$$\mathbb{E}(ZX) = \mathbb{E}(Z\mathbb{E}[X \,|\, \mathcal{G}]).$$

Furthermore, $\mathbb{E}[X \,|\, \mathcal{G}]$ is unique up to a $\mathcal{G}$-measurable set of probability zero.

17.4.1 Properties of $\mathbb{E}[X \,|\, \mathcal{G}]$ for $X \geqslant 0$

Lemma 17.4.2

a) If X and X' are non-negative random variables, and $a, b \geqslant 0$:

$$\mathbb{E}\big[aX + bX' \,\big|\, \mathcal{G}\big] = a\mathbb{E}[X \,|\, \mathcal{G}] + b\mathbb{E}\big[X' \,\big|\, \mathcal{G}\big].$$

b) If $X \geqslant 0$ and $\mathcal{G}$-measurable, then $\mathbb{E}[X \,|\, \mathcal{G}] = \mathbb{E}(X)$.

c) For any $X \geqslant 0$, $\mathbb{E}(\mathbb{E}[X \,|\, \mathcal{G}]) = \mathbb{E}(X)$.

d) If $\{X_n\}_{n\in\mathbb{N}}$ is a non-decreasing sequence of non-negative random variables with pointwise limit $X_n \uparrow X$, then $\forall_{\mathbb{P}}\omega$:

$$\mathbb{E}[X \,|\, \mathcal{G}] = \lim_{n\to\infty} \uparrow \mathbb{E}[X_n \,|\, \mathcal{G}].$$

As a useful consequence, if $\{Y_n\}_{n\in\mathbb{N}}$ is a sequence of non-negative random variables, we have:

$$\mathbb{E}\left[\sum_{n\in\mathbb{N}} Y_n \,\middle|\, \mathcal{G}\right] = \sum_{n\in\mathbb{N}} \mathbb{E}[Y_n \,|\, \mathcal{G}]$$

e) If $\{X_n\}_{n\in\mathbb{N}}$ is any sequence of non-negative random variables, then $\forall_{\mathbb{P}}\omega$:

$$\mathbb{E}\Big[\liminf_{n\to\infty} X_n \,\Big|\, \mathcal{G}\Big] \leqslant \liminf_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}].$$

f) Let $\{X_n\}_{n\in\mathbb{N}}$ be any sequence of integrable random variables that converges $\mathbb{P}$-a.s. to X. Assume that there's a non-negative random variable Z s.t. $|X_n| \leqslant Z$ a.s. for every $n \in \mathbb{N}$, and $\mathbb{E}(Z) < \infty$, then

$$L^1 \ni \mathbb{E}[X \,|\, \mathcal{G}] = \lim_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}] \text{ a.s.}$$

g) Jensen's Inequality for Conditional Expectations i.e. if $f\colon \mathbb{R} \to [0, +\infty)$ is convex, and $X \in L^1$ then

$$\mathbb{E}[f(X) \,|\, \mathcal{G}] \geqslant f(\mathbb{E}[X \,|\, \mathcal{G}]).$$

17.5 Conditional Expectation As Projection ($X \in L^2$)

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be an inner product space, $\mathcal{K} \subseteq \mathcal{H}$ be a linear subspace. Then there exists an orthogonal projection map onto $\mathcal{K}$

$$P_{\mathcal{K}} \colon \mathcal{H} \to \mathcal{K}$$

satisfying the following:

- $P_{\mathcal{K}}$ is surjective
- $P_{\mathcal{K}}(v) = v$ for every $v \in \mathcal{K}$ (fixes all vectors in $\mathcal{K}$)
- $P_{\mathcal{K}}(v) = 0$ if $v \in \mathcal{K}^{\perp} \coloneqq \{w \in V : \langle w, v \rangle = 0 \text{ for all } v \in \mathcal{K}\}$

Geometrically, the picture is clear. We have some vector $v \in \mathcal{H}$ and we wish to orthogonally project it onto K. We achieve this by looking at the orthogonal complement to $\mathcal{K}$, and project down along that subspace onto $\mathcal{K}$. The key is to find the vector $a = v - P_{\mathcal{K}}(v)$ which is orthogonal to $\mathcal{K}$. This vector uniquely specifies the orthogonal projection.

Given a Hilbert space $\mathcal{H}$ and a **closed** linear subspace $\mathcal{K} \subseteq \mathcal{H}$, for any $X \in \mathcal{H}$, $\exists! Z \in \mathcal{K}$, called the orthogonal projection of X onto $\mathcal{K}$, and denoted by $P_{\mathcal{K}}(X)$, satisfying the following equivalent conditions:

- $P_{\mathcal{K}}(X)$ is the unique $Z \in \mathcal{K}$ minimising $\|X - Z\|$
- $P_{\mathcal{K}}(X)$ is the unique $Z \in \mathcal{K}$ s.t. $(X - Z) \perp \mathcal{K}$

i.e. $\exists!$ linear transformation $P_{\mathcal{K}} \colon \mathcal{H} \to \mathcal{K}$ s.t.

1. $P_{\mathcal{K}}$ is Lipschitz-continuous (i.e. a bounded operator with operator norm $\leqslant 1$)
2. $P_{\mathcal{K}}$ fixes all the vectors in $\mathcal{K}$
$$\forall Y \in \mathcal{K} \colon P_{\mathcal{K}}(Y) = Y$$
3. Kills all the vectors orthogonal to $\mathcal{K}$
$$\forall Z \in \mathcal{K}^{\perp} \colon P_{\mathcal{K}}(Z) = 0$$
4. Is self-adjoint with respect to $\langle \cdot, \cdot \rangle$ i.e.
$$\forall X, Y \in \mathcal{H} \colon \langle P_{\mathcal{K}}(X), Y \rangle = \langle X, P_{\mathcal{K}}(Y) \rangle$$

17.5.1 Application to Conditional Expectation

$L^2(\Omega, \mathcal{F}, \mathbb{P}) = \mathcal{H}$ is a Hilbert space when equipped with the L^2 inner product

$$\langle f, g \rangle_{L^2} \coloneqq \mathbb{E}(XY) \coloneqq \int_{\Omega} XY \, \mathrm{d}\mathbb{P}.$$

In our case, we can isometrically identify $\mathcal{K} = L^2(\Omega, \mathcal{G}, \mathbb{P})$ with a closed subspace of $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mathbb{P})$ — namely, the subspace consisting of all elements of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ that have at least one representative that is $\mathcal{G}$-measurable. Thus, we can make sense of the orthogonal projection of an element of $\mathcal{H}$ onto the closed subspace $\mathcal{K}$:

Theorem 17.5.1 (11.5 [3]) If $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathbb{E}[X \,|\, \mathcal{G}]$ is the orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Proof. Jensen's inequality shows that $\mathbb{P}$-a.s.

$$(\mathbb{E}[X \,|\, \mathcal{G}])^2 \leqslant \mathbb{E}\big[X^2 \,\big|\, \mathcal{G}\big]$$

from which it follows that

$$\mathbb{E}\big((\mathbb{E}[X \,|\, \mathcal{G}])^2\big) \leqslant \mathbb{E}\big(\mathbb{E}\big[X^2 \,\big|\, \mathcal{G}\big]\big) = \mathbb{E}\big(X^2\big) < \infty$$

so $\mathbb{E}[X \,|\, \mathcal{G}] \in L^2(\Omega, \mathcal{G}, \mathbb{P})$. On the other hand, by the averaging/characteristic property of $\mathbb{E}[X \,|\, \mathcal{G}]$, we have that for every bounded $\mathcal{G}$-measurable Z

$$\mathbb{E}(ZX) = \mathbb{E}(Z\mathbb{E}[X \,|\, \mathcal{G}])$$

which implies that

$$0 = \mathbb{E}(ZX) - \mathbb{E}(Z\mathbb{E}[X \,|\, \mathcal{G}]) = \mathbb{E}(Z(X - \mathbb{E}[X \,|\, \mathcal{G}])) =: \langle Z, X - \mathbb{E}[X \,|\, \mathcal{G}]\rangle_{L^2}$$

so $X - \mathbb{E}[X \,|\, \mathcal{G}]$ is orthogonal to the space of all bounded $\mathcal{G}$-measurable random variables — a space which is dense in $L^2(\Omega, \mathcal{G}, \mathbb{P})$. It follows[8] that $X - \mathbb{E}[X \,|\, \mathcal{G}]$ is orthogonal to $L^2(\Omega, \mathcal{G}, \mathbb{P})$. ■

Thus, we may interpret the conditional expectation $\mathbb{E}[X \,|\, \mathcal{G}]$ as the best approximation of X by a $\mathcal{G}$-measurable function in the L^2 sense that for any other $\mathcal{G}$-measurable random variable Y:

$$\mathbb{E}\big((Y - X)^2\big) \geqslant \mathbb{E}\big((\mathbb{E}[X \,|\, \mathcal{G}] - X)^2\big).$$

17.6 More Properties of Conditional Expectation ($X \geqslant 0$ or $X \in L^1$)

Proposition 17.6.1 (Nested σ-Algebras, 11.7 [3]) Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two sub-σ-algebras of $\mathcal{F}$ s.t. $\mathcal{G}_1 \subseteq \mathcal{G}_2$. Then, for every non-negative (or integrable) random variable X, we have that

$$\mathbb{E}(\mathbb{E}[X \,|\, \mathcal{G}_2] \,|\, \mathcal{G}_1) = \mathbb{E}[X \,|\, \mathcal{G}_1].$$

Proof. Consider the case where $X \geqslant 0$. Let Z be a non-negative $\mathcal{G}_1$-measurable random variable. We wish to show that $\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}_2] \,|\, \mathcal{G}_1]$ is $\mathcal{G}_1$-measurable and satisfies the characteristic property of $\mathbb{E}[X \,|\, \mathcal{G}_1]$ in order to conclude the proof. Thus, consider

$$\begin{aligned}
&= \mathbb{E}(Z\,\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}_2] \,|\, \mathcal{G}_1]) \\
&= \mathbb{E}(Z\,\mathbb{E}[X \,|\, \mathcal{G}_2]) \quad \text{by the law of total expectation} \\
&= \mathbb{E}(ZX),
\end{aligned}$$

where the final equality follows from the averaging property of $\mathbb{E}[X \,|\, \mathcal{G}_2]$, noting that $Z \in \mathcal{G}_1 \implies Z \in \mathcal{G}_2$. ■

Remarks 17.6.2 Note that we also have, under the same assumptions as in the proposition, that

$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}_1] \,|\, \mathcal{G}_2] = \mathbb{E}[X \,|\, \mathcal{G}_1]$$

but this is trivial since $\mathbb{E}[X \,|\, \mathcal{G}_1]$ is $\mathcal{G}_2$-measurable (since it's already $\mathcal{G}_1$-measurable and $\mathcal{G}_1 \subseteq \mathcal{G}_2$).

[8] By some limiting argument?

CHAPTER 18

Conditional Probability

Elementary Conditional Probability

Let $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Recall that the naïve/elementary definition of the conditional probability of A given B is given by

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

This definition applies equally well to events corresponding to a pair of discrete random variables X, Y:

Definition 18.0.1 For a pair of discrete random variables X, Y, the **conditional probability of the event $\{X = x\}$ given $\{Y = y\}$** is defined by

$$\mathbb{P}(\{X = x\} \,|\, \{Y = y\}) = \frac{\mathbb{P}(\{X = x\} \cap \{Y = y\})}{\mathbb{P}(\{Y = y\})}, \tag{$\dagger$}$$

where $\mathbb{P}(\{Y = y\}) > 0$.

Since † only holds for events $\{Y = y\}$ with positive probability, the definition doesn't easily generalise to the case where Y is not discrete. For example, if Y is absolutely continuous then every $\{Y = y\}$ has probability zero. However, one can try to make sense of the expression by considering an interval about y e.g. $\{y - h \leqslant Y \leqslant y + h\}$ for some $h > 0$ and investigating the limiting expression of a modified version of †:

$$\mathbb{P}(\{X = x\} \,|\, \{y - h \leqslant Y \leqslant y + h\}) := \lim_{h \to 0} \frac{\mathbb{P}(\{X = x\} \cap \{y - h \leqslant Y \leqslant y + h\})}{\mathbb{P}(\{y - h \leqslant Y \leqslant y + h\})}$$

if such a limit exists.[1]

The above development is seen at the undergraduate level. At graduate level, the opposite route is taken to develop the theory — define conditional expectation and use it to define other conditional quantities e.g. a generalisation of conditional probability from the discrete case.

18.1 Conditional Probability of A given $\mathcal{G}$

Definition 18.1.1 The **conditional probability of $A \in \mathcal{F}$ given a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$** is denoted[2] by $\mathbb{P}[A \,|\, \mathcal{G}]$ and is defined to be the conditional expectation of $\mathbb{1}_A$ given $\mathcal{G}$:

$$\mathbb{P}[A \,|\, \mathcal{G}] := \mathbb{E}[\mathbb{1}_A \,|\, \mathcal{G}].$$

Note that $\mathbb{P}[A \,|\, \mathcal{G}]$ is an equivalence class of $\mathcal{G}$-measurable random variables that satisfy the defining property i.e. $\forall B \in \mathcal{G}$:

$$\mathbb{E}(\mathbb{1}_B \mathbb{P}[A \,|\, \mathcal{G}]) = \mathbb{E}(\mathbb{1}_B \mathbb{1}_A).$$

which simplifies to

$$\int_B \mathbb{P}[A \,|\, \mathcal{G}] \, \mathrm{d}\mathbb{P} = \mathbb{P}(A \cap B).$$

[1] I wonder why one considers a symmetric interval $[y - h, y + h]$ about y and whether it makes a difference. I imagine L'Hôpital's rule could be useful in recovering the limit if the density of Y is differentiable at y, and maybe the type of interval corresponds to the type (e.g. left-differentiable) of differentiability at said point y.

[2] Recall my convention that round parentheses in a probability-related map denote that the object is a number e.g. $\mathbb{E}(X)$, and square parentheses indicate that the object is a random variable e.g. $\mathbb{E}[X \,|\, \mathcal{G}]$.

Remarks 18.1.2

- The above definition is in the same spirit as $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A)$, but this time we're defining a *random variable* $\mathbb{P}[A \,|\, \mathcal{G}]$ by the conditional expectation.
- $\mathbb{P}[A \,|\, \mathcal{G}]$ can be thought of as encoding the updated belief about A given the partial information $\mathcal{G}$, from which (via integration) we can reproduce the joint probabilities $\mathbb{P}(A \cap B)$ for $B \in \mathcal{G}$.
- Whether or not A occurs, encoded by the random variable $\mathbb{1}_A$, may or may not be something we can determine from $\mathcal{G}$ alone i.e. $\mathbb{1}_A \in \mathcal{G}$ is unknown. If not, we turn to $\mathbb{P}[A \,|\, \mathcal{G}]$ as the best $\mathcal{G}$-measurable approximation to the indicator $\mathbb{1}_A$.

I will abuse notation and often write $\mathbb{P}[A \,|\, \mathcal{G}](\omega)$ for a version of the conditional probability. The context *should* be clear when I do this, but if not, I will use something like f_A to denote such a version:

Definition 18.1.3 By varying over $A \in \mathcal{F}$, we can define the **conditional probability on $\mathcal{F}$ given $\mathcal{G}$** by the map $\kappa \colon \mathcal{F} \times \Omega \to [0,1]$ which is defined for each $A \in \mathcal{F}$ by:

$$\omega \longmapsto \kappa(A, \omega) := \mathbb{P}[A \,|\, \mathcal{G}](\omega).$$

Example 18.1.4 Suppose that $\mathcal{G}$ is generated by B i.e. $\mathcal{G} = \sigma(B) = \{\emptyset, B, B^{\mathsf{c}}, \Omega\}$. Any version f_A of the conditional probability $\mathbb{P}[A \,|\, \mathcal{G}] := \mathbb{E}[\mathbb{1}_A \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_A \,|\, \sigma(B)]$ is therefore constant on each cell of the partition $\{B, B^{\mathsf{c}}\}$ of Ω. By Lemma 17.3.2, we can write f_A as the sum

$$f_A = a_B \mathbb{1}_B + a_{B^{\mathsf{c}}} \mathbb{1}_{B^{\mathsf{c}}}.$$

We can determine the constants a_B and $a_{B^{\mathsf{c}}}$ by the defining property of conditional expectation.

- For the event $B \in \mathcal{G}$, the property tells us that

$$\mathbb{P}(A \cap B) = \int_B (a_B \mathbb{1}_B + a_{B^{\mathsf{c}}} \mathbb{1}_{B^{\mathsf{c}}}) \, \mathrm{d}\mathbb{P} = a_B \mathbb{P}(B) + a_{B^{\mathsf{c}}} \int_\Omega \mathbb{1}_\emptyset \, \mathrm{d}\mathbb{P} = a_B \mathbb{P}(B)$$

- Similarly, for $B^{\mathsf{c}} \in \mathcal{G}$, we obtain $a_{B^{\mathsf{c}}} \mathbb{P}(B^{\mathsf{c}}) = \mathbb{P}(A \cap B^{\mathsf{c}})$.

Now one notes that if $\mathbb{P}(B) = 0$, then we have that $\mathbb{P}(A \cap B) = 0$. This tells us nothing about how to determine a_B. Recall that any version f_A of $\mathbb{P}[A \,|\, \mathcal{G}]$ is uniquely defined up to a $\mathbb{P}$-null set so its behaviour on B is arbitrary i.e. we can choose any value on a cell of measure zero and this won't change that $f_A = a_B \mathbb{1}_B + a_{B^{\mathsf{c}}} \mathbb{1}_{B^{\mathsf{c}}}$ is a valid version of $\mathbb{P}[A \,|\, \mathcal{G}]$.

If we suppose that $\mathbb{P}(B), \mathbb{P}(B^{\mathsf{c}}) > 0$, we may recover the classical expressions

$$a_B = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{and} \quad a_{B^{\mathsf{c}}} = \frac{\mathbb{P}(A \cap B^{\mathsf{c}})}{\mathbb{P}(B^{\mathsf{c}})}$$

and so our version f_A is given by

$$f_A = \underbrace{\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}_{=: \mathbb{P}(A \,|\, B)} \mathbb{1}_B + \underbrace{\frac{\mathbb{P}(A \cap B^{\mathsf{c}})}{\mathbb{P}(B^{\mathsf{c}})}}_{=: \mathbb{P}(A \,|\, B^{\mathsf{c}})} \mathbb{1}_{B^{\mathsf{c}}}.$$

For the sake of completeness, checking the defining property for the other two events $\emptyset, \Omega$ of $\mathcal{G}$ are trivialities:

- For $\emptyset$, we have $0 = \int_\emptyset f_A \, \mathrm{d}\mathbb{P} = \mathbb{P}(A \cap \emptyset) = 0$.

- For Ω, we have $\int_\Omega f_A \, d\mathbb{P} = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A)$ which is just the law of total expectation $\mathbb{E}(\mathbb{P}[A \,|\, \mathcal{G}]) = \mathbb{E}(\mathbb{E}[\mathbb{1}_A \,|\, \mathcal{G}]) = \mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$.

Corollary 18.1.5 For a countable partition $\{B_i\}_{i \in J \subseteq \mathbb{N}} \subseteq \mathcal{F}$ of Ω, the conditional probability of $A \in \mathcal{F}$ given $\mathcal{G} = \sigma(\{B_i\})$ is given by

$$\mathbb{P}[A \,|\, \mathcal{G}] = \sum_{i \in J} \mathbb{P}(A \,|\, B_i) \mathbb{1}_{B_i}.$$

Furthermore, the map $\kappa \colon \mathcal{F} \times \Omega \to [0, 1]$ defined for each $A \in \mathcal{F}$ by:

$$\omega \longmapsto \kappa(A, \omega) := \mathbb{P}[A \,|\, \mathcal{G}](\omega) = \sum_{i \in J} \mathbb{P}(A \,|\, B_i) \mathbb{1}_{B_i}(\omega)$$

is a conditional probability on $\mathcal{F}$ given $\mathcal{G} = \sigma(\{B_i\})$.

18.2 Regular Conditional Probability

From the above definition, we understand that for each $A \in \mathcal{F}$, the map $\kappa(A, \cdot)$ is a $\mathcal{G}$-measurable random variable — in particular it's a version of $\mathbb{P}[A \,|\, \mathcal{G}]$. That's only half of the map and this begs the question: What kind of object should $\kappa(\cdot, \omega)$ be defined by κ for each ω?

As some motivation for this question, Le Gall alludes (in [3, p. 248]) to the idea that we can use κ to model a stochastic process (which can loosely be thought of as a collection of states and probabilities of moving between states). Namely, for *every* "starting state" $\omega \in \Omega$, if we can guarantee that $\kappa(\cdot, \omega)$ is a probability measure then we can use it to model how we choose an "arrival point" ω' in a random manner on the *entire* outcome space.

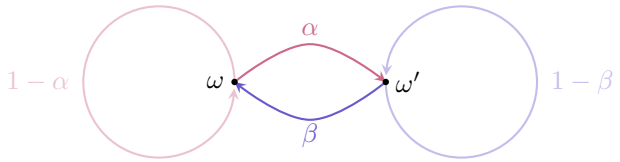


Figure 18.1: $\Omega = \{\omega, \omega'\}$ and a state diagram representing the corresponding transition probabilities. At each state, the probabilities must sum to 1.

If we can do this for **every** $\omega \in \Omega$, then we will call κ **regular**. We'll now attempt to verify the probability axioms for $\kappa(\cdot, \omega)$.

- It's clear that $\mathbb{P}[\emptyset \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_\emptyset \,|\, \mathcal{G}] = 0$ almost surely since $\mathbb{1}_\emptyset$ is the constant random variable with value zero. Thus, we may choose the constant function $\kappa(\emptyset, \omega) = 0$ (pointwise) as our version of $\mathbb{P}[\emptyset \,|\, \mathcal{G}]$.
- Similarly, $\mathbb{P}[\Omega \,|\, \mathcal{G}] = \mathbb{E}[\mathbb{1}_\Omega \,|\, \mathcal{G}] = 1$ almost everywhere. By the same logic, we may choose the constant function $\kappa(\Omega, \cdot) = 1$ as our version of $\mathbb{P}[\Omega \,|\, \mathcal{G}]$.

- For a collection $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ of pairwise disjoint events, we have that

$$\begin{aligned}
\mathbb{P}\left[\bigsqcup_{n\in\mathbb{N}} A_n \,\middle|\, \mathcal{G}\right] &:= \mathbb{E}\left[\mathbb{1}_{\bigsqcup_{n\in\mathbb{N}} A_n} \,\middle|\, \mathcal{G}\right] \\
&= \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbb{1}_{A_n} \,\middle|\, \mathcal{G}\right] \\
&= \mathbb{E}\left[\lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{1}_{A_n} \,\middle|\, \mathcal{G}\right] \\
&= \lim_{N\to\infty} \mathbb{E}\left[\sum_{n=1}^{N} \mathbb{1}_{A_n} \,\middle|\, \mathcal{G}\right] \quad \text{by the conditional MCT} \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{E}[\mathbb{1}_{A_n} \,|\, \mathcal{G}] \quad \text{by linearity of } \mathbb{E}[\cdot \,|\, \mathcal{G}] \colon L^1(\Omega,\mathcal{F},\mathbb{P}) \to L^1(\Omega,\mathcal{G},\mathbb{P}) \\
&= \sum_{n\in\mathbb{N}} \mathbb{E}[\mathbb{1}_{A_n} \,|\, \mathcal{G}] \\
&=: \sum_{n\in\mathbb{N}} \mathbb{P}[A_n \,|\, \mathcal{G}]
\end{aligned}$$

  where the MCT was applied with $\mathbb{1}_{\bigsqcup_{n=1}^{N} A_n} = \sum_{n=1}^{N} \mathbb{1}_{A_n} \uparrow \mathbb{1}_{\bigsqcup_{n\in\mathbb{N}} A_n}$ as $N \to \infty$.

Remarks 18.2.1

- The order of quantifiers is important here. For each! particular collection $\{A_n\}_{n\in\mathbb{N}}$, these corresponding equality demonstrated above, crucially, hold **only** $\mathbb{P}$-a.e. so the convergence of the series $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n \,|\, \mathcal{G}]$ is only almost sure.
- Let $\omega \in \Omega$. For each $A \in \mathcal{F}$, fix a version f_A of $\mathbb{P}[A \,|\, \mathcal{G}]$. Does the map $A \longmapsto f_A(\omega)$ define a measure?

  Let $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ be any pairwise disjoint family. Denote by A their union $\sqcup_{n\in\mathbb{N}} A_n$. The conditional version of the Monotone Convergence Theorem tells us that

$$f_A(\omega) = \sum_{n\in\mathbb{N}} f_{A_n}(\omega). \qquad (\sigma_{\text{a.s.}})$$

  holds $\mathbb{P}$-almost surely i.e.

$$\exists N_{\{A_n\}_{n\in\mathbb{N}}} \in \mathcal{G} \text{ that is } \mathbb{P}\text{-null}$$

  s.t. $(\sigma_{\text{a.s.}})$ holds pointwise for $\omega \in \Omega \setminus N_{\{A_n\}_{n\in\mathbb{N}}}$.

  Since in general there are potentially uncountably many such families $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$, then there must also be potentially countably many corresponding null sets in $\mathcal{G}$ whose union $N \in \mathcal{G}$ may not necessarily be $\mathbb{P}$-negligible, and so we cannot guarantee the existence of a $\mathbb{P}$-null set $N \in \mathcal{G}$ s.t. $(\sigma_{\text{a.s.}})$ holds true for every collection $\{A_n\}_{n\in\mathbb{N}}$ simultaneously on $\Omega \setminus N$.

Despite there being no guarantee in general that the assignment $A \longmapsto f_A(\omega)$ is a measure for $\mathbb{P}$-almost every $\omega \in \Omega$, we can demonstrate that the series $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n \,|\, \mathcal{G}]$ converges in the L^1 sense, from which we deduce that $A \longmapsto \mathbb{P}[A \,|\, \mathcal{G}]$ is an $(L^1(\Omega,\mathcal{G},\mathbb{P}), \|\cdot\|_{L^1})$-valued *vector measure* — a generalisation of a finite measure taking values in $[0,+\infty]$:

Definition 18.2.2 Let $(\Omega,\mathcal{F})$ be a measurable space, and $(B, \|\cdot\|_B)$ be a Banach space. A map $\mu \colon \Omega \to B$ is called a **vector measure** if:

- $\mu(\emptyset) = 0$

- For any pairwise disjoint collection $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$:

$$\mu\Big(\bigsqcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n),$$

where the series converges in $(B, \|\cdot\|_B)$.

Proof. The series $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n \,|\, \mathcal{G}]$ converges absolutely in $L^1(\Omega, \mathcal{G}, \mathbb{P})$ because

$$\begin{aligned}
\sum_{n\in\mathbb{N}} \|\mathbb{P}[A_n \,|\, \mathcal{G}]\|_{L^1} &= \sum_{n\in\mathbb{N}} \mathbb{E}(|\mathbb{P}[A_n \,|\, \mathcal{G}]|) \\
&= \sum_{n\in\mathbb{N}} \mathbb{E}(|\mathbb{E}[\mathbb{1}_{A_n} \,|\, \mathcal{G}]|) \\
&= \sum_{n\in\mathbb{N}} \mathbb{E}(\mathbb{E}[\mathbb{1}_{A_n} \,|\, \mathcal{G}]) \quad \text{since } \mathbb{E}[\mathbb{1}_{A_n} \,|\, \mathcal{G}] \text{ is non-negative } \mathbb{P}\text{-a.e.} \\
&= \sum_{n\in\mathbb{N}} \mathbb{E}(\mathbb{1}_{A_n}) \quad \text{by the law of total expectation} \\
&= \sum_{n\in\mathbb{N}} \mathbb{P}(A_n) \\
&= \mathbb{P}\Big(\bigsqcup_{n\in\mathbb{N}} A_n\Big) \leqslant 1 < \infty.
\end{aligned}$$

Since $(L^1(\Omega, \mathcal{G}, \mathbb{P}), \|\cdot\|_{L^1})$ is a Banach space, the original series converges to some element in the same space — exactly what we think it is:

$$\begin{aligned}
\left\| \mathbb{P}\left[\bigsqcup_{n\in\mathbb{N}} A_n \,\middle|\, \mathcal{G}\right] - \sum_{n=1}^{N} \mathbb{P}[A_n \,|\, \mathcal{G}] \right\|_{L^1} &:= \left\| \mathbb{E}\left[\sum_{n\in\mathbb{N}} \mathbb{1}_{A_n} \,\middle|\, \mathcal{G}\right] - \sum_{n=1}^{N} \mathbb{E}[\mathbb{1}_{A_n} \,|\, \mathcal{G}] \right\|_{L^1} \\
&= \left\| \mathbb{E}\left[\sum_{n=N+1}^{\infty} \mathbb{1}_{A_n} \,\middle|\, \mathcal{G}\right] \right\|_{L^1} \\
&= \mathbb{P}\left(\bigsqcup_{n=N+1}^{\infty} A_n\right) \overset{N\to\infty}{\longrightarrow} 0
\end{aligned}$$

■

- The vector measure structure will permit one to use the tools of functional analysis to investigate $A \mapsto \mathbb{P}[A \,|\, \mathcal{G}]$ further.

Small tangent on vector measures aside, the problem for arbitrary $\mathcal{F}$ is that there are too many sets in $\mathcal{F}$. The appropriate corrective step is to construct a map (if it exists) for which the assignment $A \longmapsto \mathbb{P}[A \,|\, \mathcal{G}](\omega)$ is a probability measure for **at least** $\mathbb{P}$-a.e. $\omega \in \Omega$.

Definition 18.2.3 A function $\kappa\colon \mathcal{F} \times \Omega \to [0,1]$ is called a **regular conditional probability on $\mathcal{F}$ given $\mathcal{G}$** if:

1. For $\mathbb{P}$-a.e. $\omega \in \Omega$, $\kappa(\cdot, \omega)$ is a probability measure on $(\Omega, \mathcal{F})$.
2. For each $A \in \mathcal{F}$, the mapping $\omega \longmapsto \kappa(A, \omega)$ is a version of $\mathbb{P}[A \,|\, \mathcal{G}]$ i.e.
   - $\kappa(A, \cdot)$ is $\mathcal{G}$-measurable, agrees $\mathbb{P}$-a.s. with $\mathbb{P}[A \,|\, \mathcal{G}]$, and
   - satisfies the **averaging property** $\forall B \in \mathcal{G}$:

$$\mathbb{P}(A \cap B) = \mathbb{E}(\mathbb{1}_B \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_B\, \kappa(A, \cdot)) = \int_B \kappa(A, \omega)\, \mathrm{d}\mathbb{P}(\omega).$$

If a regular conditional probability exists (with our $\mathbb{P}$-a.e definition), then we can easily upgrade it to a pointwise statement on Ω. Suppose that N is the null set outside of which (i.e. for every $\omega \in \Omega \setminus N$) $\kappa(\cdot, \omega)$ every is a probability measure. Let $\xi \in \Omega$. For every $\omega \in N$, and for every $A \in \mathcal{F}$, re-define

$$\mathbb{P}[A \,|\, \mathcal{G}](\omega) := \mathbb{1}_A(\xi).$$

Then condition 1. certainly holds, and so too 2. for every $\omega \in \Omega$.

On the other hand, if we state the definition with $\kappa(\cdot, \omega)$ being a probability measure for every $\omega \in \Omega$, then it certainly follows $\forall_{\mathbb{P}}\, \omega \in \Omega$. Thus, both definitions are **equivalent** when the weaker (a.e.) statement is true.

Henceforth, the definition of a regular conditional probability will state that $\kappa(\cdot, \omega)$ is a probability measure on $(\Omega, \mathcal{F})$ for **every** $\omega \in \Omega$.

This convention sidesteps some technical difficulties.

18.2.1 Regular conditional distribution of X given $\mathcal{G}$

In the particular case that $\mathcal{F}$ is generated by a random variable X, we give a new name to a particular instance of regular conditional probability as per Remark 12.3.1 from [13, p. 393]:

Definition 18.2.4 When $\mathcal{F} = \sigma(X)$, the regular conditional probability of $\mathcal{F}$ given $\mathcal{G}$ is also called the **regular conditional probability distribution of X given $\mathcal{G}$**.

Theorem 18.2.5 (B.32 [5]) If X maps into a Borel space, then there exists a regular conditional distribution of X (equivalently $\mathcal{F} = \sigma(X)$) given any sub-σ-algebra $\mathcal{G} \subseteq \sigma(X)$.

18.2.2 $\mathcal{G}$ gen. by partition

We can use the conditional probability κ from **Corollary 18.1.5** as an example. It is indeed regular.

Proof. Without loss of generality, the partition can take the form $\{B_i\}_{i \in \{0\} \cup J \subseteq \mathbb{N}}$ where $\mathbb{P}(B_i) > 0$ for every $i \in J$, and B_0 is $\mathbb{P}$-null but not necessarily empty.

By construction, for each $A \in \mathcal{F}$ we already have that $\omega \mapsto \kappa(A, \omega) = \mathbb{P}\big[A \,\big|\, \sigma(\{B_i\}_{i \in \{0\} \cup J})\big](\omega)$ is already a conditional probability that decomposes into a countable collection of probability measures:

$$\mathbb{P}[A \,|\, \mathcal{G}](\omega) = \mathbb{P}(A \,|\, B_0)\mathbb{1}_{B_0}(\omega) + \sum_{i \in J} \mathbb{P}(A \,|\, B_i)\mathbb{1}_{B_i}(\omega).$$

What remains to be seen is that for each $\omega \in \Omega$, the map $A \mapsto \mathbb{P}[A \,|\, \mathcal{G}](\omega)$ is a probability measure. This is certainly the case because any fixed ω is a member of at most one of the cells, e.g. B_{i_0}, and if so, all other terms in the decomposition with $\mathbb{1}_{B_j}(\omega)$ where $j \neq i_0$ vanish, leaving the probability measure $\mathbb{P}(\cdot \,|\, B_{i_0})$ behind. ■

18.2.3 Defining conditional expectation via r.c.p on $\mathcal{F}$ given $\mathcal{G}$-arbitrary

Back to the general case of an arbitrary sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Take for granted, for now, that a regular conditional probability κ exists.

By property 2 of κ, for each $\omega \in \Omega$, $\kappa(\cdot, \omega)$ is a probability measure on $(\Omega, \mathcal{F})$ and so we may

write[3] for any $A \in \mathcal{F}$:

$$\kappa(A,\omega) = \int_A \mathrm{d}\kappa(\cdot,\omega)(\omega') = \int_\Omega \mathbb{1}_A(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega').$$

Thus, we may write the integral property of $\kappa(A,\cdot)$ in the suggestive form:

$$\begin{aligned}\mathbb{E}(\mathbb{1}_B\mathbb{1}_A) &= \mathbb{E}(\mathbb{1}_B\,\kappa(A,\cdot)) \\ &= \int_B \kappa(A,\omega)\,\mathrm{d}\mathbb{P}(\omega) \\ &= \int_B\int_\Omega \mathbb{1}_A(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega')\,\mathrm{d}\mathbb{P}(\omega).\end{aligned}$$

This equality expresses that the conditional expectation $\kappa(A,\omega)$ of $\mathbb{1}_A$ given $\mathcal{G}$ is equal ($\mathbb{P}$-a.s.) to the Lebesgue integral of $\mathbb{1}_A$ with respect to the probability measure $\kappa(\cdot,\omega)$ at said ω.

Since this holds for indicators, by linearity we can extend to simple functions, then non-negative random variables, and finally any integrable random variable X. This is demonstrated in the following proof:

Claim Suppose that an r.c.p. κ on $\mathcal{F}$ given $\mathcal{G}$ exists, and that $X \in L^1(\Omega,\mathcal{F},\mathbb{P})$. Then for $\mathbb{P}$-a.e. $\omega \in \Omega$:

$$\mathbb{E}[X\,|\,\mathcal{G}](\omega) = \int_\Omega X(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega').$$

Proof. Denote the RHS by

$$h(\omega) \coloneqq \int_\Omega X(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega').$$

We want to show that h is a version of the conditional expectation $\mathbb{E}[X\,|\,\mathcal{G}]$.

1. For $X = \mathbb{1}_A$, this has already been shown. Note, in particular, that $\forall_{\mathbb{P}}\,\omega$: $h(\omega) = \kappa(A,\omega)$ which is certainly $\mathcal{G}$-measurable.

2. Now consider a simple function $X = \sum_{i=1}^n a_i\mathbb{1}_{A_i}$ where every $A_i \in \mathcal{F}$. Then, by linearity

$$h(\omega) = \int_\Omega \sum_{i=1}^n a_i\mathbb{1}_{A_i}(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega') = \sum_{i=1}^n a_i\kappa(A_i,\omega)$$

which is a sum of $\mathcal{G}$-measurable functions so $h \in \mathcal{G}$.

For any $B \in \mathcal{G}$:

$$\begin{aligned}\mathbb{E}(\mathbb{1}_B h) &= \mathbb{E}\left(\mathbb{1}_B \sum_{i=1}^n a_i\kappa(A_i,\cdot)\right) \\ &= \sum_{i=1}^n a_i\mathbb{E}(\mathbb{1}_B\kappa(A_i,\cdot)) \\ &= \sum_{i=1}^n a_i\mathbb{E}(\mathbb{1}_B\mathbb{1}_{A_i}) \quad \text{by the averaging prop. of } \kappa(A_i,\cdot) \text{ which is a version of } \mathbb{P}[A_i\,|\,\mathcal{G}] \\ &= \mathbb{E}\left(\mathbb{1}_B \sum_{i=1}^n a_i\mathbb{1}_{A_i}\right) \\ &= \mathbb{E}(\mathbb{1}_B X)\end{aligned}$$

[3] In the exact same sense as

$$\mu(A) = \int_A \mathrm{d}\mu(x) = \int_\Omega \mathbb{1}_A(x)\,\mathrm{d}\mu(x).$$

3. Let X be a non-negative random variable. There exists a non-decreasing sequence of simple measurable functions $\{X_n\}_{n\in\mathbb{N}}$ with pointwise limit $\lim_{n\to\infty} X_n(\omega') = X(\omega')$ for every $\omega' \in \Omega$. Since for each $\omega \in \Omega$, the map $\kappa(\cdot,\omega)$ is a probability measure on $(\Omega,\mathcal{F})$, define

$$h_n(\omega) := \int_\Omega X_n(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega')$$

so that $\{h_n(\omega)\}_{n\in\mathbb{N}}$ is a non-decreasing sequence of non-negative $\mathcal{G}$-measurable functions. By the Monotone Convergence theorem,

$$\begin{aligned}\lim_{n\to\infty} h_n(\omega) &= \int_\Omega \lim_{n\to\infty} X_n(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega') \\ &= \int_\Omega X(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega') =: h(\omega).\end{aligned}$$

As the pointwise limit (for $\mathbb{P}$-a.e. $\omega \in \Omega$), h is $\mathcal{G}$-measurable. Now let $B \in \mathcal{G}$. Noting that for $\mathbb{P}$-a.e. $\omega \in \Omega$, $\{\mathbb{1}_B\, h_n\}_{n\in\mathbb{N}} \uparrow \mathbb{1}_B\, h$ and so:

$$\begin{aligned}\mathbb{E}(\mathbb{1}_B\, h) &= \mathbb{E}\Big(\lim_{n\to\infty} \mathbb{1}_B\, h_n\Big) \\ &= \lim_{n\to\infty} \mathbb{E}(\mathbb{1}_B\, h_n) \quad \text{by the MCT on } \{\mathbb{1}_B\, h_n\}_{n\in\mathbb{N}} \\ &= \lim_{n\to\infty} \mathbb{E}(\mathbb{1}_B X_n) \quad \text{by Step 2 since } X_n\text{-simple} \\ &= \mathbb{E}\Big(\lim_{n\to\infty} \mathbb{1}_B X_n\Big) \quad \text{by the MCT again but this time on } \{\mathbb{1}_B X_n\}_{n\in\mathbb{N}}. \\ &= \mathbb{E}(\mathbb{1}_B X)\end{aligned}$$

4. We may write any $X \in L^1(\Omega,\mathcal{F},\mathbb{P})$ as $X = X^+ - X^-$ where $X^+ = \max(0,X)$ and $X^- = \min(0,-X)$ are non-negative integrable random variables. Now denote by

$$h^\pm(\omega) := \int_\Omega X^\pm(\omega')\,\mathrm{d}\kappa(\cdot,\omega)(\omega').$$

Therefore $h = h^+ - h^-$ which is the difference of two $\mathcal{G}$-measurable functions, is also $\mathcal{G}$-measurable. For any $B \in \mathcal{G}$, we have that:

$$\begin{aligned}\mathbb{E}(\mathbb{1}_B h) &= \mathbb{E}\big(\mathbb{1}_B(h^+ - h^-)\big) \\ &= \mathbb{E}\big(\mathbb{1}_B h^+\big) - \mathbb{E}\big(\mathbb{1}_B h^-\big) \quad \text{by linearity of } \mathbb{E}(\,\cdot\,) \\ &= \mathbb{E}\big(\mathbb{1}_B X^+\big) - \mathbb{E}\big(\mathbb{1}_B X^-\big) \quad \text{by Step 3} \\ &= \mathbb{E}(\mathbb{1}_B X) \quad \text{by linearity once more.}\end{aligned}$$

■

Corollary 18.2.6 For any measurable $h\colon (E,\mathcal{E}) \to (\mathbb{R},\mathcal{B}_\mathbb{R})$ s.t. $\mathbb{E}(|h(X)|) < \infty$, one has $\forall_\mathbb{P}\,\omega \in \Omega$ that

$$\mathbb{E}[h(X)\,|\,\mathcal{G}](\omega) = \int_\Omega h(X(\omega'))\,\mathrm{d}\kappa(\cdot,\omega)(\omega').$$

18.3 I'm Disintegrating

The setup of this section is as follows:

$$(\Omega, \mathcal{F}, \mathbb{P}) \xrightarrow{\quad Y \quad} (E, \mathcal{E}, \mathbb{P}_Y)$$

where Y is $(\mathcal{F}, \mathcal{E})$-measurable. The overall goal is to *disintegrate* our probability measure $\mathbb{P}$ i.e. to find a **disintegration of $\mathbb{P}$ with respect to $\sigma(Y)$** — a (conditional-)probability-measure-valued map $y \longmapsto \mathbb{P}^y$ that is $\mathcal{E}$-measurable and satisfies

$$\mathbb{P} = \int \mathbb{P}^y \, \mathrm{d}\mathbb{P}_Y(y),$$

and $\mathbb{P}^y$ is concentrated on $Y^{-1}(\{y\})$ for $\mathbb{P}_Y$-a.e. y.
This integral condition will later present itself for any $A \in \mathcal{F}$ and $D \in \mathcal{E}$ as

$$\mathbb{P}\big(A \cap Y^{-1}(D)\big) = \int_D \mathbb{P}^y(A) \, \mathrm{d}\mathbb{P}_Y(y).$$

The above integral is a type of Fubini integral in the sense that:

- we first compute the conditional probability measures of A on the level sets (fibres) $Y^{-1}(\{y\})$ of Y
- and then integrate the result over a measurable subset $D \in \mathcal{E}$ of the "base space"[4] E with respect to $\mathbb{P}^y$ in the variable y.

The requirement that makes this possible is the existence of a regular conditional probability on $\mathcal{F}$ given $\mathcal{G} = \sigma(Y)$:

Theorem 18.3.1 (Theorem 10.4.5 [14]) Suppose that $\mathcal{F}$ is countably generated and that $\mathbb{P}$ has a compact approximating class in $\mathcal{F}$. Then for every sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, there exists a regular conditional probability on $\mathcal{F}$ given $\mathcal{G}$.

Proof. A proof can be found in [14, pp. 359–361]. I leave this empty space for when I work through it myself, and any comments I may have. ■

Let's investigate what a regular conditional probability on $\mathcal{F}$ given $\mathcal{G} = \sigma(Y)$ looks like:

Definition 18.3.2 A function $\kappa \colon \mathcal{F} \times \Omega \to [0, 1]$ is called a **regular conditional probability on $\mathcal{F}$ given $\mathcal{G} = \sigma(Y)$** if:

1. For every $\omega \in \Omega$, $\kappa(\cdot, \omega)$ is a probability measure on $(\Omega, \mathcal{F})$.
2. For each $A \in \mathcal{F}$, the mapping $\omega \longmapsto \kappa(A, \omega)$ is a version of $\mathbb{P}[A \,|\, \sigma(Y)]$ i.e.
   - $\kappa(A, \cdot)$ is $\sigma(Y)$-measurable, agrees $\mathbb{P}$-a.s. with $\mathbb{P}[A \,|\, \sigma(Y)]$, and
   - satisfies the averaging property $\forall B \in \sigma(Y)$:

$$\mathbb{P}(A \cap B) = \mathbb{E}(\mathbb{1}_B \mathbb{1}_A) = \mathbb{E}(\mathbb{1}_B \kappa(A, \cdot)) = \int_B \kappa(A, \omega) \, \mathrm{d}\mathbb{P}(\omega).$$

Pretty much the same as the general definition but we slot in $\sigma(Y)$ for $\mathcal{G}$. However, we can do more:

Remarks 18.3.3

[4] I believe in the language of category theory, one can think of Ω as being layered over E.

- Since $B \in \sigma(Y) = \{Y^{-1}(D) \colon D \in \mathcal{E}\}$, there exists some $D \in \mathcal{E}$ s.t. $B = Y^{-1}(D)$ and we may re-write the averaging property for each $A \in \mathcal{F}$ and $D \in \mathcal{E}$ as

$$\mathbb{P}\big(A \cap Y^{-1}(D)\big) = \underbrace{\mathbb{E}\big(\mathbb{1}_{Y^{-1}(D)}\mathbb{1}_A\big)}_{=\mathbb{E}(\mathbb{1}_D(Y)\mathbb{1}_A)} = \mathbb{E}\big(\mathbb{1}_{Y^{-1}(D)}\,\kappa(A,\cdot)\big) = \int_{Y^{-1}(D)} \kappa(A,\omega)\,\mathrm{d}\mathbb{P}(\omega).$$

- It's clear to see that when $(E,\mathcal{E}) = (\Omega,\mathcal{G})$, and $Y = \mathrm{id}\colon (\Omega,\mathcal{F}) \to (\Omega,\mathcal{G})$, then κ reduces to **Definition 18.2.3**.

We can see the rightmost integral begin to take the form of the Fubini-type integral at the beginning of this section. If we can somehow write the $\sigma(Y)$-*measurable* function $\omega \mapsto \kappa(A,\omega)$ as *a function* of Y, it almost looks like a change of variable away...

Indeed, this is the key observation. The $\sigma(Y)$-measurable function $\omega \mapsto \kappa(A,\omega)$ factors through Y (via Doob-Dynkin) to give a re-parameterisation of $\kappa\colon \mathcal{F}\times\Omega \to [0,1]$ to $\kappa_Y\colon \mathcal{F}\times E \to [0,1]$.

Explicitly, since $\kappa(A,\cdot)$ is a $\mathcal{G}$-measurable function, the setup of Theorem A.42 is as follows:

$(S_1,\mathcal{A}_1) = (\Omega,\mathcal{F})$

$(S_2,\mathcal{A}_2) = (E,\mathcal{E})$

$(S_3,\mathcal{A}_3) = ([0,1],\mathcal{B}_{[0,1]})$ is a measurable space whose σ-algebra contains all singletons

$f = Y \in \mathrm{Meas}_{\mathcal{F},\mathcal{E}}(\Omega\,;E)$

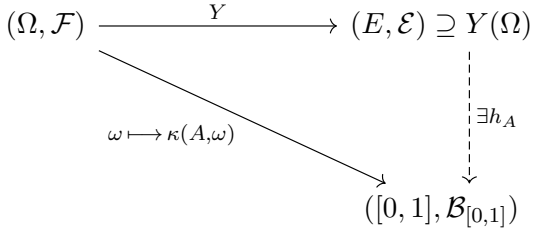
$g = (\omega \mapsto \kappa(A,\cdot))$

$\mathcal{A}_* = \{Y(\Omega)\cap D\colon D\in\mathcal{E}\}$

Then the theorem states that

$$g \in \mathrm{Meas}_{\sigma(Y),\mathcal{B}_{[0,1]}}(\Omega\,;[0,1]) \iff \exists h_A \in \mathrm{Meas}_{\mathcal{A}_*,\mathcal{B}_{[0,1]}}(Y(\Omega)\,;[0,1]) \text{ s.t. } \kappa(A,\cdot) = h_A\circ Y.$$

Pictorially:



$\forall \omega \in \Omega$ we have that $\kappa(A,\omega) = h_A(Y(\omega))$.

Let $y \in Y(\Omega)$. A fundamental observation here is that for any $\omega_1,\omega_2 \in Y^{-1}(\{y\})$, we have that

$$\kappa(A,\omega_1) = h_A(Y(\omega_1)) = h_A(y) = h_A(Y(\omega_2)) = \kappa(A,\omega_2)$$

and so we conclude that $\kappa(A,\cdot)$ is constant on $Y^{-1}(\{y\})$. This allows us to define, for each $y \in Y(\Omega)$:

$$\kappa_Y(A,y) \coloneqq h_A(y).$$

Notice that this re-parameterisation only applies to $y \in Y(\Omega)$. We wish to extend this to all of $y \in E$ so that we may have a satisfactory theory of (dis)integration over any element of $\mathcal{E}$. Furthermore, $Y(\Omega)$ may not be an element of $\mathcal{E}$, putting the measurability into question. Once this is rectified and we've extended to all of E, we will have defined the following object:

Definition 18.3.4 A **system of regular conditional probabilities generated by** $Y \in \mathrm{Meas}_{\mathcal{F},\mathcal{E}}(\Omega\,;E)$ is a function $\kappa_Y\colon \mathcal{F}\times E\to[0,1]$ such that:

1. For every $y\in E$, $\kappa_Y(\cdot,y)$ is a measure on $\mathcal{F}$,
2. For each $A\in\mathcal{F}$, the mapping $y\longmapsto\kappa_Y(A,y)$ is:
   - an $\mathcal{E}$-measurable function satisfying for $\mathbb{P}$-a.e. $\omega\in\Omega$:
   $$\kappa_Y(A,Y(\omega))=\mathbb{P}[A\,|\,\sigma(Y)](\omega),$$
   - and is $\mathbb{P}_Y$-integrable, satisfying the following **disintegration formula**: For all $A\in\mathcal{F}$ and $D\in\mathcal{E}$:
   $$\mathbb{P}\big(A\cap Y^{-1}(D)\big)=\int_D \kappa_Y(A,y)\,\mathrm{d}\mathbb{P}_Y(y).$$

Indeed, the remainder of the construction relies on some topological assumptions about our spaces. For completeness, we collect all the assumptions in one theorem:

Theorem 18.3.5 (Theorem 10.4.8 [14]) Let $Y\colon(\Omega,\mathcal{F})\to(E,\mathcal{E})$ be measurable with respect to $(\mathcal{F}_{\mathbb{P}},\mathcal{E})$. Suppose further that $Y(\Omega)\in\mathcal{E}_{\mathbb{P}_Y}$, where $\mathcal{E}_{\mathbb{P}_Y}$ is the completion of $\mathcal{E}$ with respect to the subscripted measure. If $\mathcal{F}$ is countably generated and $\mathbb{P}$ has a compact approximating class in $\mathcal{F}$, then there exists a system of regular conditional probability measures generated by Y on $\mathcal{F}$.

Proof. Fix $A\in\mathcal{F}$ and let $\mathcal{G}=\sigma(Y)\subseteq\mathcal{F}$ be a sub-σ-algebra. The assumptions of Theorem 10.4.5 are satisfied so there exists a regular conditional probability κ on $\mathcal{F}$ given $\mathcal{G}$. As previously explained, by Theorem A.42 for any $A\in\mathcal{F}$ the map $\kappa(A,\cdot)$ factors through Y i.e.

$$\exists h_A\in\mathrm{Meas}_{\mathcal{A}_*,\mathcal{B}_{[0,1]}}(Y(\Omega)\,;[0,1])\text{ s.t. }\kappa(A,\omega)=h_A(Y(\omega))\text{ for every }\omega\in\Omega.$$

For the topological assumption, since $Y(\Omega)\in\mathcal{E}_{\mathbb{P}_Y}$, there exist measurable sets $E_0,E_1\in\mathcal{E}$ s.t. $E_0\subseteq Y(\Omega)\subseteq E_1$ with $\mathbb{P}_Y(E_1\setminus E_0)=0$. Therefore, $\mathbb{P}_Y(Y(\Omega)\setminus E_0)=0$ and we conclude that E_0 is a measurable subset of $Y(\Omega)$ with full $\mathbb{P}_Y$ measure.

Construction of κ_Y**.**

For $y\in E_0$:

- Define $\kappa_Y(A,y):=h_A(y)$ for any $y\in E_0$. Now note that
  $$\kappa_Y(A,y)=h_A(Y(\omega))=\kappa(A,\omega).$$
  This is well-defined by the earlier observation about $\kappa(A,\cdot)$ being constant on the fibres $Y^{-1}(\{y\})$. Namely:
  - Thus, for any $y\in E_0$, $A\longmapsto\kappa_Y(A,y)=\kappa(A,\omega)$ is a probability measure.
  - The trace $(\mathcal{A}_*,\mathcal{B}_{[0,1]})$-measurability comes from h_A.

For the remaining points $y\in E\setminus E_0$, we can define $\kappa_Y(\cdot,y)$ arbitrarily because $E\setminus E_0$ is a $\mathbb{P}_Y$-null set. Let $\kappa_Y(\cdot,y)=\mathbb{P}$ for these y. Note:

- For any $A\in\mathcal{F}$, the mapping $y\mapsto\kappa(A,y)=\mathbb{P}(A)$ is constant and hence measurable because $E\setminus E_0\in\mathcal{E}$.
- The mapping $A\longmapsto\kappa_Y(A,y)=\mathbb{P}(A)$ is clearly a probability measure on $\mathcal{F}$.

Disintegration Formula.

For the averaging property, let $A \in \mathcal{F}$ and $D \in \mathcal{E}$. We wish to show that

$$\mathbb{P}\big(A \cap Y^{-1}(D)\big) = \int_D \kappa_Y(A, y)\, \mathrm{d}\mathbb{P}_Y(y).$$

Note that

$$\begin{aligned}
\int_D \kappa_Y(A, y)\, \mathrm{d}\mathbb{P}_Y(y) &= \int_{D\cap E_0} \kappa_Y(A, y)\, \mathrm{d}\mathbb{P}_Y(y) + \cancelto{0}{\int_{D\cap(E\setminus E_0)} \kappa_Y(A, y)\, \mathrm{d}\mathbb{P}_Y(y)} \\
&= \int_{D\cap E_0} h_A(y)\, \mathrm{d}\mathbb{P}_Y(y) \\
&\overset{\text{CVF}}{=} \int_{Y^{-1}(D\cap E_0)} h_A(Y(\omega))\, \mathrm{d}\mathbb{P}(\omega) \\
&= \int_{Y^{-1}(D\cap E_0)} \kappa(A, \omega)\, \mathrm{d}\mathbb{P}(\omega) \\
&= \mathbb{E}\big(\mathbb{1}_{Y^{-1}(D\cap E_0)}\, \kappa(A, \cdot)\big) \\
&= \mathbb{E}\big(\mathbb{1}_{Y^{-1}(D\cap E_0)}\, \mathbb{1}_A\big) \quad \text{averaging property of } \mathbb{P}[A \,|\, \sigma(Y)] \\
&= \mathbb{P}\big(A \cap Y^{-1}(D \cap E_0)\big) \\
&= \mathbb{P}\big(A \cap Y^{-1}(D)\big) \quad \text{since } E_0 \text{ is a subset of full measure.}
\end{aligned}$$

■

Remarks 18.3.6

- To emphasise that each $\kappa_Y(\cdot, y)$ is a probability measure, we often denote them by $\mathbb{P}^y(\cdot)$ and so the collection $\{\mathbb{P}^y\}_{y\in E}$ is the system of conditional probability measures that disintegrates $\mathbb{P}$.
- If we let $D = E$, then the averaging property takes the form

  $$\mathbb{P}(A) = \int_E \kappa_Y(A, y)\, \mathrm{d}\mathbb{P}_Y(y)$$

  which is precisely the disintegration identity we set out to achieve at the start of this section.

18.3.1 Extending the Disintegration Formula

A chain of equalities written in a particular way helped me to realise how to proceed from this point.

$$\begin{aligned}
\mathbb{E}\big(\mathbb{1}_{Y^{-1}(D)}\, \mathbb{E}[\mathbb{1}_A \,|\, \sigma(Y)]\big) &= \mathbb{E}\big(\mathbb{1}_{Y^{-1}(D)}\, \mathbb{1}_A\big) \quad \text{by the averaging property} \\
&:= \mathbb{P}\big(A \cap Y^{-1}(D)\big) \\
&= \int_\Omega \mathbb{1}_{Y^{-1}(D)}(\omega)\mathbb{1}_A(\omega)\, \mathrm{d}\mathbb{P}(\omega) \quad \text{by the disintegration formula} \\
&= \int_D \kappa_Y(A, y)\, \mathrm{d}\mathbb{P}_Y(y) \\
&= \int_D \left(\int_\Omega \mathbb{1}_A(\omega')\, \mathrm{d}\kappa_Y(\cdot, y)(\omega') \right) \mathrm{d}\mathbb{P}_Y(y)
\end{aligned}$$

One can extend this equality from indicators $\mathbb{1}_A$ to any integrable random variable $X \colon \Omega \to \mathbb{R}$.

Theorem 18.3.7 (Integral Form of Disintegration Theorem) For any integrable $X\colon \Omega \to \mathbb{R}$ and for any $D \in \mathcal{E}$:

$$\int_{Y^{-1}(D)} X(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \int_D \left(\int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\right)\mathrm{d}\mathbb{P}_Y(y).$$

Before proving this, I will make an immediate remark. Note that the LHS of this integral form is $\mathbb{E}\big(\mathbb{1}_{Y^{-1}(D)}\,X\big)$ which is equal to $\mathbb{E}\big(\mathbb{1}_{Y^{-1}(D)}\,\mathbb{E}[X\,|\,\sigma(Y)]\big)$. Thus, we may write

$$\int_{Y^{-1}(D)} \mathbb{E}[X\,|\,\sigma(Y)]\,\mathrm{d}\mathbb{P}(\omega) = \int_D \left(\int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\right)\mathrm{d}\mathbb{P}_Y(y).$$

Since $\mathbb{E}[X\,|\,\sigma(Y)]$ is $\sigma(Y)$-measurable, by **Theorem 16.1.1** there exists some

$$h_Y \in \mathrm{Meas}_{\mathcal{E}|_{Y(\Omega)},\mathbb{R}}(Y(\Omega)\,;\mathbb{R}) \text{ s.t. } \mathbb{E}[X\,|\,\sigma(Y)] = h_Y \circ Y.$$

This means that the LHS can be written as

$$\begin{aligned}\int_{Y^{-1}(D)} \mathbb{E}[X\,|\,\sigma(Y)]\,\mathrm{d}\mathbb{P}(\omega) &= \int_{Y^{-1}(D)} (h_Y \circ Y)(\omega)\,\mathrm{d}\mathbb{P}(\omega)\\ &= \int_D h_Y(y)\,\mathrm{d}\mathbb{P}_Y(y) \quad \text{by the CVF with } y = Y(\omega)\end{aligned}$$

from which we conclude that for $\mathbb{P}_Y$-a.e. $y \in E$

$$h_Y(y) = \int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega').$$

Definition 18.3.8 We call the assignment h_Y the **pointwise conditional expectation function** of X given $Y = y$, and we also denote it by

$$y \longmapsto h_Y(y) := \int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega') =: \mathbb{E}[X\,|\,Y = y].$$

Proof of Theorem 18.3.7.

1. For an indicator function $X = \mathbb{1}_A$, where $A \in \mathcal{F}$, the claim has already been shown.
2. Let $X = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ be a simple, measurable function. Then:

$$\begin{aligned}\int_{Y^{-1}(D)} X(\omega)\,\mathrm{d}\mathbb{P}(\omega) &= \sum_{i=1}^n a_i \int_{Y^{-1}(D)} \mathbb{1}_{A_i}(\omega)\,\mathrm{d}\mathbb{P}(\omega)\\ &= \sum_{i=1}^n a_i \int_D \left(\int_\Omega \mathbb{1}_{A_i}(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\right)\mathrm{d}\mathbb{P}_Y(y) \quad \text{by Step 1}\\ &= \int_D \left(\int_\Omega \sum_{i=1}^n a_i \mathbb{1}_{A_i}(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\right)\mathrm{d}\mathbb{P}_Y(y) \quad \text{by linearity}\\ &= \int_D \left(\int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\right)\mathrm{d}\mathbb{P}_Y(y)\end{aligned}$$

3. Let X be non-negative and measurable. Then X is the pointwise limit of a non-decreasing

sequence of simple functions X_n.

$$\begin{aligned}
\int_{Y^{-1}(D)} X(\omega)\,\mathrm{d}\mathbb{P}(\omega) &= \int_{Y^{-1}(D)} \lim_{n\to\infty} X_n(\omega)\,\mathrm{d}\mathbb{P}(\omega) \\
&= \lim_{n\to\infty} \int_{Y^{-1}(D)} X_n(\omega)\,\mathrm{d}\mathbb{P}(\omega) \quad \text{by the MCT on } \{\mathbb{1}_{Y^{-1}(D)}X_n\}_{n\in\mathbb{N}} \\
&= \lim_{n\to\infty} \int_D \underbrace{\left(\int_\Omega X_n(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\right)}_{=:\,g_n(y)} \mathrm{d}\mathbb{P}_Y(y) \quad \text{by Step 2 with } X_n\text{-simple} \\
&= \int_E \lim_{n\to\infty} \mathbb{1}_D(y) g_n(y)\,\mathrm{d}\mathbb{P}_Y(y) \quad \text{by the MCT on } \{\mathbb{1}_D g_n\}_{n\in\mathbb{N}} \\
&= \int_D \int_\Omega \lim_{n\to\infty} X_n(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\,\mathrm{d}\mathbb{P}_Y(y) \quad \text{by the MCT} \\
&= \int_D \int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\,\mathrm{d}\mathbb{P}_Y(y)
\end{aligned}$$

4. For the final step, let X be $\mathbb{P}$-integrable. Then we can, as usual, write it as the difference of two non-negative measurable functions $X = X^+ - X^-$. Then,

$$\begin{aligned}
\int_{Y^{-1}(D)} X(\omega)\,\mathrm{d}\mathbb{P}(\omega) &= \int_{Y^{-1}(D)} X^+(\omega)\,\mathrm{d}\mathbb{P}(\omega) - \int_{Y^{-1}(D)} X^-(\omega)\,\mathrm{d}\mathbb{P}(\omega) \\
&= \int_D \int_\Omega (X^+(\omega') - X^-(\omega'))\,\mathrm{d}\mathbb{P}^y(\omega')\,\mathrm{d}\mathbb{P}_Y(y) \quad \text{by Step 3, and linearity} \\
&= \int_D \int_\Omega X(\omega')\,\mathrm{d}\mathbb{P}^y(\omega')\,\mathrm{d}\mathbb{P}_Y(y)
\end{aligned}$$

■

Definition 18.3.9 If for $\mathbb{P}_Y$-a.e. $y \in E$ we have $Y^{-1}(\{y\}) \in \mathcal{F}$ and the measure $\mathbb{P}^y$ is concentrated[5] on $Y^{-1}(\{y\})$, then we call the $\mathbb{P}^y$ **proper conditional measures**.

Remarks 18.3.10 It's not always the case that for $\mathbb{P}_Y$-a.e. $y \in E$ we have that $Y^{-1}(\{y\}) \in \mathcal{F}$. This is because the measurability of $Y^{-1}(\{y\})$ depends on whether $\{y\} \in \mathcal{E}$. If one is in the situation where $(E, \mathcal{E})$ is a standard Borel space, then singletons are Borel and hence measurable, so $Y^{-1}(\{y\}) \in \mathcal{F}$.

Remarks 18.3.11

- Theorem 10.4.8 is 10.4.5 for a generated sub-σ-algebra.
- Note again that these probability measures may not be concentrated on the sets $Y^{-1}(\{y\})$ and the latter may not even be measurable. A sufficient condition for the existence of proper regular conditional probability measures is as follows:

Corollary 18.3.12 Suppose that in Theorem 10.4.8, $\mathcal{E}$ is countably generated and contains all singletons. Then there exist regular conditional probabilities $\mathbb{P}^y$ on the σ-algebra $\mathcal{F}'$ generated by $\mathcal{F}$ and $Y^{-1}(\mathcal{E})$ s.t. for $\mathbb{P}_Y$-a.e. y the measure $\mathbb{P}^y$ is concentrated on the set $Y^{-1}(\{y\})$.

Furthermore, if Y has a version $\tilde{Y} \in \mathrm{Meas}_{\mathcal{F},\mathcal{E}}(\Omega; E)$ s.t. $\tilde{Y}(\mathcal{F}) \subseteq \mathcal{E}_{\mathbb{P}_Y}$ then $Y^{-1}(\{y\}) \in \mathcal{F}_{\mathbb{P}^y}$ for $\mathbb{P}_Y$-a.e. y.

[5] Let μ be a measure on a measurable space $(X, \mathcal{F})$, and $A \in \mathcal{F}$. It's said that μ **is concentrated on** A if $\mu(X \setminus A) = 0$.

18.3.2 Conditional law of X given Y

Let $\kappa_Y : \mathcal{F} \times E \to [0,1]$ be a system of regular conditional probabilities on $\mathcal{F}$ generated by $Y \in \mathrm{Meas}_{\mathcal{F},\mathcal{E}}(\Omega; E)$. **By restricting to $\mathcal{F} = \sigma(X)$ and re-labelling our events**, we obtain the conditional law of X given Y. More precisely, if we define

$$\kappa_Y(\underbrace{X^{-1}(B)}_{\in \mathcal{F}}, y) =: \kappa_Y^X(B, y)$$

for $B \in \mathcal{S}$, $y \in E$, then κ_Y^X is a system of regular conditional probabilities of X given Y:

A **system of regular conditional probabilities of $X : (\Omega, \mathcal{F}) \to (E_X, \mathcal{E}_X)$ given $Y : (\Omega, \mathcal{F}) \to (E_Y, \mathcal{E}_Y)$** is a map $\kappa_Y^X : \mathcal{E}_Y \times E_X \to [0,1]$ s.t.

1. For every $y \in E_Y$, $\kappa_Y^X(\cdot, y)$ is a probability measure on $(E_X, \mathcal{E}_X)$.
2. For each $B \in \mathcal{E}_X$, the mapping $y \longmapsto \kappa_Y^X(B, y)$ is:
   - an $\mathcal{E}_Y$-measurable function satisfying for $\mathbb{P}$-a.e. $\omega \in \Omega$:
   $$\mathbb{P}\big[X^{-1}(B) \,\big|\, \sigma(Y)\big](\omega) = \kappa_Y^X(B, Y(\omega)),$$
   - and is $\mathbb{P}_Y$-integrable, satisfying for all $B \in \mathcal{E}_X$ and $D \in \mathcal{E}_Y$:
   $$\mathbb{P}\big(X^{-1}(B) \cap Y^{-1}(D)\big) = \int_D \kappa_Y^X(B, y)\, \mathrm{d}\mathbb{P}_Y(y).$$

18.3.3 Push-forward of Markov kernels

I encountered this fairly recently (January 2026) when studying the ordinary linear regression model. Some assumptions of joint conditional normality of errors $\boldsymbol{\varepsilon}$ given predictors $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ are made, and I needed this machinery to make rigorous statements about the related conditional distribution of $Y^{(i)} = \boldsymbol{\theta}^\top \mathbf{X}^{(i)} + \varepsilon^{(i)}$ by using an affine transformation to push the conditional law of $\boldsymbol{\varepsilon}$ given the predictors forward.

The following definition offers a short and sweet way to refer to the mouthful that is a 'regular conditional probability of blah given blah.'

Definition 18.3.13 Let $(E_X, \mathcal{E}_X)$ and $(E_Y, \mathcal{E}_Y)$ be measurable spaces. A **Markov (or transition) kernel from E_X to E_Y** is a map $\kappa : \mathcal{E}_Y \times E_X \to \mathbb{R}$ s.t.

- for every $x \in E_X$, $\kappa(\cdot, x)$ is a probability measure on $(E_Y, \mathcal{E}_Y)$, and
- for every $B \in \mathcal{E}_Y$, $\kappa(B, \cdot)$ is $\mathcal{E}_X$-measurable.

Theorem 18.3.14 Given a jointly measurable map $Z : E_Y \times E_X \to E_Z$, the function $\kappa^Z : \mathcal{E}_Z \times E_X \to [0,1]$ defined for $B \in \mathcal{E}_Z$ and $x \in E_X$ by

$$\kappa^Z(B, x) \coloneqq \kappa(\{y \in E_Y : Z(y, x) \in B\}, x)$$

is a Markov kernel from E_X to E_Z.

Proof.

- Let $x \in E_X$. We wish to show that $\kappa^Z(\cdot, x)$ is a probability measure on $E_Z, \mathcal{E}_Z$. Let $B \in \mathcal{E}_Z$,

and denote by $\iota_x \colon E_Y \hookrightarrow E_Y \times E_X$ the inclusion map defined by $\iota_x(y) = (y, x)$. Then

$$\begin{aligned}
\kappa^Z(B, x) &:= \kappa(\{y \in E_Y \colon Z(y, x) \in B\}, x) \\
&= \kappa(\{y \in E_Y \colon Z(\iota_x(y)) \in B\}, x) \\
&= \kappa(\{y \in E_Y \colon (Z \circ \iota_x)(y) \in B\}, x) \\
&= \kappa((Z \circ \iota_x)-1(B)), x) \\
&= ((Z \circ \iota_x)_\sharp \kappa(\cdot, x))(B)
\end{aligned}$$

i.e. we conclude that $\kappa^Z(\cdot, x)$ is the push-forward of the probability measure $\kappa(\cdot, x)$ on $(E_Y, \mathcal{E}_Y)$ via the $\mathcal{E}_Y$-$\mathcal{E}_Z$ measurable map $(Z \circ \iota_x) \colon E_Y \to E_Z$, and is therefore a probability measure on $(E_Z, \mathcal{E}_Z)$.

- Let $B \in \mathcal{E}_Z$. We wish to show that $\kappa^Z(B, \cdot)$ is $\mathcal{E}_X$-measurable. Let $A \in \mathcal{B}_{\mathbb{R}}$. If we can show that $\kappa^Z(B, \cdot)^{-1}(A) \in \mathcal{E}_Y$, then that demonstrates the claim.

$$\begin{aligned}
&= \kappa^Z(B, \cdot)^{-1}(A) \\
&= \{x \in E_X \colon \kappa^Z(B, x) \in A\} \\
&= \{x \in E_X \colon \kappa(\{y \in E_Y \colon Z(y, x) \in B\}, x) \in A\} \\
&= \kappa(\{y \in E_Y : Z(y, x) \in B\}, \cdot)^{-1}(A)
\end{aligned}$$

If we can show that **this set** is an element of $\mathcal{E}_Y$, then the claim follows from $\kappa(D, \cdot)$ being $\mathcal{E}_X$-measurable for any $D \in \mathcal{E}_Y$. Indeed, that set is equal to $(Z \circ \iota_x)^{-1}(B)) = \iota_x^{-1}(Z^{-1}(B))$, and the joint-measurability of Z guarantees that $Z^{-1}(B) \in \mathcal{E}_Y \otimes \mathcal{E}_X$, and its pre-image under ι_x gives an element of $\mathcal{E}_Y$. The claim is proven.

■

18.3.4 Example: Conditional Density Formula

Now to recover the traditional undergraduate conditional density formula.

Exercise 11 (Exercise 9.12.48 [14]) Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be probability spaces and $f \in L^1(\mu \otimes \nu)$. Show that the image of the measure $f \cdot (\mu \otimes \nu)$ under the natural projection $X \times Y \xrightarrow{\pi_1} X$ is given by the density

$$\varrho(x) = \int_Y f(x, y) \, \mathrm{d}\nu(y)$$

with respect to the measure μ.

Proof. The pushforward measure is defined for any $A \in \mathcal{A}$ by

$$\begin{aligned}
(\pi_1)_\sharp(f \cdot (\mu \otimes \nu))(A) &\coloneqq (f \cdot (\mu \otimes \nu))(\pi_1^{-1}(A)) \\
&= (f \cdot (\mu \otimes \nu))(A \times Y) \\
&\coloneqq \int_{A \times Y} f \, \mathrm{d}(\mu \otimes \nu) \\
&= \int_{X \times Y} \mathbb{1}_{A \times Y}(x, y) f(x, y) \, \mathrm{d}(\mu \otimes \nu)(x, y) \\
&= \int_{X \times Y} \mathbb{1}_A(x) f(x, y) \, \mathrm{d}(\mu \otimes \nu)(x, y) \\
&= \int_X \mathbb{1}_A(x) \int_Y f(x, y) \, \mathrm{d}\nu(y) \, \mathrm{d}\mu(x) \\
&= \int_A \varrho(x) \, \mathrm{d}\mu(x).
\end{aligned}$$

■

Lemma 18.3.15 (Conditional Density) Let $\mathbb{P} \ll \lambda_{\mathbb{R}^2}$ be a probability measure on $[0, 1]^2$, admitting density f. Then regular conditional measures with respect to the projection to the first coordinate axis have the form

$$\mathbb{P}^x(A) = \int_{\{y \colon (x,y) \in A\}} \frac{f(x, y)}{f_1(x)} \, \mathrm{d}y$$

where $x \in [0, 1]$ and

$$f_1(x) = \int_0^1 f(x, y) \, \mathrm{d}y.$$

Proof. We can put this in the framework of disintegration.

- $(\Omega, \mathcal{F}) = ([0, 1]^2, \mathcal{B}_{[0,1]^2})$
- $(E, \mathcal{E}) = ([0, 1], \mathcal{B}_{[0,1]})$
- $Y = \pi_1 \colon \Omega \to E$ is the projection onto the first coordinate

Note that $\mathcal{E} = \mathcal{B}_{[0,1]}$ in our example is countably generated and contains all singletons of $E = [0, 1]$. By the corollary above, there exists a system of regular conditional probabilities $\{\mathbb{P}^x\}_{x \in E = [0,1]}$ on $\mathcal{F}$ generated by $Y = \pi_1$, and this system is proper i.e. for $\mathbb{P}_{\pi_1}$-a.e. $x \in E = [0, 1]$, the probability measure $\mathbb{P}^x$ is concentrated on $\pi_1^{-1}(\{x\}) = \{x\} \times [0, 1]$. We use x in this example because we'll appeal to Euclidean space and x is the first coordinate.

By the disintegration theorem, for any $A \in \mathcal{F} = \mathcal{B}_{[0,1]^2}$ and $D \in \mathcal{E} = \mathcal{B}_{[0,1]}$:

$$\mathbb{P}\big(A \cap \pi_1^{-1}(D)\big) = \int_D \mathbb{P}^x(A) \, \mathrm{d}\mathbb{P}_{\pi_1}(x)$$

The LHS is

$$\begin{aligned}
&\mathbb{P}\big(A \cap \pi_1^{-1}(D)\big) \\
&= \mathbb{P}(A \cap (D \times [0,1])) \\
&= \int_{A \cap (D \times [0,1])} f \, \mathrm{d}\lambda_{\mathbb{R}^2} \quad \text{since } \mathbb{P} \text{ has density } f \\
&= \int_D \left(\int_{\{y\colon (x,y) \in A\}} f(x,y) \, \mathrm{d}y \right) \mathrm{d}x \quad \text{by Fubini's theorem.}
\end{aligned}$$

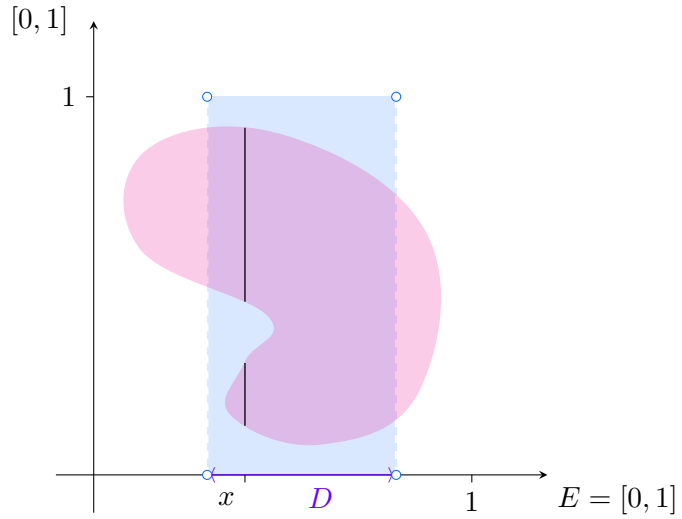


Figure 18.2: An example of a vertical slice (for fixed $x \in D$) in $A \cap (D \times [0,1])$ over which we integrate.

For the RHS, we use Exercise 9.12.48 [14]. Since $\mathbb{P} \ll \lambda_{\mathbb{R}^2}$, we can write it as $\mathbb{P}(A) = \int_A f \, \mathrm{d}\lambda^2 = \int_A f \, \mathrm{d}(\lambda \otimes \lambda)$ i.e. $\mathbb{P} = f \cdot (\lambda \otimes \lambda)$ and so we're in the situation where

- $(X, \mathcal{A}, \mu) = (Y, \mathcal{B}, \nu) = ([0,1], \mathcal{B}_{[0,1]}, \lambda)$
- f is the given density function of $\mathbb{P}$, and
- $f \cdot (\mu \otimes \nu) = f \cdot (\lambda \otimes \lambda) = \mathbb{P}$

Thus, the image measure $\mathbb{P}_{\pi_1}$ has density ϱ defined by

$$\varrho(x) = \int_{[0,1]} f(x,y) \, \mathrm{d}\lambda(y),$$

and so the RHS of our integral equality is given by

$$\int_D \mathbb{P}^x(A) \, \mathrm{d}\mathbb{P}_{\pi_1}(x) = \int_D \mathbb{P}^x(A) \varrho(x) \, \mathrm{d}x.$$

Finally, the LHS and RHS are equal as integrals, so their difference is equal to 0 and so their respective integrands are equal for λ-a.e. $x \in [0,1]$:

$$\int_{\{y\colon (x,y) \in A\}} f(x,y) \, \mathrm{d}y = \varrho(x) \mathbb{P}^x(A).$$

If $\rho(x)$ is equal to 0 for some x, we don't have enough information to determine $\mathbb{P}^x(A)$. However, as in the construction in Theorem 10.4.8, one may choose any probability measure as $\mathbb{P}^x$. ■

Remarks 18.3.16

- This lemma is Example 10.4.20 from [14, pp. 367–368].
- The attached comment from Bogachev is that the measure $\mathbb{P}^x$ is concentrated on the vertical interval $\{x\} \times [0,1]$, and is given by the density $y \longmapsto f(x,y)/f_1(x)$ with respect to the natural Lebesgue measure on this interval.
- As for my personal commentary:
  - The attached comment should say for all $x \in [0,1]$ s.t. $f_1(x) > 0$.
  - The example also states that 'we set $\frac{f(x,y)}{f_1(x)} = 0$ when $f_1(x) = 0$' but I think this is wrong or a throwaway comment instead of something rigorous. This is because the problem setup guarantees the existence of $\mathbb{P}^x$ for every $x \in [0,1]$, of which $\mathbb{P}_{\pi_1}$-almost every conditional probability measure is concentrated on its respective fibre $\pi_1^{-1}(\{x\})$ so up to some $\mathbb{P}_{\pi_1}$-null set of x, the disintegration formula is unaffected. Indeed, the proof of Theorem 10.4.8 says that we may simply set $\mathbb{P}^x$ to be some probability measure on $(\Omega, \mathcal{F})$. Setting the quotient to be zero would give you $\mathbb{P}^x$ as the zero measure but that's not a probability measure.

18.3.5 So many random variables

I'll make a previous exercise 7.72 [6], 7.98 [7] about the density of the t-distribution rigorous.

Example 18.3.17 Recall that we defined T to be the ratio $\frac{Z}{\sqrt{W/\nu}}$ where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2_\nu$, with Z and W independent.

(a) If W is fixed at w, then T is given by Z/c where $c = \sqrt{w/\nu}$. Use this idea to find the conditional density of T for a fixed $W = w$.

(b) Find the joint density of T and W using $f(t,w) = f(t \mid w) f(w)$.

Proof. The setup is that $\mathbb{P}$ represents the joint distribution of Z and W:

$$\begin{array}{ccc} (\Omega, \mathcal{F}, \mathbb{P}) & & \\ \Big\downarrow{\scriptstyle (T,W)} & \searrow^{W} & \\ (\mathbb{R} \times \mathbb{R}_{\geqslant 0}, \mathcal{B}_{\mathbb{R}} \otimes \mathcal{B}_{\mathbb{R}_{\geqslant 0}}, (T,W)_\sharp \mathbb{P}) & \xrightarrow[\pi_2]{} & (\mathbb{R}_{\geqslant 0}, \mathcal{B}_{\mathbb{R}_{\geqslant 0}}, (\pi_2)_\sharp((T,W)_\sharp \mathbb{P}) = \mathbb{P}_W) \end{array}$$

We can disintegrate in two equivalent ways because $W = \pi_2 \circ (T,W)$:

- Disintegrate $\mathbb{P}$ with respect to W
- Disintegrate the push-forward measure $(T,W)_\sharp \mathbb{P}$ with respect to the projection map onto the second coordinate.

Both approaches lead to the same integral equality i.e. $\forall B \in \mathcal{B}_{\mathbb{R}}$, $\forall D \in \mathcal{B}_{\mathbb{R}_{\geqslant 0}}$:

$$\mathbb{P}(\{T \in B\} \cap \{W \in D\}) = \mathbb{P}\big((T,W)^{-1}(B \times D)\big) = ((T,W)_\sharp \mathbb{P})(B \times D) = \int_D \mathbb{P}^w(T \in B) \, \mathrm{d}\mathbb{P}_W(w).$$

We are given that W admits a density $f_W(w)$ and so we may write

$$((T,W)_\sharp \mathbb{P})(B \times D) = \int_D \mathbb{P}^w(T \in B) f_W(w) \, \mathrm{d}\lambda(w).$$

If we can demonstrate that $\mathbb{P}^w(T \in \cdot)$ admits a density, then it follows that (T,W) admits a density, from which the undergraduate formula reveals itself:

- Begin by observing that $\mathbb{P}^w(T \in \cdot) = \mathbb{P}^w(g(Z) \in \cdot)$ is the push-forward $(g \circ Z)_\sharp \mathbb{P}^w$, where $g(z) = z/\sqrt{w/\nu}$.
- The mutual independence of Z and W tells us that the conditional density of Z given W remains unchanged. For any fixed w, $T = Z/\sqrt{w/\nu}$ and so $T = Z/c$ where $c = \sqrt{w/\nu}$ and $Z \sim \mathcal{N}(0,1)$.

Thus, we conclude that for every w the conditional probability measure $\mathbb{P}^w(T \in \cdot)$ of T given W is absolutely continuous with respect to the Lebesgue measure λ and its density is

$$f_{T|W}(t \mid w) = \sqrt{\frac{w}{2\pi\nu}} \exp\left(\tfrac{-w}{2\nu} t^2\right).$$

Therefore,

$$\begin{aligned}
((T,W)_\sharp \mathbb{P})(B \times D) &= \int_D \mathbb{P}^w(T \in B) f_W(w) \,\mathrm{d}\lambda(w) \\
&= \int_D \int_B \mathrm{d}\mathbb{P}^w(T \in \cdot)(\omega') f_W(w) \,\mathrm{d}\lambda(w) \\
&= \int_D \int_B f_{T|W}(t \mid w) \,\mathrm{d}\lambda(t) f_W(w) \,\mathrm{d}\lambda(w) \\
&= \int_{B \times D} f_{T|W}(t \mid w) f_W(w) \,\mathrm{d}(\lambda \otimes \lambda)(t,w)
\end{aligned}$$

so the joint density of T and W is given by

$$f_{(T,W)}(t,w) = f_{T|W}(t \mid w) f_W(w).$$

Finally, we follow the algebraic manipulations as before to demonstrate the relevant expressions. ■

A more general case follows.

CHAPTER 19

Disintegration (Separable Case)

Example 18.1.4 from earlier began with the conditional probability κ on $\mathcal{F}$ given a σ-algebra $\mathcal{G}$ generated by a partition $\{B_n\}_{j\in\{0\}\cup J\subseteq\mathbb{N}}$ of Ω, and it was later demonstrated in 18.2.2 that for each $A \in \mathcal{F}$, κ decomposes into a countable sum of conditional probability measures:

$$\kappa(A,\omega) = \mathbb{P}(A \mid B_0)\mathbb{1}_{B_0}(\omega) + \sum_{i\in J} \mathbb{P}(A \mid B_i)\mathbb{1}_{B_i}(\omega)$$

for $\mathbb{P}$-a.e. $\omega \in \Omega$. This decomposition allowed us to conclude that κ is regular.

The condition that $\mathcal{G}$ is generated by a partition is very specific and can be relaxed.

Definition 19.0.1 A σ-algebra is called **separable** if it's generated by a countable class of sets.

Example 19.0.2 Any σ-algebra generated by a partition is an *example* of a separable σ-algebra. The countable generating class of a separable σ-algebra need not be pairwise disjoint, and so the definition is more general.

We can make a slightly more general claim than the partition case that serves as a partial converse. Namely, if we suppose that there exists a separable sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, and that κ is a *regular* conditional probability, **then** there exists a partition of Ω and the decomposition of κ follows.

Proving this will be the aim of this chapter!

Some preliminary facts about the atomic structure of σ-algebras can't be avoided:

19.1 Atoms

Let $\mathcal{G}$ be a σ-algebra and $\mathcal{G}' \subseteq \mathcal{G}$ be a sub-σ-algebra.

Definition 19.1.1 A $\mathcal{G}$-atom is a set $A \in \mathcal{G}$ s.t.

$$\forall B \in \mathcal{G}: \quad \text{either } B \cap A = A \text{ or } B \cap A = \emptyset. \tag{$\star_1$}$$

Definition 19.1.2 A $\mathcal{G}$-atom is a set $A \in \mathcal{G}$ s.t.

$$\forall B \in \mathcal{G},\ B \subseteq A: \quad \text{either } B = A \text{ or } B = \emptyset. \tag{$\star_2$}$$

Claim The two definitions above are equivalent.

Proof.

$1 \implies 2$:

Suppose that A satisfies $(\star_1)$. Then take $B \in \mathcal{G}$ with $B \subseteq A$. This means that either

$$A \cap B = A \text{ or } A \cap B = \emptyset.$$

Coupled with $B \subseteq A$, we conclude that either

$$B = A \text{ or } B = \emptyset.$$

$\underline{2 \implies 1}$:

Suppose that A satisfies $(\star_2)$. Then take any $B \in \mathcal{G}$. We want to show that either

$$A \cap B = A \text{ or } A \cap B = \emptyset.$$

Note that $(A \cap B) \in \mathcal{G}$. Also, $(B \cap A) \subseteq A$. By $(\star_2)$, we have that

$$B \cap A = A \text{ or } (B \cap A) = \emptyset.$$

■

The crucial fact that will give us our decomposition is the following

Lemma 19.1.3 (Exercise 12, [15, p. 207]) If $\mathcal{G}$ is a separable σ-algebra, then there exist at most continuum many atoms $B_t \in \mathcal{G}$ (i.e. $B \subseteq B_t$, $B \in \mathcal{G} \implies B \in \{\emptyset, B_t\}$) such that $\bigcup_t B_t = \Omega$ and each element of $\mathcal{G}$ is a union of these atoms.

Proof. Let $\{F_j\}_{j \in J \subseteq \mathbb{N}}$ be the generators of $\mathcal{G}$. An outcome's location in Ω relative to the F_j is completely determined by its membership in every F_j i.e. we can define a mapping $\chi \colon \Omega \to \{0,1\}^J$ for any $\omega \in \Omega$ by

$$\chi(\omega) = (\mathbb{1}_{F_1}(\omega), \mathbb{1}_{F_2}(\omega), \ldots) \in \{0,1\}^J.$$

If two outcomes ω, ω' share the same value $\chi(\omega) = \chi(\omega')$, then they are indistinguishable from the perspective of the countable collection. This gives us a natural way to partition Ω. Define $\sim$ over Ω by $\omega \sim \omega' \iff \chi(\omega) = \chi(\omega')$. It follows immediately that $\sim$ is an equivalence relation. Thus, for any $t = (t_1, t_2, \ldots) \in \{0,1\}^J$ we may define the equivalence class

$$\begin{aligned} B_t \coloneqq \chi^{-1}(\{t\}) &= \{\omega \in \Omega \colon \mathbb{1}_{F_j}(\omega) = t_j \text{ for all } j \in J\} \\ &= \bigcap_{j \in J} E_j \end{aligned}$$

where each E_j is defined by

$$E_j = \begin{cases} F_j & \text{if } t_j = 1 \\ F_j^{\mathsf{c}} & \text{if } t_j = 0. \end{cases}$$

These sets B_t are pairwise disjoint and their union is all of Ω. Since every B_t is the countable intersection of $E_j \in \mathcal{G}$, and $\mathcal{G}$ is a σ-algebra, we conclude that $B_t \in \mathcal{G}$ and so the $\{B_t\}_t$ form a partition of Ω.

Now we wish to show that the B_t are $\mathcal{G}$-atoms. For any $A \in \mathcal{G}$, we wish to show that if $A \subseteq B_t$ then either $A = \emptyset$ or $A = B_t$. Since $A \in \mathcal{G}$, we may write it as $\chi^{-1}(S)$ for some Borel subset $S \subseteq \{0,1\}^J$. (Why? Why.[1]) If $A \neq \emptyset$, then $A \subseteq B_t$ means that $t \in S$. Thus, A also contains all points ω with $\chi(\omega) = t$ i.e. $B_t \subseteq A$. Therefore, $A = B_t$. Otherwise, A is empty. ■

Given two σ-algebras $\mathcal{G}' \subseteq \mathcal{G}$, there is a relationship between their atoms.

- If $A \in \mathcal{G}$ can't be decomposed (i.e. written as a disjoint union of elements of $\mathcal{G}$), then it certainly can't be decomposed by (less) sets in $\mathcal{G}' \subseteq \mathcal{G}$.

[1] This mapping is measurable with respect to $\mathcal{G}$. We must choose an appropriate σ-algebra to equip $\{0,1\}^J$ with. The natural choice, since we have a coordinate-wise map is the product σ-algebra $\mathcal{C}$ which generated by the cylindrical sets. This is the coarsest σ-algebra for which the coordinate maps $\pi_j \circ \chi = \mathbb{1}_{F_j}$ measurable. Indeed, by **Proposition 6.0.3** we conclude that

$$\chi \in \mathrm{Meas}_{\Omega,\, \{0,1\}^J}(\mathcal{G}\,; \sigma(\mathcal{C}))$$

and note that the pullback σ-algebra under χ is given by

$$\sigma(\chi) = \sigma(\{\chi^{-1}(C_j) \colon j \in J\}) = \sigma(\{F_j \colon j \in J\}) = \mathcal{G}.$$

- If such an $A \in \mathcal{G}$ is also declared to be an element of $\mathcal{G}'$, then A is also a $\mathcal{G}'$-atom.

Since κ is a conditional probability on $\mathcal{F}$ given $\mathcal{G}$, the functions $\omega \longmapsto \mathbb{P}[A \,|\, \mathcal{G}](\omega)$ are $\mathcal{G}$-measurable. In particular, this means that for every $S \in \mathcal{B}_{[0,1]}$, events of the form $(\mathbb{P}[A \,|\, \mathcal{G}])^{-1}(S)$ are elements of $\mathcal{G}$. The σ-algebra generated by the collection of such events

$$\mathcal{G}' := \sigma(\{(\mathbb{P}[A \,|\, \mathcal{G}])^{-1}(S) \colon S \in \mathcal{B}_{[0,1]}\}_{A \in \mathcal{F}})$$

is a sub-σ-algebra of $\mathcal{G}$.

Another consequence of the $\mathcal{G}$-measurability of the functions $\omega \longmapsto \mathbb{P}[A \,|\, \mathcal{G}](\omega)$ is that they reduce to constants on the atoms of $\mathcal{G}$. In fact, they reduce to constants on possibly larger events. Namely, on atoms of the σ-algebra $\mathcal{G}' \subseteq \mathcal{G}$. It suffices to take events of the form $(\mathbb{P}[A \,|\, \mathcal{G}])^{-1}((-\infty, r))$ where $r \in \mathbb{Q}_{>0}$.

Definition 19.1.4

- The atoms of $\mathcal{G}'$ will be called **$\kappa_{\mathcal{G}}$-atoms**.
- **Every event contained in a $\kappa_{\mathcal{G}}$-atom** will be called **$\kappa_{\mathcal{G}}$-indecomposable**.

19.2 Decomposition Theorem

Theorem 19.2.1 (Decomposition Theorem A [16, p. 22]) If κ is a regular conditional probability, and $\mathcal{G}$ contains a separable σ-algebra $\mathcal{G}'$ whose atoms are $\kappa_{\mathcal{G}}$-indecomposable, then there exists a partition

$$\Omega = N \sqcup \left(\bigsqcup_{t \in T} B_t \right)$$

with $T \subseteq \mathbb{R}$ and $\mathbb{P}(N) = 0$ s.t. except on $\mathcal{F} \times N$:

$$\kappa_{\mathcal{G}} = \sum_{t \in T} \mathbb{P}(\cdot \,|\, B_t) \mathbb{1}_{B_t}$$

where the $\mathbb{P}(\cdot \,|\, B_t)$ are probabilities on $\mathcal{F}$ and $\mathbb{P}(B_t \,|\, B_t) = 1$.

Proof. Denote by $\{F_j\}$ the countable class that generates $\mathcal{G}' \subseteq \mathcal{G}$. Since $\mathcal{G}'$ is separable, by **Lemma 19.1.3** there exists a partition

$$\Omega = \bigsqcup_{t \in T'} B_t$$

into atoms $B_t \in \mathcal{G}'$. Since $\mathbb{P}(\Omega) = 1$, at most countably many of these atoms have $\mathbb{P}(B_t) > 0$.

For each generator F_j, by definition of the r.c.p. we have that

$$\kappa_{\mathcal{G}}(F_j, \omega) = \mathbb{E}\big[\mathbb{1}_{F_j} \,\big|\, \mathcal{G}\big](\omega) = \mathbb{1}_{F_j}(\omega)$$

for $\mathbb{P}$-a.e. $\omega \in \Omega$ because $\mathbb{1}_{F_j}$ is $\mathcal{G}$-measurable. Denote by N_j the $\mathbb{P}$-null set on which that equality fails i.e.

$$N_j := \{\omega \in \Omega \colon \kappa_{\mathcal{G}}(F_j, \omega) \neq \mathbb{1}_{F_j}(\omega)\} \in \mathcal{G}.$$

Then their (countable) union $N = \bigcup_j N_j$ is an element of $\mathcal{G}$ and is also $\mathbb{P}$-null.

This means that for every generator F_j, and every $\omega \in \Omega \setminus N$, the following equality holds

$$\kappa_{\mathcal{G}}(F_j, \omega) = \mathbb{1}_{F_j}(\omega).$$

This defines an equality of measures on the generators of $\mathcal{G}'$. Thus, they agree on $\mathcal{G}'$. In particular, for every B_t and $\omega \in B_t \setminus N$:

$$\kappa_{\mathcal{G}}(\omega, B_t) = 1. \tag{19.1}$$

Every $\mathcal{G}'$-atom B_t is assumed to be $\kappa_{\mathcal{G}}$-indecomposable. By the $\kappa_{\mathcal{G}}$-indecomposability of every B_t, we know that for each $A \in \mathcal{F}$ the functions $\omega \longmapsto \kappa(A, \omega) = \mathbb{P}[A \,|\, \mathcal{G}](\omega)$ are almost surely constant on every B_t.

- Now let's consider only those atoms B_t for which $\mathbb{P}(B_t) > 0$. By the averaging property of conditional expectation, for each such B_t:

$$\begin{aligned}\mathbb{E}(\mathbb{1}_{B_t}\mathbb{1}_A) = \mathbb{P}(A \cap B_t) &= \int_\Omega \mathbb{P}[A \,|\, \mathcal{G}](\omega)\mathbb{1}_{B_t}(\omega)\,\mathrm{d}\mathbb{P}(\omega) = \mathbb{E}(\mathbb{P}[A \,|\, \mathcal{G}]\mathbb{1}_{B_t}) \\ &= \int_{B_t} \mathbb{P}[A \,|\, \mathcal{G}](\omega)\,\mathrm{d}\mathbb{P}(\omega) \\ &= \mathbb{P}[A \,|\, \mathcal{G}](\omega)\int_{B_t} \mathrm{d}\mathbb{P}(\omega) \text{ since it's constant a.s. on } B_t \\ &= \mathbb{P}[A \,|\, \mathcal{G}](\omega)\mathbb{P}(B_t)\end{aligned}$$

  i.e.

$$\forall_{\mathbb{P}}\,\omega \in B_t\colon\ \mathbb{P}(A \cap B_t) = \kappa(A,\omega)\mathbb{P}(B_t). \tag{19.2}$$

  Since $\mathbb{P}(B_t) > 0$, we can divide through to obtain the expression

$$\kappa(A,\omega) = \underbrace{\frac{\mathbb{P}(A \cap B_t)}{\mathbb{P}(B_t)}}_{=:\,\mathbb{P}_{B_t}(A).}$$

  Since κ is regular, each $\kappa(\cdot,\omega)$ is a probability measure on $\mathcal{F}$ for $\omega \in B_t \setminus N$, and so $\mathbb{P}_{B_t}$ is a probability measure on $\mathcal{F}$.

- For those atoms B_t that are $\mathbb{P}$-null, equation 19.2 breaks down. We cannot divide by 0. However, equation 19.1 tells us that $\kappa_{\mathcal{G}}$ assigns full conditional probability to B_t. We can define $\kappa_{\mathcal{G}}(\cdot,\omega)$ to be any probability measure $\mathbb{P}_{B_t}$ supported on B_t.

Thus, we arrive at the decomposition

$$\Omega = N \sqcup \bigsqcup_{t\in T} B_t$$

and the regular conditional probability can be written in the form

$$\kappa_{\mathcal{G}}(A,\omega) = \sum_{t\in T} \mathbb{P}_{B_t}(A)\mathbb{1}_{B_t}(\omega), \text{ for } \omega \in \Omega \setminus N.$$

■

Now I'll return to reading Chapter 6 from [1] — the book that kicked off this large tangent because I didn't know, rigorously, what conditional probability is.

CHAPTER 20

Approximation

The keen eye will notice that we've been following a single workflow:

- Suppose that a population Π follows a particular distribution — this is represented by a probability space. Then we formalise the act of drawing one element from Π by a random variable X defined on Π. The distribution of this random variable is the distribution of the population e.g. writing $X \sim \mathcal{N}(\mu, \sigma^2)$ means the population is normal.
- Sample from this population randomly by realising n i.i.d. copies of X (namely $X_1, \ldots, X_n$)
- Decide on a quantity/population parameter to estimate e.g. μ or σ^2 in this case. Let's say we focus on μ *assuming* σ^2 *is known.*
- Consider a statistic T so that the estimator $T(X_1, \ldots, X_n)$ can be used to estimate said parameter
  - e.g. $\overline{X} = T(X_1, \ldots, X_n) \coloneqq \dfrac{X_1 + \ldots + X_n}{n}$ to estimate μ.
- Investigate the sampling distribution of $T \circ \mathbf{X}$
- Use this sampling distribution to make some inference/decision
  - e.g. In **Example 15.1.2**, we rewrote $\mathbb{P}\big(|\overline{X} - \mu| \leqslant 0.3\big)$ as $\mathbb{P}(-0.9 \leqslant Z \leqslant 0.9)$ by observing that $Z \coloneqq (\overline{X} - \mu)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0, 1)$, and then read off a statistical table for the probability that the sample mean is within 0.3 ounces of the population (true) mean.

If σ is unknown, we can estimate σ with $S = \sqrt{S^2}$, consider $T(X_1, \ldots, X_n) = (\overline{X} - \mu)/(S/\sqrt{n})$, investigate its sampling distribution (which is the t_{n-1} distribution from last chapter) and then make some inference. Again, these steps follow the above framework.

20.1 Approximation

The behaviour of certain sample quantities as the sample size $n \to \infty$ can offer some useful approximations for the finite-sample case (because expressions often become simplified in the limit) despite an infinite sample being a theoretical artefact.

In particular, we can look at the behaviour of

$$\overline{X}_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} X_i$$

as n grows. Investigating this statistic will lead us to the main idea of this chapter — the central limit theorem which is concerned with giving an asymptotic (i.e. for large, fixed n) approximation for the sampling distribution of the (standardised) sample mean of a random sample of n observations drawn from **any** population (irrespective of the the population distribution).

We build up to it by discussing modes of convergence (of random variables) and some preliminary results on the behaviour of the sample mean $\overline{X}_n$ (of $X_1, \ldots, X_n$).

What follows is several sections on the types of convergence of random variables *defined on the same probability space.*

20.2 Convergence in Probability

Definition 20.2.1 A sequence of random variables $X_1, X_2, \ldots$ (without stipulation on their distribution) **converges in probability** to a random variable X if $\forall \varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geqslant \varepsilon) = 0.$$

Equivalently, we can write $\mathbb{P}(|X_n - X| < \varepsilon) \overset{n\to\infty}{\longrightarrow} 1$.

The following theorem will serve to make some sense of the above definition.

Theorem 20.2.2 (Weak Law of Large Numbers) Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then $\overline{X}_n$ converges in probability to (the constant random variable) μ i.e. $\forall \varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\big(|\overline{X}_n - \mu| < \varepsilon\big) = 1.$$

The weak law of large numbers says the measure (probability) of the set (of $\omega \in \Omega$) on which there are "large" ε deviations of $\overline{X}_n$ from μ goes to zero but this makes **no guarantee** that our sequence $(\overline{X}_n)_{n\in\mathbb{N}}$ **stays** within ε of μ forever after a certain point along the sequence.

Proof. Let $\varepsilon > 0$. Then

$$\begin{aligned}
\mathbb{P}\big(|\overline{X}_n - \mu| \geqslant \varepsilon\big) &= \mathbb{P}\big((\overline{X}_n - \mu)^2 \geqslant \varepsilon^2\big) \\
&\leqslant \frac{1}{\varepsilon^2}\,\mathbb{E}\big((\overline{X}_n - \mu)^2\big) \text{ by Chebyshev's Theorem} \\
&= \frac{\mathrm{Var}\big(\overline{X}_n\big)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}
\end{aligned}$$

By considering complements,

$$\mathbb{P}\big(|\overline{X}_n - \mu| < \varepsilon\big) = 1 - \mathbb{P}\big(|\overline{X}_n - \mu| \geqslant \varepsilon\big) \geqslant 1 - \frac{\sigma^2}{n\varepsilon^2} \overset{n\to\infty}{\longrightarrow} 1.$$

■

20.2.1 Consistency

The property exhibited in the Weak Law of Large Numbers, that a sequence of the "same" sample quantity ($\overline{X}_n$ in this case) converges in probability to a constant (μ in this case) as $n \to \infty$, is called **consistency** of the estimator. This is expanded upon in **Section 21.11.2**.
Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and finite variance $\sigma^2 < \infty$. Denote by

- $\overline{X}_k$ the **sample mean of the first k observations**

$$\overline{X}_k := \frac{1}{k}\sum_{i=1}^{k} X_k,$$

- and by S_k^2 the **sample variance of the first k observations**.

$$S_k^2 := \frac{1}{k-1}\sum_{i=1}^{k}(X_i - \overline{X}_k)^2.$$

Example 20.2.3 (Consistency of S^2) An analogous weak law holds for S_n^2, also by Chebyshev's theorem.

Theorem 20.2.4 Let g be a measurable function. If X_n converges to X in probability as $n \to \infty$, then $g(X_n)$ converges in probability to $g(X)$ as $n \to \infty$.

Corollary 20.2.5 Since S_n^2 is a consistent estimator of σ^2, let $g = \sqrt{\cdot}$ and then we observe that S_n is a consistent estimator of $\sigma = \sqrt{\sigma^2} =: g(\sigma^2)$.

Exercise 12 (Example 5.5.5 [1, p. 190] and Exercise 5.11 [1, p. 209]) S_n is a biased estimator of σ (but the bias disappears asymptotically).

The solution of this depends on Jensen's inequality which I shall present in the short, and self-contained subsection just below.

Proof.

$$\begin{aligned}\mathbb{E}(S_k) = \mathbb{E}\left(\sqrt{S_k^2}\right) &\leqslant \sqrt{\mathbb{E}\left(S_k^2\right)} \quad \text{by Jensen's Inequality} \\ &= \sqrt{\sigma^2} \quad \text{since } S_k^2 \text{ is an unbiased estimator of } \sigma^2 \\ &= \sigma\end{aligned}$$

■

20.2.2 Jensen's Inequality

Theorem 20.2.6 (Jensen's Inequality) Let X be a random variable.

- If $g(x)$ is a convex function, then $\mathbb{E}(g(X)) \geqslant g(\mathbb{E}(X))$.
- If $g(x)$ is a concave function, then $\mathbb{E}(g(X)) \leqslant g(\mathbb{E}(X))$.

Equality $\mathbb{E}(g(X)) \leqslant g(\mathbb{E}(X))$ holds iff $\mathbb{P}(\{g(X) = aX + b\}) = 1$ for every tangent line $ax + b$ to $g(x)$ at $x = \mathbb{E}(X)$.

Remarks 20.2.7 The condition 'for every tangent' line originally confused me but it's there to deal with situations where a tangent line isn't unique e.g. a cusp of the function, or the endpoints of a flat part of the function (at which the derivative isn't smooth).

Proof. Let $\ell(x) = ax + b$ be a tangent line to $g(x)$ at the point $x = \mathbb{E}(X)$. Since g is convex, $g(x) \geqslant \ell(x) = ax + b$ for all x. This means that for every outcome $\omega \in \Omega$, we have that $g(X(\omega)) \geqslant aX(\omega) + b$. By the monotonicity of the Lebesgue integral w.r.t. $\mathbb{P}$:

$$\begin{aligned}\mathbb{E}(g(X)) &\geqslant \mathbb{E}(aX + b) \\ &= a\mathbb{E}(X) + b \\ &=: \ell(\mathbb{E}(X)) \\ &= g(\mathbb{E}(X)) \text{ since } \ell \text{ is tangent to } g \text{ at } \mathbb{E}(X).\end{aligned}$$

Now for the equality criterion.

$\Longleftarrow$ Since $\ell(x) = ax + b$ is tangent to $g(x)$ at $x = \mathbb{E}(X)$, we have that both ℓ and g pass through the same point at $x = \mathbb{E}(X)$. By the convexity of g, $g(x) \geqslant \ell(x)$ for all x. Suppose that $\mathbb{P}(\{g(X) = aX + b\}) = 1$ i.e. $\forall_{\mathbb{P}}\, \omega \in \Omega\colon g(X(\omega)) = aX(\omega) + b$. They only differ on a set of probability zero so their expectations are equal i.e.

$$\begin{aligned}\mathbb{E}(g(X)) &= \mathbb{E}(\ell(X)) \\ &= a\mathbb{E}(X) + b \\ &=: \ell(\mathbb{E}(X)) \\ &= g(\mathbb{E}(X)) \quad \text{because they both pass through the same point.}\end{aligned}$$

$\Longrightarrow$ The "only if" direction is proven by contrapositive i.e. suppose that there exists some tangent line $\ell(x) = ax + b$ for which $\mathbb{P}(\{g(X) = aX + b\}) < 1$. Then we wish to demonstrate that $g(\mathbb{E}(X)) \neq \mathbb{E}(g(X))$.

To this end, note that since g is convex (and so $g(x) \geqslant \ell(x)$), and $g(X)$ isn't equal to $\ell(X)$ with probability 1, then $\mathbb{P}(\{g(X) > a + bX\}) > 0$. Now observe that we may partition our outcome space by

- $B := \{\omega \in \Omega \colon g(X(\omega)) > aX(\omega) + b\}$,
- and its complement $B^{\mathsf{c}} = \{\omega \in \Omega \colon g(X(\omega)) = aX(\omega) + b\}$

It follows that

$$\begin{aligned}
\mathbb{E}(g(X)) &= \int_{B \sqcup B^{\mathsf{c}}} g(X)\,\mathrm{d}\mathbb{P} \\
&= \int_B g(X)\,\mathrm{d}\mathbb{P} + \int_{B^{\mathsf{c}}} g(X)\,\mathrm{d}\mathbb{P} \\
&= \int_B g(X)\,\mathrm{d}\mathbb{P} + \int_{B^{\mathsf{c}}} (aX + b)\,\mathrm{d}\mathbb{P} \\
&> \int_B (aX + b)\,\mathrm{d}\mathbb{P} + \int_{B^{\mathsf{c}}} (aX + b)\,\mathrm{d}\mathbb{P} \\
&= \int_\Omega (aX + b)\,\mathrm{d}\mathbb{P} \\
&=: \mathbb{E}(aX + b) \\
&= a\mathbb{E}(X) + b \\
&=: \ell(\mathbb{E}(X)) \\
&= g(\mathbb{E}(X)) \quad \text{since they both pass through the same point at } x = \mathbb{E}(X).
\end{aligned}$$

■

20.3 Almost Sure Convergence

This mode of convergence is stronger than convergence in probability.

Definition 20.3.1 A sequence of random variables $X_1, X_2, \ldots$ converges **$\mathbb{P}$-almost surely** to X if $\forall \varepsilon > 0$:

$$\mathbb{P}\Big(\lim_{n\to\infty} |X_n - X| < \varepsilon\Big) = 1.$$

This is similar to point-wise convergence of functions but convergence need not happen on at most a set of probability zero. We also denote this convergence by $X_n \overset{n\to\infty}{\longrightarrow} X$ $\mathbb{P}$-a.s. or $X_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} X$ as $n \to \infty$.

Example Consider the probability space $(\Omega = [0,1], \mathcal{B}_{[0,1]}, \lambda_{[0,1]})$. Recall that the probability measure $\lambda_{[0,1]}$ is the uniform probability distribution. Define $X_n(\omega) = \omega + \omega^n$ and $X(\omega) = \omega$. For each $\omega \in [0,1)$, $\omega^n \to 0$ as $n \to \infty$ so $X_n(\omega) \to \omega = X(\omega)$ on $[0,1)$. However, for every $n \in \mathbb{N}$: $X_n(1) = 2 \neq 1 = X(1)$ so $X_n(1) \not\longrightarrow X(1)$. In summary, X_n converges to X on $[0,1)$ and $\lambda_{[0,1]}([0,1)) = 1$ so X_n converges $\lambda_{[0,1]}$-almost surely to X as $n \to \infty$.

Remarks 20.3.2

- The converse, that convergence in probability implies $\mathbb{P}$-almost sure convergence, is **not** true.
- However, a partial converse exists: If $X_1, X_2, \ldots$ converges to X in probability, then we can extract a subsequence that does converge $\mathbb{P}$-almost surely.

A stronger version of the weak law of large numbers holds.

Theorem 20.3.3 (Strong Law of Large Numbers) Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then $\overline{X}_n$ converges $\mathbb{P}$-almost surely to (the constant random variable) μ i.e. $\forall \varepsilon > 0$:

$$\mathbb{P}\Big(\lim_{n\to\infty} |\overline{X}_n - \mu| < \varepsilon\Big) = 1.$$

In the SLLN, we're saying that **the set on which** "the sequence of sample means eventually stabilises within an epsilon of the population mean" has (measure) probability 1

$$\mathbb{P}\Big(\{\omega \in \Omega\colon \lim_{n\to\infty} \overline{X}_n = \mu\}\Big) = 1.$$

20.4 Convergence in Distribution

The third mode of convergence is one we'd already seen when discussing moment-generating functions. This mode of convergence will let us formulate the main theorem of this chapter — the CLT.

Definition 20.4.1 A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all points x where $F_X(x)$ is continuous.

Convergence in distribution is implied by other types of convergence.

Theorem 20.4.2 If $X_n \to X$ in probability, then $X_n \to X$ in distribution.

The following proof follows the guideline of Exercise 5.40 [1, p. 213]:

(a) Given t and ε, show that $\mathbb{P}(\{X \leqslant t - \varepsilon\}) \leqslant \mathbb{P}(\{X_n \leqslant t\}) + \mathbb{P}(\{|X_n - X| \geqslant \varepsilon\})$. This gives a lower bound on $\mathbb{P}(\{X_n \leqslant t\})$.

(b) Use a similar strategy to get an upper bound on $\mathbb{P}(\{X_n \leqslant t\})$.

(c) By pinching, deduce that $\mathbb{P}(\{X_n \leqslant t\}) \to \mathbb{P}(\{X \leqslant t\})$.

Proof.

(a) We want a lower bound on $\mathbb{P}(\{X_n \leqslant t\})$ so we should be looking for some event A that implies $B = \{X_n \leqslant t\}$, for this would tell us that $A \subseteq B$ from which we can deduce (by the monotonicity of $\mathbb{P}$) that $\mathbb{P}(A) \leqslant \mathbb{P}(B)$. Hopefully this will give me what I need. That $X_n(\omega) \leqslant t$ certainly happens if both $X(\omega)$ is at least some ε to the left of t, and that X_n is within ε of X.

Figure 20.1: A visualisation of the shaded region where X_n can reside given the two conditions prescribed.

Let $\omega \in \Omega$ be s.t. $X(\omega) \leqslant t - \varepsilon$ and $|X_n(\omega) - X(\omega)| \leqslant \varepsilon$. It follows that

$$\begin{aligned} |X_n(\omega) - X(\omega)| \leqslant \varepsilon &\iff -\varepsilon \leqslant X_n(\omega) - X(\omega) \leqslant \varepsilon \\ &\implies X_n(\omega) \leqslant X(\omega) + \varepsilon \leqslant (t - \varepsilon) + \varepsilon = t \end{aligned}$$

so it follows that $A := \{X \leqslant t-\varepsilon\} \cap \{|X_n - X| \leqslant \varepsilon\} \subseteq B$. Taking probabilities, we have that

$$\mathbb{P}(\{X \leqslant t-\varepsilon\} \cap \{|X_n - X| \leqslant \varepsilon\}) = \mathbb{P}(C \cap D) = \mathbb{P}(A) \leqslant \mathbb{P}(B) = \mathbb{P}(\{X_n \leqslant t\}).$$

The desired formula suggests[1] that we should write $\mathbb{P}(C \cap D) = \mathbb{P}(C) - \mathbb{P}(C \cap D^{\mathsf{c}})$ because $D^{\mathsf{c}} = \{|X_n - X| \geqslant \varepsilon\}$. Thus,

$$\begin{aligned}\mathbb{P}(C \cap D) &= \mathbb{P}(C) - \mathbb{P}(C \cap D^{\mathsf{c}}) \\ &\geqslant \mathbb{P}(C) - \mathbb{P}(D^{\mathsf{c}}) \quad \text{since } \mathbb{P}(C \cap D^{\mathsf{c}}) \leqslant \mathbb{P}(D^{\mathsf{c}}).\end{aligned}$$

We conclude (a) by combining the two inequalities:

$$\mathbb{P}(\{X \leqslant t-\varepsilon\}) - \mathbb{P}(\{|X_n - X| \geqslant \varepsilon\}) \leqslant \mathbb{P}(\{X_n \leqslant t\}).$$

(b) I would argue that 'similarly' is not a good word to describe what's going on in this case. The spirit of the argument is somewhat opposite to (a). Where for the lower bound, we fixed X and asked where X_n could be, for the upper bound we fix X_n and ask where X could be (in the sense that we're looking for a containment $\{X_n \leqslant t\} \subseteq B$).

In that spirit, we fix $\{X_n \leqslant t\}$ and partition the outcome space based on where X can be:

$$\begin{aligned}\{X_n \leqslant t\} &\subseteq \{X_n \leqslant t\} \cap \Omega \\ &= \{X_n \leqslant t\} \cap (\{|X_n - X| < \varepsilon\} \sqcup \{|X_n - X| \geqslant \varepsilon\}) \\ &= (\{X_n \leqslant t\} \cap \{|X_n - X| < \varepsilon\}) \sqcup (\{X_n \leqslant t\} \cap \{|X_n - X| \geqslant \varepsilon\}) \\ &=: A_1 \sqcup A_2 \\ &\subseteq A_1 \sqcup \{|X_n - X| \geqslant \varepsilon\}\end{aligned}$$

Let $\omega \in A_1$. Then, $X(\omega) \leqslant \varepsilon + t$. Thus, we conclude a similar inequality:

$$\begin{aligned}\mathbb{P}(\{X_n \leqslant t\}) = \mathbb{P}(B) = \mathbb{P}(B \cap \Omega) &\leqslant \mathbb{P}(A_1 \sqcup \{|X_n - X| \geqslant \varepsilon\}) \\ &\leqslant \mathbb{P}(\{X \leqslant t-\varepsilon\} \sqcup \{|X_n - X| \geqslant \varepsilon\}) \\ &\leqslant \mathbb{P}(\{X \leqslant t-\varepsilon\}) + \mathbb{P}(\{|X_n - X| \geqslant \varepsilon\})\end{aligned}$$

Thus, we've proven (b):

$$\mathbb{P}(\{X_n \leqslant t\}) \leqslant \mathbb{P}(\{X \leqslant t-\varepsilon\}) + \mathbb{P}(\{|X_n - X| \geqslant \varepsilon\}).$$

(c) Combining (a) and (b) gives

$$\mathbb{P}(\{X \leqslant t-\varepsilon\}) - \mathbb{P}(\{|X_n - X| \geqslant \varepsilon\}) \leqslant \mathbb{P}(\{X_n \leqslant t\}) \leqslant \mathbb{P}(\{X \leqslant t-\varepsilon\}) + \mathbb{P}(\{|X_n - X| \geqslant \varepsilon\}).$$

Since $X_n \to X$ in probability, the terms $\mathbb{P}(\{|X_n - X\}) \overset{n\to\infty}{\longrightarrow} 0$. Thus, $\mathbb{P}(\{X_n \leqslant t\}) \longrightarrow \mathbb{P}(\{X \leqslant t\})$ as $n \to \infty$ which is what it means for X_n to converge to X in distribution.

■

Theorem 20.4.3 For a consistent estimator, convergence in probability and distribution are the same i.e. The sequence of random variables $X_1, X_2, \ldots$ converges in probability to a constant (random variable) μ iff the sequence converges in distribution to μ.

[1] I'm not very satisfied with my hand being held so much but it's 1am and it is what it is.

20.5 An Approximation for $\overline{X}$ — Classical Central Limit Theorem

The general vibe of the CLT is that it gives us an asymptotic, i.e. for large, fixed n, approximation for the sampling distribution of $\overline{X}$ regardless of the distribution of the population from which the sample is drawn.

This was a bit of a minefield to navigate, primarily because many texts present different versions (in terms of assumptions and form of conclusion) of the Central Limit Theorem. It looks like there are 3 main ones. In descending order of strength of assumptions (so increasing order of generality i.e. weaker assumptions):

- The classical CLT is Lindenberg-Lévy
  - X_i i.i.d. with $\mathbb{E}(X_i) = \mu$, finite variance $\text{Var}(X_i) = \sigma^2$.
- Lyapunov CLT
  - X_i independent, $|X_i|$ have moments of some order $(2+\delta)$ and the growth of the moments is limited by the "Lyapunov condition."
- Lindenberg-Feller CLT
  - Replace the Lyapunov condition with the weaker "Lindenberg condition."

This means that C&B and Wackerly approach a special case of the classical Lindenberg-Lévy CLT by assuming the X_i all have MGFs in some neighbourhood of the origin.

- [7] Theorem 7.5 = [1] Theorem 5.5.14
  - Lindenberg-Lévy with added MGF assumption to offer a nice proof
- [7] Theorem 7.4 = [1] Theorem 5.5.15 ("Strong CLT")
  - Just Lindenberg-Lévy which neither text proves.

If people would just unify under common names, humanity would advance so profusely.

Theorem 20.5.1 (Special case of Lindenberg-Lévy CLT) Let $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} X$ whose MGFs exist in a neighbourhood of the origin (that is, $\exists b > 0$ s.t. $M_X(t)$ exists for $|t| < b$). Let $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 > 0$. (Both μ and σ^2 are finite as a consequence of the MGFs existing)
Let $G_n(x)$ denote the CDF of the random variable

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Then for any $x \in \mathbb{R}$

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{y^2/2}\, \mathrm{d}y.$$

Limitations of the CLT include:

- We have no way of knowing how good the approximation is in general.
  - In fact, the goodness of the approximation is a function of the original distribution itself so we must check this on a case-by-case basis.
- Meta-limitation: Increased and cheaper computation power lessens the importance of finding approximations like this.

A fundamental nuance of the CLT is that for a **fixed** (but very large) n, the shape of the frequency distribution (histogram) will begin to look similar to a normal frequency distribution (the iconic bell shape). This means we can renormalise it to get the approximate distribution of the sample mean. However, the problem is that the variance of the sample means will continue to decrease in size (since $\operatorname{Var}(\overline{X}_n) = \sigma^2/n$) and so the distribution of the normalised sample mean converges to a degenerate distribution — the distribution of the constant random variable μ, as stated in the laws of large numbers.

This is what Wackerly et al. mean when they write:

> As a matter of convenience, the conclusion of the central limit theorem is often replaced with the simpler statement that $\overline{X}$ is asymptotically normally distributed with mean μ and variance σ^2/n.
>
> [7, p. 372]

Proof. We'll show that for $|t| < h$, the MGF of $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ converges to $\exp(t^2/2)$, the MGF of a standard normally distributed random variable. The quantity of interest is the moment-generating function $M_{(\overline{X}_n-\mu)/(\sigma/\sqrt{n})}(t)$ so one needs to figure out for which values of t it exists.

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma} =: \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$$

Since $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} X$, $Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} Y$ where $Y = (X - \mu)/\sigma$. Then we can see that the MGF of $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ is:

$$\begin{aligned}
M_{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}(t) &= \mathbb{E}\left(\exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right)\right) \\
&= \mathbb{E}\left(\prod_{i=1}^{n}\exp\left(\tfrac{t}{\sqrt{n}}Y_i\right)\right) \\
&\overset{7.1.3}{=} \prod_{i=1}^{n}\mathbb{E}\left(\exp\left(\tfrac{t}{\sqrt{n}}Y_i\right)\right) \\
&= \left(M_Y\left(\tfrac{t}{\sqrt{n}}\right)\right)^n \text{ by identical distribution}
\end{aligned}$$

We can go one step further and write the MGF of Y_i in terms of the MGF of X in order to determine where the MGF of $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ is defined. Appealing to **Lemma 11.0.3**,

$$M_{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}(t) = \left(M_Y\left(\tfrac{t}{\sqrt{n}}\right)\right)^n = \left(M_{(X-\mu)/\sigma}\left(\tfrac{t}{\sqrt{n}}\right)\right)^n \overset{11.0.3}{=} \left(\exp\left(\tfrac{-\mu}{\sigma}\tfrac{t}{\sqrt{n}}\right)M_X\left(\tfrac{1}{\sigma}\tfrac{t}{\sqrt{n}}\right)\right)^n$$

Since the MGFs of the X_i exist for $|t| < h$, the MGF of $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ exists for $\left|\frac{1}{\sigma}\frac{t}{\sqrt{n}}\right| < h \iff |t| < h\sigma\sqrt{n}$.
Since the MGF exists in a neighbourhood of the origin, it coincides with its Taylor series for $|t| < h\sigma\sqrt{n}$:

$$M_Y\left(\tfrac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty}\frac{\mathbb{E}(Y^k)\left(\frac{t}{\sqrt{n}}\right)^k}{k!} = \sum_{k=0}^{\infty}\frac{M_Y^{(k)}(0)\left(\frac{t}{\sqrt{n}}\right)^k}{k!} = 0 + 1 + \frac{\left(\frac{t}{\sqrt{n}}\right)^2}{2!} + \mathcal{R}_Y\left(\tfrac{t}{\sqrt{n}}\right)$$

since by construction $Y \sim \mathcal{N}(0, 1)$ and:

- $M_Y^{(0)} = \mathbb{E}\big(Y^0\big) = 1$, trivially,
- $M_Y^{(1)}(0) = \mathbb{E}\big(Y^1\big) = 0$, and
- $M_Y^{(2)}(0) = \mathbb{E}\big(Y^2\big) = \text{Var}(Y) + (\mathbb{E}(Y))^2 = 1 + 0 = 1$.

Thus, we're interested in the following limit:

$$\begin{aligned}\lim_{n\to\infty} M_{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}(t) &= \lim_{n\to\infty}\left(M_Y\big(\tfrac{t}{\sqrt{n}}\big)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \mathcal{R}_Y\big(\tfrac{t}{\sqrt{n}}\big)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{1}{n}\left(\frac{t^2}{n} + n\mathcal{R}_Y\big(\tfrac{t}{\sqrt{n}}\big)\right)\right)^n\end{aligned}$$

It's cheating a bit but the last line comes from the fact that we know the end goal should have an exponential term manifest in the form $\big(1 + \frac{a_n}{n}\big)^n$.

> Let $g \in \mathcal{C}^r(x)$. Taylor's theorem states that the remainder from the approximation $g(x) - T_r(x) = g(x) - \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i$ (where $T_r(x)$ is the Taylor polynomial of order r about a) always tends to 0 faster than the highest-order explicit term i.e.
>
> $$\lim_{x\to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

For fixed $t \neq 0$:

$$\lim_{n\to\infty} \frac{\mathcal{R}_Y\big(\frac{t}{\sqrt{n}}\big)}{\big(\frac{t}{\sqrt{n}}\big)^2} = 0.$$

Since t is fixed, we also have that

$$\lim_{n\to\infty} \frac{\mathcal{R}_Y\big(\frac{t}{\sqrt{n}}\big)}{\big(\frac{1}{\sqrt{n}}\big)^2} = \lim_{n\to\infty} n\mathcal{R}_Y\big(\tfrac{t}{\sqrt{n}}\big) = 0.$$

Since $\mathcal{R}_Y\big(\frac{0}{\sqrt{n}}\big)$ is equal to a sum of terms that have $(t/\sqrt{n})^k$ as a factor, then its limit is also zero for $t = 0$. Thus, for any fixed t we can write:

$$\begin{aligned}\lim_{n\to\infty} M_{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}(t) &= \lim_{n\to\infty}\left(M_Y\big(\tfrac{t}{\sqrt{n}}\big)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \mathcal{R}_Y\big(\tfrac{t}{\sqrt{n}}\big)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{1}{n}\left(\frac{t^2}{n} + n\mathcal{R}_Y\big(\tfrac{t}{\sqrt{n}}\big)\right)\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n \\ &= \exp\left(\lim_{n\to\infty} a_n\right) \text{ where } a_n = \tfrac{t^2}{n} + n\mathcal{R}_Y\big(\tfrac{t}{\sqrt{n}}\big) \overset{n\to\infty}{\longrightarrow} \tfrac{t^2}{2} \\ &= \exp\left(\frac{t^2}{n}\right)\end{aligned}$$

■

20.6 The Normal Approximation to the Binomial Distribution

We can view a binomially distributed random variable $Y \sim \text{Bin}(n, p)$ as the sum of a collection of

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$$

with $\mathbb{E}(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$. Consequently, when n is large, the sample fraction of successes

$$\frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

possesses an approximately normal sampling distribution with mean $\mu = \mathbb{E}(X_i) = p$ and variance $\sigma^2/n = p(1-p)/n$.

A useful rule of thumb is that the normal approximation to the binomial distribution is appropriate when $p \pm 3\sqrt{pq/n} \in (0, 1)$.

- Exercise 7.44 [6] $\implies$ a more convenient, but equivalent, criterion is that the normal approximation is adequate if
$$n > 9\left(\frac{\text{larger of } p, q}{\text{smaller of } p, q}\right)$$
- Exercise 7.45 [6] $\implies$ for some values of p, this criterion is sometimes met for more moderate values of n.

CHAPTER 21

Goodness of Estimators, Point and Interval

Statistics is about making inferences about a population based on information contained in a sample. Populations are characterised by numerical descriptive measures called parameters so the objective of many statistical investigations is to make an inference about (one or more) population parameters.

- A single number calculated from a sample which is used to estimate a population parameter is called a *point* estimate.
- *Interval* estimates can also be computed from a sample to enclose a population parameter.

In either case, we use an estimator[1] to compute an estimate.

Definition 21.0.1 Let $\mathbf{X} = (X_1, \ldots, X_n) \colon (\Omega, \mathcal{F}) \to (E, \mathcal{E}) = (\prod_1^n E_i, \otimes_1^n \mathcal{E}_i)$ be a random sample where the $X_i \colon (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$ are s.t. $X_i \underset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$, and θ is a fixed population parameter to be estimated/enclosed.

- We call $T \colon \mathbf{X}(\Omega) \to S$ a **statistic** if:
  - $(E, \mathcal{E})$ is a Borel space,
  - $\mathcal{S}$ contains all singletons,
  - T is $(\mathcal{E}|_{\mathbf{X}(\Omega)}, \mathcal{S})$-measurable, and
  - T doesn't depend on any unknown parameters (including θ).
- If T is a statistic, and the random element $T \circ \mathbf{X}$ is used to estimate a population parameter, then we call $T \circ \mathbf{X}$ an **estimator**.

21.1 Point Estimation

Given a statistic T, a **point estimator** is a random element $T \circ \mathbf{X}$ that one uses to approximate θ. **Point estimation** is thus the the endeavour to give the best single estimated value of a parameter — to find such a statistic T s.t. $T \circ \mathbf{X}$ is a good approximation of θ.

Criteria for comparing point estimators include biasedness, variance, mean square error, and Wackerly [7] mentions the **error of estimation** $\varepsilon := |\widehat{\theta} - \theta|$.

21.2 Interval Estimation

Interval estimation is the specification of a range of values within which the true parameter θ is asserted to lie, and an interval estimator is (informally) a rule specifying how one can use sample measurements to calculate two numbers that form the endpoints of such an interval. Ideally:

- the interval will contain the target parameter θ,
- and it will be relatively narrow.

[1] A rule or method of estimating a parameter of a population, usually expressed as a function of sample values.

One or both of the endpoints will vary randomly from sample to sample (i.e. at least one endpoint is a function of the random sample). Therefore, the length and location of the interval are random quantities.

The goal is to find an interval estimator capable of generating narrow intervals that have a high probability of enclosing θ.

Interval estimation is the problem of finding two statistics $L, U : \mathbf{X}(\Omega) \to S$ s.t.

$$\mathbb{P}(L \leqslant \theta \leqslant U) \geqslant 1 - \alpha.$$

The random interval $[L, U]$ is the **interval estimator** of θ with confidence level $1 - \alpha$.
The assertion that θ lies in this interval will be true, on average, in a proportion $1 - \alpha$ of the cases where the assertion is made.

There is no general algorithm for interval estimation. But there is a general outline.

Remarks (Terminological)

- **Interval estimators** are also called **confidence intervals**.
- The **upper** and **lower** endpoints of a confidence interval are called the **upper** and **lower confidence limits**, respectively.
- The probability that a confidence interval will enclose θ-fixed is called the **confidence coefficient** (or a **confidence level**), denoted by $1 - \alpha$.
- The term **significance level** is reserved for α itself.

A high confidence coefficient associated with our interval estimator means we can be highly confident that any realisation interval $I \ni \theta$.

> **Definition 21.2.1** More generally, a **confidence set** is a random subset of a parameter space which has a specified probability of containing unknown parameters under repeated sampling. The best-known example is a confidence interval, but a confidence set may comprise disjoint subsets or may lie in a set of dimension greater than 1.

Sometimes we're interested in a **one-sided** confidence interval. This occurs when either L is $-\infty$, or U is $+\infty$ (but not both at the same time).

21.3 Bias and MSE of Point Estimators

- A point estimator $\widehat{\theta}$ of a population parameter θ is called **unbiased** if $\mathbb{E}\left(\widehat{\theta}\right) = \theta$.
  - Otherwise, $\widehat{\theta}$ is said to be **biased**.
- The **bias** of a point estimator is given by $\mathbb{E}\left(\widehat{\theta}\right) - \theta$.

If an estimator is unbiased, we'd like for its variance to be as small as possible so that in repeated sampling, a higher fraction of estimates will be close to θ.

Another measure of goodness is the mean square error of a point estimator $\widehat{\theta}$ defined by:

$$\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\left((\widehat{\theta} - \theta)^2\right).$$

Lemma 21.3.1 The mean square error of $\widehat{\theta}$ is a function of both the variance and bias of $\widehat{\theta}$.

Proof.

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}\big((\widehat{\theta}-\theta)^2\big) \\
&= \mathbb{E}\big(\widehat{\theta}^2 - 2\widehat{\theta}\theta + \theta^2\big) \\
&= \mathbb{E}\big(\widehat{\theta}^2\big) - 2\mathbb{E}(\widehat{\theta})\theta + \theta^2 \\
&= \underbrace{\mathbb{E}\big(\widehat{\theta}^2\big) - \mathbb{E}(\widehat{\theta})^2}_{=\mathrm{Var}(\widehat{\theta})} + \underbrace{\mathbb{E}(\widehat{\theta})^2 - 2\mathbb{E}(\widehat{\theta})\theta + \theta^2}_{=(\mathrm{Bias}(\widehat{\theta}))^2}
\end{aligned}$$

■

Example Examples of unbiased point estimators:

- The sample mean $\overline{X}$
- The **sample proportion** $\hat{p} = Y/n$ (to estimate a binomial parameter p in $Y \sim \mathrm{Bin}(n,p)$)

We can construct more examples of unbiased estimators that follow our intuition e.g. Suppose that we have two independent random samples of n_1 and n_2 observations selected from two different populations. We can estimate the difference between means $\mu_1 - \mu_2$ by considering the difference in sample means.

To facilitate communication, we use the notation $\sigma^2_{\widehat{\theta}}$ to denote the variance of the sampling distribution of the estimator $\widehat{\theta}$. The standard deviation of the sampling distribution of the estimator $\widehat{\theta}$, $\sigma_{\hat{\theta}} = \sqrt{(\sigma_{\hat{\theta}})^2}$, is usually called the **standard error** of the estimator.

Biased estimators also exist.

Example Let $Y_1, \ldots, Y_n$ be a random sample with $\mathbb{E}(Y_i) = \mu$ and $\mathrm{Var}(Y_i) = \sigma^2$. Show that

$$S'^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

is a biased estimator for σ^2.

$$\begin{aligned}
\mathbb{E}\big(S'^2\big) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right) \\
&= \mathbb{E}\left(\frac{1}{n}\left(\left(\sum_{i=1}^{n}Y_i^2\right) - n\overline{Y})^2\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big(Y_i^2\big) - \mathbb{E}\Big(\overline{Y}^2\Big) \\
&= \mathbb{E}\big(Y_1^2\big) - \mathbb{E}\Big(\overline{Y}^2\Big) \\
&= \mathrm{Var}(Y_1) + (\mathbb{E}(Y_1))^2 - (\mathrm{Var}\big(\overline{Y}\big) + \big(\mathbb{E}\big(\overline{Y}\big)\big)^2) \\
&= \sigma^2 + \mu^2 - \sigma^2_{\overline{Y}} - (\mu_{\overline{Y}})^2 \\
&= \sigma^2\left(1 - \frac{1}{n}\right) \\
&\neq \sigma^2.
\end{aligned}$$

21.4 Evaluating the Goodness of a Point Estimator

The **error of estimation** ε is the distance between an estimator and its target parameter

$$\varepsilon := |\widehat{\theta} - \theta|.$$

As mentioned earlier, we aren't able to predict the error of estimation ε for a particular estimate since ε is a random variable, but we can make probability statements about it.

Example Suppose that $\widehat{\theta}$ is an unbiased estimator of θ, and has a known sampling distribution with density $f_{\widehat{\theta}}$. Let $b > 0$.

The probability that the error of estimation is less than b is $\mathbb{P}(|\widehat{\theta} - \theta| \leqslant b)$. Graphically, it would be represented by the area under the density curve over $(\theta - b, \theta + b)$. We can think of b as a probabilistic bound on the error of estimation. If b is small, then $\mathbb{P}(|\widehat{\theta} - \theta| < \text{small})$ can be regarded as a measure of goodness of a single estimate. This probability identifies the fraction of times, in repeated sampling, that the estimator $\widehat{\theta}$ falls within b units of θ (the target parameter).

Suppose that we wish to know the $b > 0$ s.t. $\mathbb{P}(\varepsilon < b) = 0.9$.

- If we know the sampling distribution of $\widehat{\theta}$, we can search for a value of b s.t.

$$\int_{\theta-b}^{\theta+b} f(\widehat{\theta})\,\mathrm{d}\widehat{\theta} = 0.90.$$

- Irrespective of knowing the sampling distribution of $\widehat{\theta}$, if we know that $\widehat{\theta}$ is an unbiased point estimator of θ, then we can find an approximate bound on ε by expressing b as a multiple of the **standard error** of $\widehat{\theta}$.

Example Chebyshev's theorem tells us that if we let $b = k\sigma_{\widehat{\theta}}$, then

$$\mathbb{P}(\varepsilon < b) = \mathbb{P}\Big(|\widehat{\theta} - \theta| < k\sigma_{\widehat{\theta}}\Big) \geqslant 1 - \frac{1}{k^2}.$$

A convenient value to take is $k = 2$. Hence, $\mathbb{P}\big(\varepsilon < 2\sigma_{\widehat{\theta}}\big) \geqslant 0.75$. Most random variables observed in nature lie within 2 standard deviations of their mean with probability ≈ 0.95. Chebyshev's theorem is often very conservative when it comes to the bounds for probabilities. The actual probability often exceeds the Chebyshev bounds by a lot.

21.5 Pivotal Method for Interval Estimation

A method for finding confidence intervals. The method depends on finding a pivotal quantity that possesses 2 characteristics:

- It's a function of the sample measurements and the unknown parameter θ, where θ is the **only** unknown quantity.
- Its probability distribution doesn't depend on θ.

Definition 21.5.1 A **pivotal quantity** is a function of sample values and one or more parameters with a distribution that does not depend on the parameters. The concept is mainly used in the construction of confidence intervals. [4]

Example 21.5.2 If $X \sim \mathcal{N}(\mu, s^2)$ where s^2 is an estimated variance, then $(\overline{X} - \mu)/(S/\sqrt{n})$ follows a t-distribution and may be used to define confidence limits for estimating μ.

If the probability distribution of the pivotal quantity is known, then the following logic can be used to form the desired interval estimate e.g.

1. Y-r.v., $c > 0$ constant, and $\mathbb{P}(a \leqslant Y \leqslant b) = 0.7$
2. Then certainly $\mathbb{P}(ca \leqslant cY \leqslant cb) = 0.7$
3. Similarly, $\mathbb{P}(ca + d \leqslant cY + d \leqslant cb + d) = 0.7$
4. i.e. the probability is unaffected by an affine transformation of Y

and so we can use operations like these to form the desired interval estimator.

Let $Y \sim \text{Exp}(\theta)$. We want to find a pivotal quantity $U = g(Y, \theta)$ s.t. f_U doesn't depend on θ.

Example 21.5.3 $U = g(Y, \theta) = \dfrac{1}{\theta}Y$ has distribution

$$f_U(u) = \begin{cases} e^{-u}, & u > 0 \\ 0, & \text{otherwise} \end{cases}$$

We've found a pivotal quantity. We want an interval estimator with confidence coefficient 0.90 so we're looking for numbers a and b s.t.

$$\mathbb{P}(a \leqslant U \leqslant b) = 0.90.$$

One way to do this is to choose a and b s.t.

$$\mathbb{P}(U < a) = 0.05 \quad \text{and} \quad \mathbb{P}(U > b) = 0.05$$

It follows that

$$0.90 = \mathbb{P}(0.051 \leqslant U \leqslant 2.996) = \mathbb{P}\left(\cdots \leqslant \frac{Y}{\theta} \leqslant \ldots\right).$$

Since $Y \sim \text{Exp}(\theta)$, $Y > 0$ so we can divide through by Y to get

$$= \mathbb{P}\left(\frac{\cdots}{Y} \leqslant \frac{1}{\theta} \leqslant \frac{\ldots}{Y}\right)$$

and taking reciprocals (which is permitted because $Y > 0$) gives

$$\begin{aligned} &= \mathbb{P}\left(\frac{Y}{\cdots} \geqslant \theta \geqslant \frac{Y}{\ldots}\right) \\ &= \mathbb{P}(\widehat{\theta_U} \geqslant \theta \geqslant \widehat{\theta_L}) \end{aligned}$$

Example 21.5.4 Suppose that we take a single observation from a uniformly distributed population $\mathcal{U}[0, \theta]$ where θ-unknown. Find a 95% lower confidence bound for θ.

Solution: Let $X \sim \mathcal{U}([0, \theta])$ represent the act of sampling one element from a $\mathcal{U}([0, \theta])$-distributed population. We're tasked with finding a 95% lower confidence bound for θ i.e. we wish to find a statistic L s.t. $\mathbb{P}(L \leqslant \theta) \geqslant 0.95$. We'll do so by first finding a pivotal quantity.

The distribution of X is given by

$$\begin{aligned} F_X(x) &= \mathbb{P}(\{X \leqslant x\}) \\ &= \mathbb{E}\big(\mathbb{1}_{X^{-1}((-\infty, x])}\big) \\ &= \mathbb{E}(h(X)) \quad \text{where } h = \mathbb{1}_{(-\infty, x]} \\ &= \int_\Omega h(X)\,\mathrm{d}\mathbb{P} = \int_{\mathbb{R}} h\,\mathrm{d}\mathbb{P}_X = \int_{\mathbb{R}} h\phi_X\,\mathrm{d}\lambda \\ &= \int_{\mathbb{R}} \mathbb{1}_{(-\infty, x]}(y)\phi_X(y)\,\mathrm{d}\lambda(y) \\ &= \int_{\mathbb{R}} \mathbb{1}_{(-\infty, x]}(y)\,\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(y)\,\mathrm{d}y \\ &= \int_0^x \frac{1}{\theta}\,\mathrm{d}y = \frac{x}{\theta} \end{aligned}$$

It may be worth dividing X through by θ to standardise the stretching of the density as θ increases. Try $U = X/\theta$. Then

$$\begin{aligned} F_U(u) &= \mathbb{P}(\{U \leqslant u\}) \\ &= \mathbb{P}(\{X/\theta \leqslant u\}) \\ &= \mathbb{P}(\{X \leqslant u\theta\}) \\ &= F_X(u\theta) = u \end{aligned}$$

i.e. the distribution of U is uniform on $[0,1]$ and independent of θ. Thus, U is a pivotal quantity. Now I want to use the distribution of U, and re-arrange it to get a lower confidence interval about θ. For any $c \in [0,1]$:

$$\mathbb{P}\left(\left\{\frac{X}{\theta} \leqslant c\right\}\right) = \mathbb{P}(\{X \leqslant c\theta\}) = \mathbb{P}\left(\left\{\frac{X}{c} \leqslant \theta\right\}\right)$$

Since U is uniform, this term is equal to c.

$$\therefore c = \mathbb{P}\left(\left\{\frac{X}{c} \leqslant \theta\right\}\right)$$

Now we can let $c = 0.95$ to get the 95% confidence interval.

21.6 Selecting the Sample Size

- The design of an experiment is essentially a plan for purchasing a quantity of information.
  - We should seek to minimise the cost of obtaining said information.
- The sampling procedure (or experimental design) affects the quantity of information per measurement.
  + This, together with the sample size n, controls the total amount of relevant information in a sample.

Attention will now be focused on the sample size n, with the discussion branching off into small and large samples:

This section is on a potential discussion between an experimenter and statistician, highlighting the process of selecting a sample size:
Experimenter: How many measurements should be included in the sample to calculate my estimate?

Statistician: Depends on how much "information" you want out of the sample e.g. accuracy i.e. You should specify a bound on the error of estimation.

Experimenter: I want to estimate the true average daily yield μ of a chemical, and for the error of estimation to be less than 5 tons with probability 0.95.

Statistician: Here the statistician assumes the sample is large, and so the distribution of the sample mean is asymptotically normal. Approximately 95% of the sample means will lie within $2\sigma_{\overline{Y}}$ of μ in repeated sampling, so you're asking that $2\sigma_{\overline{Y}} = 5$ i.e. $\frac{2\sigma}{\sqrt{n}} = 5$ i.e. $n = \frac{4\sigma^2}{25}$. Unless σ is known, we can estimate σ with $0.25 * \text{range}$ (since damn near every observation is within 4σ of the mean). So what's your range?

Experimenter: 84 tons!

Statistician: Therefore, $\sigma \approx 84/4 = 21$ so $n = \frac{4\sigma^2}{25} \approx 71$. So with a sample size of $n = 71$, we can be reasonably certain (with confidence coefficient $\approx .95$) that our estimate will lie within 5 tons of μ, the true average daily yield.

Note that we can't obtain an exact numerical value for $n = \frac{4\sigma^2}{25}$ because the variability of the estimator $\overline{Y}$ depends on the variability of the population σ from which the sample was drawn.

21.7 Large-Sample Confidence Intervals

It turns out that some common unbiased point estimators we've seen can all be approximated (for a large sample size) by a normal sampling distribution with standard errors as follows:

target parameter θ	sample size(s)	point estimator $\widehat{\theta}$	$\mathbb{E}(\widehat{\theta})$	standard error $\sigma_{\widehat{\theta}}$
μ	n	$\overline{X}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\widehat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\overline{X_1} - \overline{X_2}$	$\mu_1 - \mu_2$	$\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}$
$p_1 - p_2$	n_1 and n_2	$p_1 - p_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

Table 21.1: The second row is for $Y \sim \text{Bin}(n, p)$, and the bottom two rows are for independent samples.

For large samples,

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}}$$

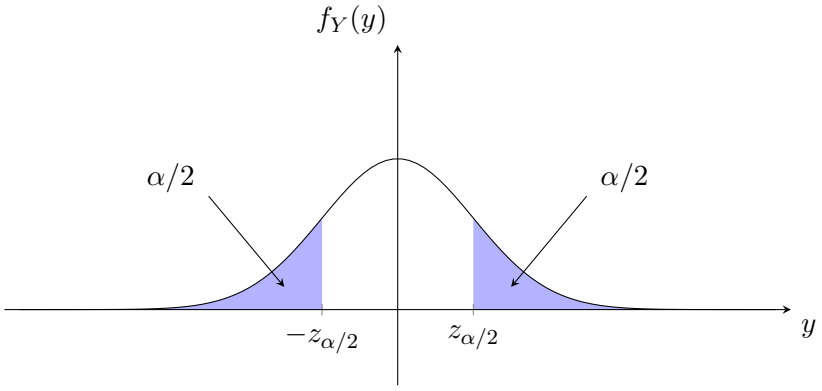
possesses approximately a standard normal distribution. Therefore, Z forms (at least approximately/asymptotically) a pivotal quantity, and so we can use the pivotal method to develop confidence intervals for the target parameter θ.

Example 21.7.1 Let $\widehat{\theta}$ be a statistic that's normally distributed with mean θ and standard error $\sigma_{\widehat{\theta}}$. Find a confidence interval for θ that possesses a confidence coefficient equal to $(1 - \alpha)$.

Solution: The quantity

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{N}(0, 1).$$

The distribution of this random variable doesn't depend on θ and $Z = g(\widehat{\theta}, \theta)$ so we can use it as a pivotal quantity. We can select values $z_{\alpha/2}$ and $-z_{\alpha/2}$



in the support of this distribution corresponding to the tails such that:

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\big(-z_{\alpha/2} \leqslant Z \leqslant z_{\alpha/2}\big) \\
&= \mathbb{P}\Big(-z_{\alpha/2} \leqslant \tfrac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \leqslant z_{\alpha/2}\Big) \\
&= \mathbb{P}\big(-z_{\alpha/2}\sigma_{\widehat{\theta}} \leqslant \widehat{\theta} - \theta \leqslant z_{\alpha/2}\sigma_{\widehat{\theta}}\big) \\
&= \mathbb{P}\big(-z_{\alpha/2}\sigma_{\widehat{\theta}} - \widehat{\theta} \leqslant -\theta \leqslant z_{\alpha/2}\sigma_{\widehat{\theta}} - \widehat{\theta}\big) \\
&= \mathbb{P}\big(z_{\alpha/2}\sigma_{\widehat{\theta}} + \widehat{\theta} \geqslant \theta \geqslant \widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}}\big) \\
&= \mathbb{P}(\widehat{\theta_U} \geqslant \theta \geqslant \widehat{\theta_L})
\end{aligned}$$

are the confidence limits for an approximately $100(1-\alpha)\%$ confidence interval (i.e. an interval estimator with confidence coefficient $1-\alpha$).

When computing an interval based on a realised sample $x_1, \ldots, x_n$ of observations, the interval either contains the true parameter or not. You're 95% confident that the interval contains the parameter because the procedure that generates intervals that contain the parameter 95% of the times the procedure is used.

21.8 Small-Sample Confidence Intervals

This section will operate **under** the assumptions of:

- a sample randomly selected from a normal population (for discussion concerning μ)
- two independent normally distributed random samples

However, if the interval estimator is to be of any value, it must work reasonably well[2] even if the population(s) from which the sample(s) is(/are) draw is/are not normal.

21.8.1 μ

Suppose that $X_1, \ldots, X_n$ is a random sample from a normally distributed population with mean μ and variance σ^2. Let $\overline{X}$ and S^2 denote the sample mean and sample variance, respectively. We'd like to construct a confidence interval for the population mean when σ^2 is unknown, and the sample size is too small to permit us to apply the large-sample techniques from earlier.

We know that $(\overline{X}-\mu)/(S/\sqrt{n})$ has a t_{n-1} distribution. Call that quantity T. This will be the pivotal quantity we use to construct a confidence interval for μ. We can find values $t_{\alpha/2}$ and $-t_{\alpha/2}$ s.t.

$$\mathbb{P}\big(-t_{\alpha/2} \leqslant T \leqslant t_{\alpha/2}\big) = 1-\alpha.$$

Similar to before, we can manipulate this probability to make μ the subject:

$$\begin{aligned}
&= \mathbb{P}\Big(-t_{\alpha/2} \leqslant \tfrac{\overline{X}-\mu}{(S/\sqrt{n})} \leqslant t_{\alpha/2}\Big) \\
&= \mathbb{P}\left(\frac{-t_{\alpha/2}\,S}{\sqrt{n}} - \overline{X} \leqslant -\mu \leqslant \frac{t_{\alpha/2}\,S}{\sqrt{n}} - \overline{X}\right) \\
&= \mathbb{P}\left(\frac{t_{\alpha/2}\,S}{\sqrt{n}} + \overline{X} \geqslant \mu \geqslant \overline{X} - \frac{t_{\alpha/2}\,S}{\sqrt{n}}\right)
\end{aligned}$$

So the resulting confidence interval for μ with confidence coefficient $1-\alpha$ is

$$\left[\overline{X} - \frac{t_{\alpha/2}\,S}{\sqrt{n}}, \overline{X} + \frac{t_{\alpha/2}\,S}{\sqrt{n}}\right].$$

21.8.2 $\mu_1 - \mu_2$

Suppose that we're interested in comparing the means of two normal populations, one with mean μ_1 and variance $(\sigma_1)^2$, and the other with mean μ_2 and variance $(\sigma_2)^2$. If the samples are independent, confidence intervals for $\mu_1 - \mu_2$ based on a t-distributed random variable can be constructed if we assume both populations have a common but unknown variance $\sigma^2 = (\sigma_1)^2 = (\sigma_2)^2$.

Let $\overline{X_1}$ and $\overline{X_2}$ be their respective sample means. The large-sample confidence interval can be developed by using

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_1)^2}{n_1}}}$$

[2]Working reasonably well means that the confidence coefficient should not be affected by modest departures from normality. For most mound-shaped population distributions, experimental studies indicate these confidence intervals maintain confidence coefficients close to the nominal values used in their calculations.

as a pivotal quantity. Since we assume the normal random samples have a shared variance

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_1}}}$$

and σ is unknown so we need to find an estimator of the common variance σ^2 so we can construct a quantity with a t-distribution.

The usual unbiased estimator of σ^2 is obtained by pooling the sample data to obtain the pooled estimator S_p^2:

$$\begin{aligned} S_p^2 &:= \frac{\sum_{i=1}^{n_1}(X_{1,i} - \overline{X_1})^2 + \sum_{i=1}^{n_2}(X_{2,i} - \overline{X_1})^2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \end{aligned}$$

This final expression is a weighted average of S_1^2 and S_2^2, giving larger weight to the sample variance associated with the larger sample size. Further,

$$W = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

is the sum of two independent χ^2-distributed random variables with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom, respectively.

$$\therefore W \sim \chi^2_{(n_1-1)+(n_2-1)}.$$

Since Z and W are independent, we can form the pivotal quantity

$$T = \frac{Z}{\sqrt{\frac{W}{\nu}}} = \ldots = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which by construction is a quantity with a t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. Proceeding as before, the confidence interval for $(\mu_1 - \mu_2)$ has the form

$$(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $t_{\alpha/2}$ is determined from the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

21.9 Confidence Intervals for σ^2

Throughout our construction of confidence intervals for μ, we used S^2 to estimate σ^2 when σ^2 was unknown. In addition to needing information about σ^2 to calculate confidence interval for μ or $\mu_1 - \mu_2$, we may be interested in forming a confidence interval for σ^2 itself:

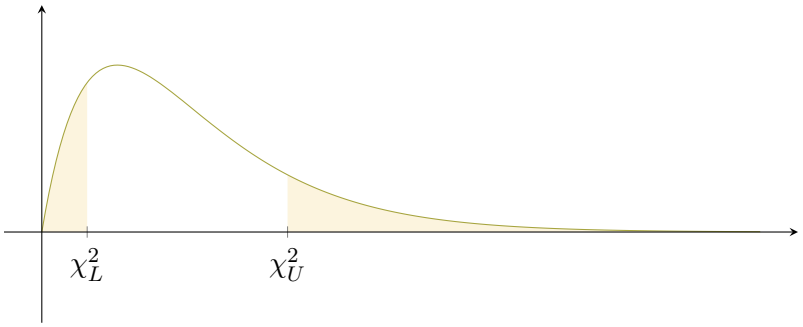
We need a pivotal quantity once more. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 both unknown. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

so we can proceed to find numbers χ_L^2 and χ_U^2 s.t.

$$\mathbb{P}\Big(\chi_L^2 \leqslant \tfrac{(n-1)S^2}{\sigma^2} \leqslant \chi_U^2\Big) = 1 - \alpha$$

for any confidence coefficient $(1 - \alpha)$. The χ^2_{n-1} density isn't symmetric so we have some freedom in choosing the lower and upper confidence limits. We'd like to find the shortest interval that includes σ^2 with probability $(1 - \alpha)$. This is complicated in general and requires a trial-and-error search for the appropriate values. We compromise by choosing points that cut off equal tail areas, as indicated below:



Re-arranging the probability statement gives

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2)}}\right) = 1-\alpha$$

and so a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-(\alpha/2)}}\right).$$

Note that the confidence intervals for σ^2 in this section can differ markedly from the nominal level if the sampled population isn't normally distributed. This observation comes from the examples in [7].

21.10 Summary

Calling back to the earlier comment on parameters having mound-shaped distributions and their confidence intervals maintaining confidence coefficients close to "the nominal values used in their calculations," I can now use the example of estimating μ (in the small-sample case) to explain:

- We assume normality of the random sample $X_1, \ldots, X_n$ so $\overline{X}$ is also normally distributed. ($\widehat{\theta} = \overline{X}$ in this case)
- Then we construct a pivotal quantity $T = (\overline{X} - \mu)/(S/\sqrt{n})$ where we approximate σ^2 with S^2.
$$T \sim t_{n-1}$$
- This leads to the nominal probability statement that for a confidence level $(1-\alpha)$, there exists some real number $t_{\alpha/2,n-1}$ such that
$$\mathbb{P}\big(-t_{\alpha/2,n-1} \leqslant T \leqslant t_{\alpha/2,n-1}\big) = 1-\alpha$$
where $(1-\alpha)$ is the nominal confidence interval.
- We re-write the statement to get a confidence interval for μ:
$$\mathbb{P}\Big(\overline{X} - t_{\alpha/2,n-1}\tfrac{S}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} + t_{\alpha/2,n-1}\tfrac{S}{\sqrt{n}}\Big) = 1-\alpha$$

However, we must keep in mind that our original formulation of the problem did not assume any normality:

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, $\mathbb{E}(X_i) = \mu$-unknown, $\text{Var}(X_i) = \sigma^2$-unknown
- $\mathbb{P}$ is a mound-shaped distribution.

In this case, for a mound-shaped distribution, the distribution of the pivotal quantity T isn't necessarily t_{n-1}. For a confidence level of $1-\alpha$, there exist quantities $Q_{\dots}$ that are the true tail probabilities of the actual distribution of T:

$$\mathbb{P}\big(T \leqslant Q_{\alpha/2,n}\big) = \tfrac{\alpha}{2}$$

$$\mathbb{P}\big(T \leqslant Q_{1-(\alpha/2),n}\big) = 1 - \tfrac{\alpha}{2}$$

In practice, we typically don't know the true distribution of T so we construct the confidence interval for the idealised normal case where $T \sim t_{n-1}$ and use the (potentially) wrong quantities $t_{\alpha/2}$ and $-t_{\alpha/2}$ which satisfy for some $\gamma \in [0,1]$, the equation:

$$\mathbb{P}\left(\overline{X} - t_{\alpha/2,n-1}\tfrac{S}{\sqrt{n}} \leqslant \mu \leqslant \overline{X} + t_{\alpha/2,n-1}\tfrac{S}{\sqrt{n}}\right) = 1-\gamma.$$

One can see the disparity below if the mound-shape is sufficiently skewed.

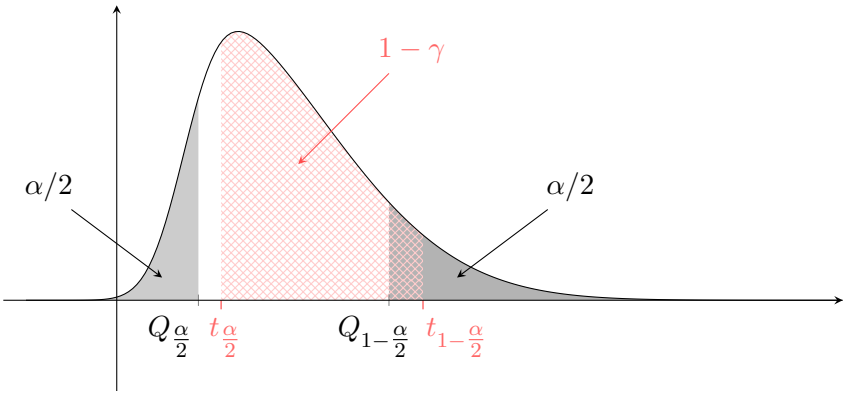


Figure 21.1: The true density of T which is mound-shaped. The true tail quantiles $Q_{\alpha/2}$ and $Q_{1-(\alpha/2)}$ form a confidence interval for σ^2 with confidence coefficient $1-\alpha$.

21.11 Properties of Point Estimators

21.11.1 Relative Efficiency

It's possible to obtain more than one unbiased estimator for the same target parameter θ. We often prefer the estimator with the smaller variance. That is, if both estimators $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are unbiased, $\widehat{\theta_1}$ is relatively more efficient than $\widehat{\theta_2}$ if

$$\operatorname{Var}(\widehat{\theta_1}) < \operatorname{Var}(\widehat{\theta_2}).$$

In fact, we use the ratio of their variance to define the relative efficiency of $\widehat{\theta_1}$ and $\widehat{\theta_2}$ i.e. the **efficiency of $\widehat{\theta_1}$ relative to $\widehat{\theta_2}$**, denoted $\operatorname{eff}(\widehat{\theta_1}, \widehat{\theta_2})$ is defined by

$$\operatorname{eff}(\widehat{\theta_1}, \widehat{\theta_2}) := \frac{\operatorname{Var}(\widehat{\theta_1})}{\operatorname{Var}\left(\widehat{\theta_2}\right)}.$$

21.11.2 Consistency

e.g. Suppose that a coin, which has a probability p of resulting in a head, is tossed n times. If the tosses are independent, then the number of heads $Y \sim \operatorname{Bin}(n,p)$. If p is unknown, $Y/n =: \hat{p}$ is an estimator for p. As $n \to \infty$, Y/n should get closed to the true value of p i.e. for large n:

$$\mathbb{P}\left(\left|\frac{Y}{n} - p\right| \leqslant \varepsilon\right) \quad \text{should be close to 1.}$$

Definition 21.11.1 The estimator $\widehat{\theta_n}$ is said to be a **consistent estimator** of θ if $\forall \varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\theta_n} - \theta| \leqslant \varepsilon) = 1.$$

Remarks

- The above definition really just says that $\widehat{\theta_n}$ is a consistent estimator of θ if, as the sample size n goes to ∞, $\widehat{\theta_n}$ converges in probability to its estimand θ.
- A stronger notion is that of $\widehat{\theta_n}$ being a **strongly consistent** estimator — called as such if it converges $\mathbb{P}$-almost surely, as the sample size increases, to its estimand.

Theorem 21.11.2 (Theorem 9.1 [1]) An unbiased estimator $\widehat{\theta_n}$ for θ is a consistent estimator for θ if

$$\lim_{n\to\infty} \operatorname{Var}(\widehat{\theta_n}) = 0.$$

Proof. Let Y be a random variable with $\mathbb{E}(Y) = \mu$ and $\operatorname{Var}(Y) = \sigma^2 < \infty$. Let $k > 0$. By Chebyshev's theorem:

$$\mathbb{P}\big(|Y - \mu| \geqslant k\sigma^2\big) \leqslant \frac{1}{k^2}.$$

For the random variable $\widehat{\theta_n}$:

$$\mathbb{P}(|\widehat{\theta_n} - \underbrace{\mathbb{E}(\widehat{\theta_n})}_{=\theta}| \geqslant k\sigma^2_{\hat{\theta_n}}) \leqslant \frac{1}{k^2}$$

where we note that the estimator is unbiased. Now let n be a fixed sample size. $k > 0$ can be written in the form

$$k = \frac{\varepsilon}{\sigma_{\hat{\theta_n}}} > 0.$$

Thus, for any fixed n:

$$0 \leqslant \mathbb{P}(|\widehat{\theta_n} - \theta| \geqslant \varepsilon) \leqslant \frac{\sigma^2_{\hat{\theta_n}}}{\varepsilon^2} = \frac{\operatorname{Var}(\widehat{\theta_n})}{\varepsilon^2}.$$

If $\lim_{n\to\infty} \operatorname{Var}(\widehat{\theta_n}) = 0$, then the above expression goes to 0 i.e. $\widehat{\theta_n}$ is a consistent estimator of θ. ■

Lemma 21.11.3 (The Algebra of Convergence in Probability) Suppose that $\widehat{\theta_n}$ converges in probability to θ, and $\widehat{\varphi_n}$ converges in probability to φ. Then:

1. $\widehat{\theta_n} + \widehat{\varphi_n}$ converges in probability to $\theta + \varphi$
2. $\widehat{\theta_n}\widehat{\varphi_n}$ converges in probability to $\theta\varphi$
3. Given that $\varphi \neq 0$, $\frac{\hat{\theta_n}}{\hat{\varphi_n}}$ converges in probability to θ/φ
4. If g is real-valued and continuous, then $g(\widehat{\theta_n})$ converges in probability to $g(\theta)$.

The rigorous justification for the large-sample confidence interval discussion for the mean of any random sample $X_1, \ldots, X_n$ involved the creation of a confidence interval $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ with confidence coefficient approximately equal to $1 - \alpha$. If σ^2 is known, this interval can and should be calculated. If not, we can estimate σ^2 with S^2 without significant loss in accuracy. The following theorem provides the theoretical justification:

Theorem 21.11.4 (Theorem 9.3 from [1]) Suppose that U_n has a distribution function that converges to a standard normal distribution as $n \to \infty$. If W_n converges in probability to 1, then the distribution function of U_n/W_n converges to a standard normal distribution function.

Example 21.11.5 Suppose that $X_1, \ldots, X_n$ is a random sample of size n from a distribution with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Define

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Show that the distribution of $(\overline{X} - \mu)/(S_n/\sqrt{n})$ converges to a standard normal distribution.

Solution:

CHAPTER 22

Data Reduction

Let $\mathbf{X}\colon (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ be a random sample, and T be a statistic of $\mathbf{X}$. Thus, $T\colon (\mathbf{X}(\Omega), \mathcal{E}|_{\mathbf{X}(\Omega)}) \to (S, \mathcal{S})$ is $\mathcal{E}|_{\mathbf{X}(\Omega)}$-$\mathcal{S}$-measurable by **Definition 16.2.1**. It follows that the composition $T \circ \mathbf{X}$ is $(\sigma(\mathbf{X}), \mathcal{S})$-measurable. By the $\sigma(\mathbf{X})$-measurability of $T \circ \mathbf{X}$, it's true that the following inclusion is one of sigma-algebras:

$$\sigma(T \circ \mathbf{X}) \subseteq \sigma(\mathbf{X}).$$

Since σ-algebras have a natural interpretation as amounts of information (in this case, required to fully determine a random variable), the inclusion above is what is meant by a statistic T providing a form of data-reduction (under the guise of summarising samples in a functional sense i.e. two realisations $\boldsymbol{x}$ and $\boldsymbol{y}$ of $\mathbf{X}$ are indistinguishable from the perspective of T if $T(\boldsymbol{x}) = T(\boldsymbol{y})$).

There are three important principles of data reduction; partitioned into two sub-flavours:

- methods of data reduction that don't discard important information about the unknown parameter θ
- methods that successfully discard information that's irrelevant as far as gaining knowledge about θ is concerned

The Sufficiency Principle promotes a method of data reduction that doesn't discard information about θ while achieving some summary of the data.

22.1 Sufficiency

Informally, a sufficient statistic is a function that, when composed with the data, provides as much information about a parameter of interest as the entire dataset does. An undergraduate definition commonly found in undergraduate books/dictionaries is as follows:

Definition 22.1.1 A property of an estimator defined by Fisher (1992). A statistic T is said to be sufficient for a parameter θ if the distribution of a sample $X_1, \ldots, X_n$ given $T = t$ does not depend on θ. [4]

At first glance, I couldn't claim to understand what was being said in this definition. The following general definition in Billingsley [17, p. 450] helped me connect the dots:

> **Notation**: Denote by $\mathbb{P}_\theta[A \,|\, \mathcal{G}]$ and $\mathbb{E}_\theta[X \,|\, \mathcal{G}]$ the conditional probabilities and expected values computed with respect to the probability measure $\mathbb{P}_\theta$ on $(\Omega, \mathcal{F})$.

Suppose that for each $\theta \in \Theta$, where Θ is an indexing set, $\mathbb{P}_\theta$ is a probability measure on a measurable space $(\Omega, \mathcal{F})$. In statistics, the problem is to draw inferences about the unknown parameter θ from an observation.

Definition 22.1.2 A sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ is called **sufficient for the family** $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ if versions of $\mathbb{P}_\theta[A \,|\, \mathcal{G}]$ can be chosen that are independent of θ — that is, if there exists a function $p\colon \mathcal{F} \times \Omega \to [0, 1]$ s.t. $\forall \theta \in \Theta$:

$$\forall A \in \mathcal{F},\ p(A, \cdot) \text{ is a version of } \mathbb{P}_\theta[A \,|\, \mathcal{G}].$$

In the above definition, there's no requirement that $p(\cdot, \omega)$ is a measure for fixed ω. The key takeaway is the order of quantifiers — there exists a **single kernel** p that works for all θ.

> The general idea in Billingsley's definition is that although there may be information in $\mathcal{F}$ not already contained in $\mathcal{G}$, this information is irrelevant to the drawing of inferences about θ.

This got me thinking about how the two definitions are connected. Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample of $X_i \colon (\Omega, \mathcal{F}) \to (E_i, \mathcal{E}_i)$ s.t. $X_i \underset{\text{i.i.d.}}{\sim} \mathbb{P}_\theta$ i.e. $(X_i)_\sharp \mathbb{P} = \mathbb{P}_\theta$. Thus, $\mathbb{P}_{\mathbf{X}} = \otimes_n \mathbb{P}$.

A statistic T is called **sufficient** for θ iff $\sigma(T)$ is a sufficient σ-algebra for the family $\{\otimes_n \mathbb{P}_\theta\}_{\theta \in \Theta}$ defined on $(\mathbf{X}(\Omega), \mathcal{E}|_{\mathbf{X}(\Omega)})$.

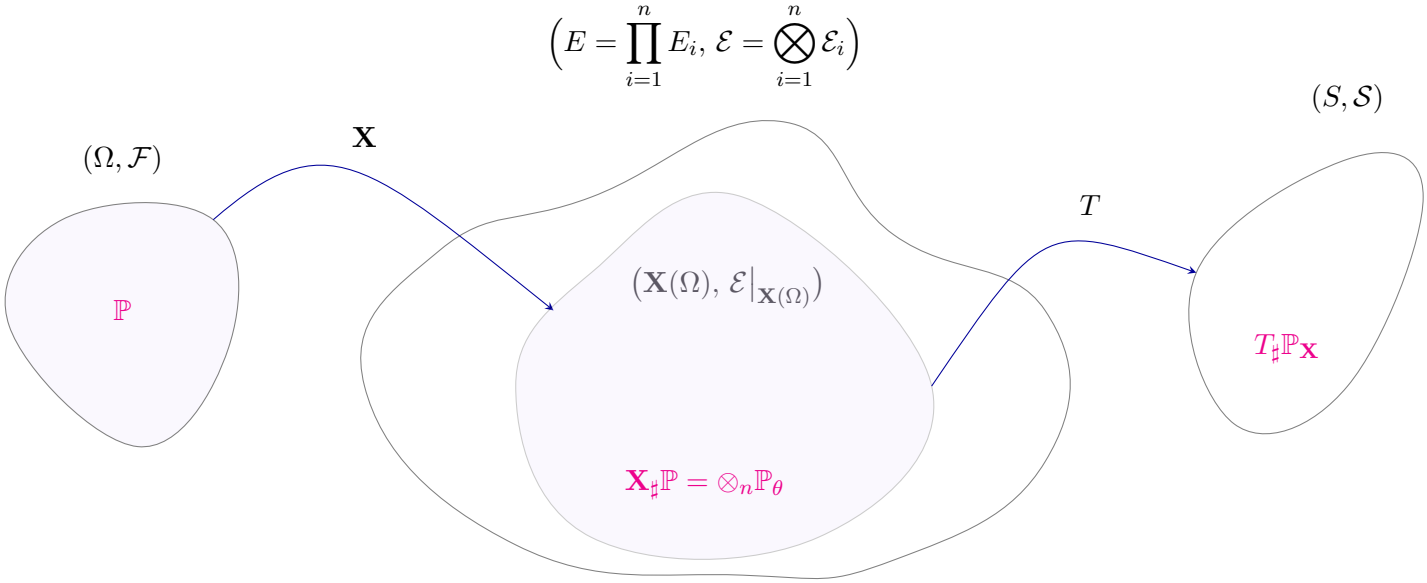


Figure 22.1: A visualisation to keep track of the spaces, maps between them, and the probability measures that live on their respective σ-algebras.

Now we'll derive an expression that characterises the conditional distribution of the sample $\mathbf{X}$ given T in two[1] equivalent ways. Afterwards, I'll explain the implications of sufficiency of T for θ.

Approach 1 — Reason through $(\Omega, \mathcal{F}, \mathbb{P})$

By our definition of a statistic, $T \in \text{Meas}_{\mathcal{E}|_{\mathbf{X}(\Omega)}, \mathcal{S}}(\mathbf{X}(\Omega); S) \iff T \circ \mathbf{X} \in \text{Meas}_{\sigma(\mathbf{X}), \mathcal{S}}(\Omega; S)$, and the latter in particular[2] implies $(\mathcal{F}, \mathcal{S})$-measurability of $T \circ \mathbf{X}$. Also, the earlier discussion of data reduction tells us that $\sigma(T \circ \mathbf{X}) \subseteq \sigma(\mathbf{X})$ as σ-algebras, and so we have a viable sub-σ-algebra to condition on. In practice, $(S, \mathcal{S})$ is some Borel space so suppose that's the case. Then the conditions of **Corollary 18.3.12** are satisfied by $Y = T \circ \mathbf{X}$, and so there exists a system of **proper** regular conditional probability measures generated by $T \circ \mathbf{X}$ on $\mathcal{F} = \sigma(\mathbf{X})$ i.e.

$\exists \kappa_\theta^{T \circ \mathbf{X}} \colon \sigma(\mathbf{X}) \times S \to [0, 1]$ s.t.

1. $\forall_{\mathbb{P}_{T \circ \mathbf{X}}}\, t \in S$, we have that $\kappa_\theta^{T \circ \mathbf{X}}(\cdot\,, t)$ is a probability measure **concentrated on** $(T \circ \mathbf{X})^{-1}(\{t\})$
2. for each $A \in \sigma(\mathbf{X})$, $\kappa_\theta^{T \circ \mathbf{X}}(A, \cdot)$ is a version of $\mathbb{P}[A \,|\, \mathcal{G} = \sigma(T \circ \mathbf{X})]$, and is $\mathbb{P}_{T \circ \mathbf{X}}$-integrable, and
3. $\forall A \in \sigma(\mathbf{X})$, $\forall D \in \mathcal{S}$, the following disintegration formula holds:

$$\mathbb{P}\big(A \cap (T \circ \mathbf{X})^{-1}(D)\big) = \int_D \kappa_\theta^{T \circ \mathbf{X}}(A, t)\, \mathrm{d}\mathbb{P}_{T \circ \mathbf{X}}(t).$$

[1] There is no need to do it both ways but it's a good sign that they're consistent.
[2] If we let $D \in \mathcal{S}$, then $(T \circ \mathbf{X})^{-1}(D) \in \sigma(\mathbf{X}) \subseteq \mathcal{F}$.

Approach 2 — Reason through $\left(\mathbf{X}(\Omega), \mathcal{E}\big|_{\mathbf{X}(\Omega)}, \otimes_n \mathbb{P}_\theta\right)$

Begin with the pushed-forward probability space in the subsection title. It's immediate from the measurability of T that $\sigma(T)$ is a sub-σ-algebra of $\mathcal{E}|_{\mathbf{X}(\Omega)}$. With the same assumption that $(S, \mathcal{S})$ is Borel, the same corollary with $Y = T$ this time gives us the existence of a system of proper regular conditional probabilities on $\mathcal{E}|_{\mathbf{X}(\Omega)}$ generated by T i.e.

$\exists \kappa_\theta^T : \mathcal{E}\big|_{\mathbf{X}(\Omega)} \times S \to [0,1]$ s.t.

1. for $T_\sharp \otimes_n \mathbb{P}_\theta$-a.e. $t \in S$, we have that $\kappa_\theta^T(\cdot, t)$ is a probability measure concentrated on $T^{-1}(\{t\})$,
2. for each $A \in \mathcal{E}\big|_{\mathbf{X}(\Omega)}$, $\kappa_\theta^T(A, \cdot)$ is a version of $(\otimes_n \mathbb{P}_\theta)[A \mid \sigma(T)]$, and is $T_\sharp \otimes_n \mathbb{P}_\theta$-integrable, and
3. $\forall B \in \mathcal{E}\big|_{\mathbf{X}(\Omega)}$, $\forall D \in \mathcal{S}$, the following disintegration formula holds:

$$\mathbb{P}_{\mathbf{X}}\big(B \cap T^{-1}(D)\big) = (\otimes_n \mathbb{P}_\theta)(B \cap T^{-1}(D)) = \int_D \kappa_\theta^T(B, t)\, \mathrm{d}(T_\sharp \otimes_n \mathbb{P}_\theta)(t).$$

These approaches are equivalent. Why?

The first disintegration formula holds for any $A \in \sigma(\mathbf{X})$ and $D \in \mathcal{S}$. Note that there exists some $B \in \mathcal{E}$ s.t. $A = \mathbf{X}^{-1}(B)$. Now we take the LHS from the disintegration formula and re-write it like so:

$$\begin{aligned}
\mathbb{P}\big(A \cap (T \circ \mathbf{X})^{-1}(D)\big) &= \mathbb{P}\big(\mathbf{X}^{-1}(B) \cap \mathbf{X}^{-1}(T^{-1}(D))\big) \\
&= \mathbb{P}\big(\mathbf{X}^{-1}(B \cap T^{-1}(D))\big) \\
&= \mathbb{P}_{\mathbf{X}}\big(B \cap T^{-1}(D)\big).
\end{aligned}$$

From this equality, we conclude that

$$\int_D \kappa_\theta^{T \circ \mathbf{X}}(A, t)\, \mathrm{d}\mathbb{P}_{T \circ \mathbf{X}}(t) = \int_D \kappa_\theta^T(B, t)\, \mathrm{d}(T_\sharp \mathbb{P}_{\mathbf{X}})(t)$$

i.e. that for every $A = \mathbf{X}^{-1}(B)$, we have that

$$\kappa_\theta^{T \circ \mathbf{X}}(A, t) = \kappa_\theta^T(B, t) \quad \text{for } \mathbb{P}_{T \circ \mathbf{X}}\text{-almost every } t \in S.$$

This mathematical consistency is entirely expected but good to verify nonetheless because it raises an important point. Though it's good to keep in mind that the underlying probability space is there (Approach 1), **one can typically avoid the pushforward measure** $\mathbf{X}_\sharp \mathbb{P} = \otimes_n \mathbb{P}_\theta$ by outright specifying n i.i.d. random variables, since one can then simply begin with the respective product distribution on $(E, \mathcal{E})$ which is typically $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$.

This obfuscation of the underlying probability space seems to be **the** way mathematical statistics is written.

What Sufficiency Means

A statistic T is called sufficient for θ iff $\sigma(T)$ is a sufficient σ-algebra for the family of probability measures $\{\mathbb{P}_{\mathbf{X}} = \otimes_n \mathbb{P}_\theta\}_{\theta \in \Theta}$, all defined on $(\mathbf{X}(\Omega), \mathcal{E}\big|_{\mathbf{X}(\Omega)})$. These measures represent the possible laws of the sample $\mathbf{X}$. Thus, the undergraduate definition of sufficiency for a parameter is truly a statement about sufficiency for a collection of probability measures that are indexed by $\theta \in \Theta$. To summarise:

The undergraduate statement that

> "the conditional distribution of the sample
> $\mathbf{X}$ given $T = t$ is independent of θ"

is a short way of saying that for every $\theta \in \Theta$, if such a conditional distribution

$$\kappa_\theta^T \colon \mathcal{E}\big|_{\mathbf{X}(\Omega)} \times S \to [0,1]$$

exists, then we can run through all $B \in \mathcal{E}|_{\mathbf{X}(\Omega)}$ and pick versions of $\mathbb{P}_{\mathbf{X}}[B \,|\, \sigma(T)]$ that are independent of θ. We denote this choice of kernel by dropping the θ and simply write

$$\kappa \colon \mathcal{E}\big|_{\mathbf{X}(\Omega)} \times S \to [0,1].$$

This manifests in the disintegration formula of the conditional probability as follows:

$$\mathbb{P}_{\mathbf{X}}\big(B \cap T^{-1}(D)\big) = \int_D \kappa(B,t)\,\mathrm{d}(T_\sharp \otimes_n \mathbb{P}_\theta)(t)$$

and so the θ-dependence of the law of $\mathbf{X}$ is entirely determined by $T_\sharp \mathbb{P}_{\mathbf{X}}$ i.e. $\mathbb{P}_{T\circ\mathbf{X}}$. This is precisely what Borovkov means in the following passage:

> Knowing $T \circ \mathbf{X}$ is sufficient to construct an estimator for the parameter θ; the rest of the data contained in the sample $\mathbf{X}$ is useless.
>
> [18, p. 116]

And in simpler terms, sufficiency is a statement about how the conditional distribution of $\mathbf{X}$ given $T \circ \mathbf{X}$ doesn't depend on θ since θ has already been accounted for in the information $\sigma(T \circ \mathbf{X})$ provides:

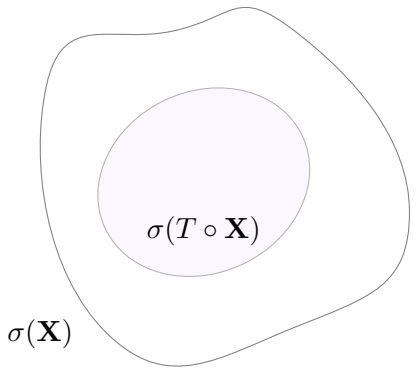


Figure 22.2: All the information required to determined θ is within the sub-σ-algebra $\sigma(T \circ \mathbf{X})$. Conditioning on this information (taking it as given), means the resulting conditional distribution of the sample $\mathbf{X}$ does not depend on θ i.e. the remaining information $\sigma(X) \setminus \sigma(T \circ \mathbf{X})$ is superfluous with respect to determining θ.

22.1.1 Factorisation Theorem

This is another way to characterise sufficient statistics. We start with the disintegration formula. For any $B \in \mathcal{E}\big|_{\mathbf{X}(\Omega)}$, and for any $D \in \mathcal{S}$:

$$\mathbb{P}_{\mathbf{X},\theta}\big(B \cap T^{-1}(D)\big) = \int_D \kappa_\theta^T(B,t)\,\mathrm{d}(T_\sharp \mathbb{P}_{\mathbf{X},\theta})(t).$$

Ideally, I'd like the following:

- For every $\mathbb{P}_{\mathbf{X},\theta}$ to have a density (i.e. that there exists some σ-finite measure μ_θ s.t. $\mathbb{P}_{\mathbf{X},\theta} \ll \mu$ so the LHS of the disintegration formula looks like the Lebesgue integral of some f_θ with respect to μ.
    - Preferably, I'd like for there to exist some dominating measure μ independent of θ for which $\mathbb{P}_{\mathbf{X},\theta} \ll \mu$ for every $\theta \in \Theta$. I believe this is possible because we're working with a Borel space $(\mathbf{X}(\Omega), \mathcal{E}|_{\mathbf{X}(\Omega)})$.
- For conditions that allow me to write the RHS as a Lebesgue integral with respect to the same measure μ.

Under these conditions, I can then compare f_θ with a product of two functions as per the famous Neyman-Fisher factorisation theorem for sufficient statistics.

Let's assume that there exists a dominating measure μ for the family $\{\mathbb{P}_{\mathbf{X},\theta}\}_{\theta\in\Theta}$, and so every $\mathbb{P}_{\mathbf{X},\theta}$ has density for μ-a.e. $x \in \mathbf{X}(\Omega)$:

$$f_\theta(x) = \frac{\mathrm{d}\mathbb{P}_{\mathbf{X},\theta}}{\mathrm{d}\mu}.$$

Furthermore, since T is sufficient for the family $\{\mathbb{P}_{\mathbf{X},\theta}\}_{\theta\in\Theta}$, there exists a choice of kernel $\kappa\colon \mathcal{E}|_{\mathbf{X}(\Omega)} \times S \to [0,1]$ that is independent of θ. With these two pieces of information, the disintegration formula becomes:

$$\int_{B\cap T^{-1}(D)} f_\theta(x)\,\mathrm{d}\mu(x) = \int_D \kappa(B,t)\,\mathrm{d}(T_\sharp \mathbb{P}_{\mathbf{X},\theta})(t).$$

Foresight tells me two things:

- For $T_\sharp\mathbb{P}_{\mathbf{X},\theta}$-a.e. $t \in S$, $\kappa(\cdot,t)$ is a probability measure, I can write

$$\kappa(B,t) = \int_B \mathrm{d}\kappa(\cdot,t)(x).$$

    - If I assume further that each $\kappa(\cdot,t)$ is dominated by the same μ as earlier, then for μ-a.e. x its density is given by
    $$h(\cdot,x) = \frac{\mathrm{d}\kappa(\cdot,t)}{\mathrm{d}\mu}.$$
    - There seems to be some formula that says because $\kappa(\cdot,\cdot)$ is measurable in the second coordinate, the Radon-Nikodym derivatives $(x,t) \longmapsto h(x,t)$ are jointly measurable.
    - Since the $\kappa(\cdot,t)$ are each concentrated on $T^{-1}(\{t\})$, each one assigns full mass to $T^{-1}(\{t\})$, and so it follows that
    $$0 = \kappa((T^{-1}(\{t\}))^{\mathsf{c}},t) = \int_{(T^{-1}(\{t\}))^{\mathsf{c}}} h(x,t)\,\mathrm{d}\mu(x)$$
    which implies that
    $$h(x,t) = 0 \text{ for } \mu\text{-a.e. } x \in (T^{-1}(\{t\}))^{\mathsf{c}}$$
    i.e. $h(x,t) > 0$ for μ-a.e. $x \in T^{-1}(\{t\})$.
    Thus, we conclude that $h(x,t) > 0 \implies T(x) = t$. Equivalently, one can write this as
    $$h(x,t) = h(x,T(x))\mathbb{1}_{T^{-1}(\{t\})}(x).$$

- If I then assume that the family $\{T_\sharp \mathbb{P}_{\mathbf{X},\theta}\}_{\theta\in\Theta}$ is dominated by some measure ν, then at some point I will need to exchange the order of integration (by Fubini-Tonelli) to get the Lebesgue integral of some expression (that I wish to compare with f_θ) with respect to μ.

If I take the above into account, I have:

$$\begin{aligned}
\int_{B\cap T^{-1}(D)} f_\theta(x)\,\mathrm{d}\mu(x) &= \int_D \kappa(B,t)\,\mathrm{d}(T_\sharp\mathbb{P}_{\mathbf{X},\theta})(t)\\
&= \int_D\int_B \mathrm{d}\kappa(\cdot,t)(x)\,\mathrm{d}(T_\sharp\mathbb{P}_{\mathbf{X},\theta})(t)\\
&= \int_D\int_B h(x,t)\,\mathrm{d}\mu(x)\,\mathrm{d}(T_\sharp\mathbb{P}_{\mathbf{X},\theta})(t)\\
&= \int_D\int_B h(x,t)\,\mathrm{d}\mu(x) g_\theta(t)\,\mathrm{d}\nu(t)\\
&= \int_B\int_D h(x,t) g_\theta(t)\,\mathrm{d}\nu(t)\,\mathrm{d}\mu(x) \qquad \text{by Fubini-Tonelli}\\
&= \int_B\int_D h(x,T(x))\mathbb{1}_{T^{-1}(\{t\})}(x) g_\theta(t)\,\mathrm{d}\nu(t)\,\mathrm{d}\mu(x)\\
&= \int_B h(x,T(x)) g_\theta(T(x))\,\mathrm{d}\mu(x)
\end{aligned}$$

and so we conclude that $f_\theta(x) = h(x,T(x))g_\theta(T(x))$ for μ-a.e. $x\in\mathbf{X}(\Omega)$.

Schervish [5, p. 89] states the theorem as follows (I've adapted the notation to suit my own conventions):

Theorem 22.1.3 (Theorem 2.21) Assume that $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is a parametric family such that $\mathbb{P}_\theta \ll \nu$ (σ-finite) for all θ and $\mathrm{d}\mathbb{P}_\theta/\,\mathrm{d}\nu(x) = f_\theta(x)$. Then $T(X)$ is sufficient for Θ iff there are functions m_1 and m_2 such that

$$f_\theta(x) = m_1(x)m_2(T(x),\theta), \text{ for all } \theta.$$

What's curious is that the proof of this theorem doesn't make so many domination assumptions, but the proof involves constructing a dominating measure that the author calls ν^*. Not really sure what to think of this, but I certainly believe my proof makes more assumptions than necessary.

According to [1], future topics on data reduction include:

1. **The Likelihood Principle** describes a function of the parameter, determined by the observed sample, that contains all the information about θ that's available from the sample.
2. **The Equivariance Principle** prescribes yet another method of data reduction that still preserves some important features of the model.

2025-12-12

I'm putting a halt to my studies on data reduction to study some Stochastic Processes and Machine Learning. I imagine the latter will bring me back to data reduction techniques.

CHAPTER 23

Temp

23.1 2025-09-21, Decomposing Spaces

I need to recover the link to the `.pdf` by Michael Betancourt but it offered some nice structure and notation to describe partitioning sets.

Consider a finite partition $\mathcal{P} = \{B_1, \ldots, B_n\}$ of a space X. The indexing implicitly defines a bijective index function that maps each cell to its corresponding integer index:

$$\begin{aligned} b_{\mathcal{P}} &: \mathcal{P} \longrightarrow \{1, \ldots, n\} \\ &: B_i \longmapsto i. \end{aligned}$$

We can also define an inclusion function that maps each point x in the ambient space X into the partition cell that contains it:

$$\begin{aligned} \iota_{\mathcal{P}} &: X \longrightarrow \mathcal{P} \\ &: x \longmapsto \{B_i \in \mathcal{P} : x \in B_i\}. \end{aligned}$$

Composing these functions defines a third that maps points to partition cell indices.

$$\begin{aligned} \phi_{\mathcal{P}} &: X \longrightarrow \{1, \ldots, n\} \\ &: x \longmapsto \{i \in \{1, \ldots, n\} : x_i \in B_i \in \mathcal{P}\}. \end{aligned}$$

Since $\mathcal{P}$ is a partition, each $x \in X$ belongs to only one partition cell. This means that $\phi_{\mathcal{P}}$ is surjective. The level set of $\phi_{\mathcal{P}}$ for a given index is

$$\begin{aligned} (\phi_{\mathcal{P}})^{-1}(\{i\}) &= \{x \in X : \phi_{\mathcal{P}}(x) = i\} \\ &= B_i. \end{aligned}$$

Consequently, we can completely reconstruct[1] the cells of the partition $\mathcal{P}$ from these level sets as follows:

$$\mathcal{P} = \{B_1 = (\phi_{\mathcal{P}})^{-1}(\{1\}), \ldots, B_n(\phi_{\mathcal{P}})^{-1}(\{n\})\}.$$

Different permutations of the labels define different index functions $b_{\mathcal{P}}$ and hence different composite maps $\phi_{\mathcal{P}}$. The level sets of these functions, however, are always the same, allowing us to work with whichever indexing might be most convenient in a given application. This implicit definition of a partition by a surjective function immediately generalises to any type of partition (countable, uncountable).

Every function $f : X \to Y$ decomposes the set X into level sets $f^{-1}(\{y\})$ that are not only disjoint, but also cover X i.e.

$$X = \bigsqcup_{y \in Y} f^{-1}(\{y\}).$$

If f is surjective, then $f^{-1}(\{y\}) \neq \emptyset$ for every $y \in Y$. Consequently, the level sets of every surjective function implicitly defines a partition where each cell is indexed by a unique output value.

- If $|Y| < \infty$, then the level sets of f define a finite partition (resp. countable, uncountable).

Example 23.1.1 An uncountable partition defined this way is given by the function

$$\begin{aligned} f &: \mathbb{R}^2 \longrightarrow \mathbb{R} \\ &: (x, y) \longmapsto \sqrt{x^2 + y^2} =: r. \end{aligned}$$

Note that the level sets $f^{-1}(\{r\})$ are concentric circles with centre $(0, 0)$ and radius r.

[1] The cells in a partition are unordered so the exact indexing we use is arbitrary.

Partitions consisting of measurable sets are particularly important in probability theory. Let $(X, \mathcal{F}, \mathbb{P})$ be a probability space. When the cells of a partition are all elements of $\mathcal{F}$, then $\mathcal{P} \subseteq \mathcal{F}$ and we call such a $\mathcal{P}$ a **measurable partition**.

Only when the output space is equipped with a σ-algebra that contains all the singletons, does a measurable, surjective function define a measurable partition.

23.1.1 Application to Conditioning

When it comes to defining conditional probability, any measurable subset $A \in \mathcal{F}$ that completely overlaps with the conditioning partition cell B_i, $A \cap B_j = B_j$, is allocated full conditional probability $\mathbb{P}(A \mid B_j) = 1$. If A doesn't overlap with B_j, the it's allocated zero conditional probability. Indeed, $\mathbb{P}(\cdot \mid B_j)$ is a probability measure, and is more "singular" than one would expect of a probability measure over X in the sense that all the conditional probability concentrates within B_j itself.

Intuitively, this suggests that we may interpret a conditional probability distribution given B_j as a restriction of $\mathbb{P}$ to a particular cell, and via the subspace σ-algebra we may view $\mathbb{P}(\cdot \mid B_j)$ as a map : $\mathcal{F}|_{B_j} \to [0, 1]$. Thus, we have two valid interpretations of $\mathbb{P}(\cdot \mid B_j)$ as a probability measure on $\mathcal{F}$ that concentrates on B_j, or a probability measure on $\mathcal{F}|_{B_j}$.

23.2 2025-10-13, Lebesgue-Stieltjes Measure

Proposition 23.2.1 Let $F \colon \mathbb{R} \to \mathbb{R}$ be non-decreasing and right continuous. Define

$$F(\pm\infty) = \lim_{x \to \pm\infty} F(x).$$

Set

- $\mathcal{A} = \{\emptyset\} \cup \left\{ \bigcup_{j=1}^{n} (a_j, b_j] : n \in \mathbb{N}, -\infty \leqslant a_1 < b_1 < a_2 < \ldots < b_n \leqslant \infty \right\}$
- $\mu_0(\emptyset) = 0$
- $\mu_0\Big(\bigcup_{j=1}^{n} (a_j, b_j] \Big) = \sum_{j=1}^{n} \big(F(b_j) - F(a_j) \big)$. for every element in $\mathcal{A}$.

In the above, replace $(a, b]$ by (a, b) when $b = \infty$.

Then, μ_0 is a pre-measure on $\mathcal{A}$.

From the above lemma, for any non-decreasing and right continuous $F \colon \mathbb{R} \to \mathbb{R}$, μ_0 is a pre-measure, and we know that $\mathcal{B}_{\mathbb{R}}$ is generated by the collection of half-open intervals $(a, b]$. Furthermore, μ_0 is σ-finite since $\mathbb{R} = \cup_{n \in \mathbb{N}} (n, n+1]$ and $\mu_0((n, n+1]) = 1 < \infty$. By Caratheodory's extension theorem, there exists a unique Borel measure μ_F on $\mathbb{R}$ that extends μ_0 to $\mathcal{B}_{\mathbb{R}}$.

Let $G \colon \mathbb{R} \to \mathbb{R}$ be another non-decreasing and right continuous function. Then $\mu_F = \mu_G$ iff $F - G$ is constant.

Proof.

$$\begin{aligned}
\mu_F = \mu_G &\iff \text{the corresponding pre-measures are equal} \\
&\iff \mu_{F_0}((a, b]) = \mu_{G_0}((a, b]) \quad \text{for all } a < b \\
&\iff F(b) - F(a) = G(b) - G(a) \quad \text{for all } a < b \\
&\iff \underbrace{F(b) - G(b)}_{=: (F-G)(b)} = \underbrace{F(a) - G(a)}_{=: (F-G)(a)} \quad \text{for all } a < b
\end{aligned}$$

■

Theorem 23.2.2 (Characterisation) If μ is a Borel measure on $\mathbb{R}$, and is finite on all bounded Borel sets, then $\exists F\colon \mathbb{R} \to \mathbb{R}$ that is right continuous and non-decreasing s.t.

$$\mu = \mu_F.$$

This F is given by

$$F(x) = \begin{cases} \mu((0,x]) & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -\mu((x,0]) & \text{if } x > 0. \end{cases}$$

Proof.

$$\mu_F((a,b]) = F(b) - F(a) = \begin{cases} \mu((0,b]) - \mu((0,a]) & \text{if } 0 \leqslant a < b \\ \mu((0,b]) + \mu((a,0]) & \text{if } a < 0 \leqslant b \\ -\mu((b,0]) + \mu((a,0]) & \text{if } a < b < 0. \end{cases}$$

■

23.3 2025-11-11, Hypothesis Testing

These are the things I learned about hypothesis testing from `brawthy` and `Catullus` in the Statistics Discord server. The original question from user `Cartesian` was:

> Are p-values just conditional probabilities? My understanding of them has always been
>
> $\mathbb{P}(\text{this sample statistic or one more extreme} \mid \text{null hypothesis})$.

This opened up a lot of questions. I don't know what a p-value is. I don't know what a null hypothesis is. However, I wondered if p-values involve testing the validity of a hypothesis? One of the responses was from `Catullus`:

> Suppose we have a family $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ of probability measures on some measurable space, we have some test statistic T, and we wish to test the null $\theta = \theta_0$ for some known value θ_0. Then the p value is $\mathbb{P}_{\theta_0}(T \geqslant t)$ where t is the value of T in your specific sample.
>
> (Note in particular that we're not conditioning, we're just working with a specific probability measure that corresponds to the null.)

This looks similar to the formulation of sufficiency, something I can understand apart from "we wish to test the null ... for some known value θ_0". What type of *mathematical object* is "the null"?

> `brawthy` set up a toy example as follows:
>
> Imagine you're conscripted to determine the rate of failure for GPUs. You have an assembly line and are told you can sample n number of GPUs to test per day. You talk to business management about what you think is an acceptable number of failures — maybe 5 in 20.
>
> Now you have some assumptions about the process that leads you to believe the Bernoulli model is acceptable for this (it's probably not a good model because I'd expect most manufacturing in this space to have some weird correlations with respect to the fabrication process but it works here). So you've

> packaged a statement about some population parameter, a model (I guess we should say likelihood because what is a model of not a combination of parameters and a likelihood) you think is acceptable (the Bernoulli — a probability measure!) and now you go forth and want to compute the p value under the assumption of correct specification of the model and some range of potential failure rate is acceptable.

I'd formalise this as follows. You have a sample $\mathbf{X}$ (random element) of GPUs, and we assume the underlying process is governed by some true probability measure in the family $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$, where $\mathbb{P}_\theta$ is shorthand for the distribution of the sample[2] $\mathbf{X}$. So the test statistic is some function, say T, of the sample values that represents something we want to test. Practically speaking, I guess the choice of a particular test statistic is two-fold:

1. We don't wish to lose any information in the summary of the sample values by applying this test statistic T, and
2. We wish for the CDF of the push-forward distribution $(T_\sharp\mathbb{P}_\theta)$ to be easier to interpret.

> `Catullus` corrected my final remark: [...] the point of using a test statistic is because we want some statistic whose distribution under the null is known and because we want a single value so we can compare it.

Furthermore, "the null" in our example is a statement hypothesised about the "tolerable number of failures." Mathematically speaking, the null hypothesis is a statement specifying some subspace of Θ in which the true parameter is posited (or hypothesised) to live e.g. $\theta \in \{\theta_0\}$ for some fixed θ_0 is the subspace corresponding to the null hypothesis that $\theta = \theta_0$ (in our case, that the failure rate is θ_0).

It's late in these notes to mention this but so far, we've been operating under the assumption that we've correctly specified the family of laws the sample's distribution could possibly live in. In other words, we assume that the "true" law that generated our sample is in the family we chose. This is called **the assumption of correct specification**.

Then a (the?) p-value is

$$\mathbb{P}_{\theta_0}(T \geqslant t) = \mathbb{P}_{\theta_0}\big(T^{-1}([t,+\infty))\big) = (T_\sharp\mathbb{P}_{\theta_0})([t,+\infty)).$$

[2] This is $\mathbf{X}_\sharp\mathbb{P}$ if we wish to specify the underlying probability space that governs the randomness of the experiment.

Chapter A

Rings $\overset{?}{\leftrightarrow}$ Algebras

The build-up of the theory for constructing measures across the literature I've encountered is not consistent; some authors use semi-rings, others use semi-algebras. There must be some common thread. I aim to reconcile both paths (or at least understand each one separately if they are indeed not connected).

Definition A.0.1 A **semi-ring** is a collection $\mathcal{I} \subseteq 2^X$ s.t.

- $\emptyset \in \mathcal{I}$,
- $A, B \in \mathcal{I} \implies A \cap B \in \mathcal{I}$,
- For $A, B \in \mathcal{I}$ s.t. $A \subseteq B$, we can write $B \setminus A$ as a finite pairwise disjoint union of $\{C_j\}_{j=1}^n \subseteq \mathcal{I}$ i.e.
$$B \setminus A = \bigsqcup_{j=1}^{n} C_j.$$

Definition A.0.2 A **ring** is a collection $\mathcal{R} \subseteq 2^X$ that is closed under finite unions and relative complements:

- $A, B \in \mathcal{R} \implies A \cup B \in \mathcal{R}$,
- $A, B \in \mathcal{R}$ s.t. $A \subseteq B \implies B \setminus A \in \mathcal{R}$.

Note that a ring is not defined to have the whole space X as a member. For our purposes, we have and will be defining outer measures using set functions from covers of our space. For example, every algebra $\mathcal{A}$ on X is trivially a cover (since $\emptyset \in \mathcal{A} \ni X$). The forthcoming statements about pre-measures on algebras equally apply when replacing 'algebra' with 'ring' because of the following lemma:

Lemma A.0.3 Let $\mathcal{A} \subseteq 2^X$. Then TFAE:

- $\mathcal{A}$ is an algebra.
- $\mathcal{A}$ is a ring and contains the whole space X.

Proof.

$\implies$ Suppose that $\mathcal{A}$ is an algebra. Then $X \in \mathcal{A}$ and $\mathcal{A}$ is closed under finite unions by assumption. All that remains to show is closure under relative complements. Let $A, B \in \mathcal{A}$ s.t. $A \subseteq B$. We wish to show that $B \setminus A \in \mathcal{A}$. Since $B \setminus A = B \cap A^{\mathsf{c}}$ and $A \in \mathcal{A} \implies A^{\mathsf{c}} \in \mathcal{A}$, we conclude that $B \setminus A \in \mathcal{A}$ by the closure of $\mathcal{A}$ under finite intersections.

$\impliedby$ For the converse, suppose that $\mathcal{A}$ is a ring s.t. $X \in \mathcal{A}$. Then $\mathcal{A}$ is certainly closed under finite unions by assumption. What remains to show is closure under (absolute) complements. Let $A \in \mathcal{A}$. We wish to show that $X \setminus A \in \mathcal{A}$. This is automatic from a ring's closure under relative complements since $A, X \in \mathcal{A}$ are s.t. $A \subseteq X$.

■

CHAPTER B

Extending Properties

$X \neq \emptyset$

Definition B.0.1 A subset $\mathcal{M}$ of 2^X is called a **λ-system**[1] if it satisfies:

(a) $X \in \mathcal{M}$

(b) Closure under relative complements: If $A, B \in \mathcal{M}$ with $A \subseteq B$, then $B \setminus A \in \mathcal{M}$

(c) Closure under increasing (monotone) sequences: If for every $n \in \mathbb{N}$: $A_n \in \mathcal{M}$ and $A_n \subseteq A_{n+1}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{M}$.

Proposition B.0.2 The intersection of any two λ-systems is a λ-system.

Definition B.0.3 The **λ-system generated by a collection $\mathcal{C} \subseteq 2^X$**, denoted by $\lambda(\mathcal{C})$, is defined as the intersection of all λ-systems $\mathcal{M}$ on X that contain $\mathcal{C}$.

B.1 π-λ Theorem

Theorem B.1.1 (π-λ Theorem) Let $\mathcal{C} \subseteq 2^X$ be closed under finite intersections.[2] Then $\lambda(\mathcal{C}) = \sigma(\mathcal{C})$.

Proof.

$\subseteq$ Any σ-algebra is a λ-system, so certainly $\lambda(\mathcal{C}) \subseteq \sigma(\mathcal{C})$.

$\supseteq$ It's enough to verify that $\lambda(\mathcal{C})$ is a σ-field (from which it follows that $\sigma(\mathcal{C})$ being the smallest σ-field containing $\mathcal{C}$ tells us that $\sigma(\mathcal{C}) \subseteq \lambda(\mathcal{C})$). Since $\lambda(\mathcal{C})$ is already a λ-system, it suffices to show that it's also closed under finite intersections.

For every $A \in 2^X$, define $\mathcal{M}_A \coloneqq \{B \in \lambda(\mathcal{C}) \colon A \cap B \in \lambda(\mathcal{C})\}$.

1. Fix $A \in \mathcal{C}$.

   If we can show that $\mathcal{C} \subseteq \mathcal{M}_A$ and that $\mathcal{M}_A$ is a λ-system, then we'll be able to conclude that $\lambda(\mathcal{C}) \subseteq \mathcal{M}_A$ i.e.

   $$(\forall A \in \mathcal{C})(\forall B \in \lambda(\mathcal{C})) \quad A \cap B \in \lambda(\mathcal{C}) \tag{*}$$

   i.e. $\lambda(\mathcal{C})$ is closed under intersection with A.

   - For the first part, $\mathcal{C}$ is closed under finite intersections so for any $B \in \mathcal{C}$, $B \cap A \in \mathcal{C}$ i.e. $\mathcal{C} \subseteq \mathcal{M}_A$.
   - To show that $\mathcal{M}_A$ is a λ-system.
     (a) Since $A \in \mathcal{C} \subseteq \lambda(\mathcal{C})$ and $\lambda(\mathcal{C})$ is closed under intersections, $A \cap X = A \in \lambda(\mathcal{C})$ i.e. $X \in \mathcal{M}_A$.

[1] This is the **monotone form** of a λ-system. The original **Dynkin form** is characterised by a collection $\mathcal{D}$ satisfying

- $\emptyset \in \mathcal{D}$,
- closure under absolute complements i.e. $A \in \mathcal{D} \implies X \setminus A \in \mathcal{D}$, and
- closure under countable unions of pairwise disjoint sets.

[2] This is known as a **π-system**.

(b) Suppose that $B_1, B_2 \in \mathcal{M}_A$ and $B_1 \subseteq B_2$. We wish to show that $B_2 \setminus B_1 \in \mathcal{M}_A$ i.e. that $A \cap (B_2 \setminus B_1) \in \lambda(\mathcal{C})$.

$$A \cap (B_2 \setminus B_1) = (A \cap B_2) \setminus (A \cap B_1)$$

which is the relative complement of two sets in $\lambda(\mathcal{C})$, and is thus in $\lambda(\mathcal{C})$.

(c) For closure under increasing limits, let $\{B_n\}_{n\in\mathbb{N}} \subseteq \mathcal{M}_A$ with $B_n \uparrow B$. We wish to show that $B \in \mathcal{M}_A$. Since $\lambda(\mathcal{C})$ is a λ-system, $\bigcup_{n\in\mathbb{N}} B_n =: B \in \lambda(\mathcal{C})$. Now note that

$$A \cap B = A \cap \left(\bigcup_{n\in\mathbb{N}} B_n \right) = \bigcup_{n\in\mathbb{N}} (A \cap B_n)$$

is a union of increasing elements of $\lambda(\mathcal{C})$ because for all $n \in \mathbb{N}$: $B_n \in \mathcal{M}_A$ and $B_n \subseteq B_{n+1} \implies A \cap B_n \subseteq A \cap B_{n+1}$. Thus, $B \in \mathcal{M}_A$.

2. Now we fix $A \in \lambda(\mathcal{C})$.
To show that $\mathcal{C} \subseteq \mathcal{M}_A$, let $C \in \mathcal{C}$. In step 1, we concluded that $\mathcal{M}_C \supseteq \lambda(\mathcal{C})$ i.e. (*):

$$(\forall C \in \mathcal{C},\ \forall A \in \lambda(\mathcal{C}))\quad C \cap A \in \lambda(\mathcal{C}).$$

Since $A \in \lambda(\mathcal{C})$, it follows that $A \in \mathcal{M}_C$ i.e. $C \cap A \in \lambda(\mathcal{C})$. But, this is precisely what it means for $C \in \mathcal{M}_A$. Hence, for every $C \in \mathcal{C}$, $C \in \mathcal{M}_A$ i.e. $\mathcal{C} \subseteq \mathcal{M}_A$. A similar argument to step 1 says that $\mathcal{M}_A$ is a λ-system, and it follows that for every $A \in \lambda(\mathcal{C})$, we have that $\lambda(\mathcal{C}) \subseteq \mathcal{M}_A$ i.e. $\lambda(\mathcal{C})$ is closed under finite intersections.

■

Corollary B.1.2 For *any* λ-system $\mathcal{M}$ containing $\mathcal{C}$, we get $\sigma(\mathcal{C}) = \lambda(\mathcal{C}) \subseteq \mathcal{M}$.

Chapter C

The Big Three

This entire section is based on the **videos** by Nicolas Lanchier on these three theorems.

C.1 Monotone Convergence Theorem

Theorem C.1.1 (Monotone Convergence Theorem) Let $\{X_n\}_{n\in\mathbb{N}}$ be a non-decreasing[1] sequence of non-negative measurable functions with pointwise limit X. Then, X is measurable and

$$\lim_{n\to\infty}\int_\Omega X\,\mathrm{d}\mu = \int_\Omega \Big(\lim_{n\to\infty} X_n\Big)\,\mathrm{d}\mu.$$

Lemma C.1.2 Let $\{A_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ be a non-decreasing sequence of sets. Then

$$\lim_{n\to\infty}\mu(A_n) = \mu\big(\lim_{n\to\infty} A_n\big).$$

This limit is similar to the MCT — a weak monotone convergence theorem that follows from the MCT by letting $X_n = \mathbb{1}_{A_n}$ and noticing that:

$$\int_\Omega \mathbb{1}_{A_n}\,\mathrm{d}\mu = \mu(A_n).$$

Proof. Since $\{A_n\}$ is increasing, its limit is the countable union $A := \bigcup_{n\in\mathbb{N}} A_n$. We can disjointify the sequence into a new sequence $\{B_n\}_{n\in\mathbb{N}}$ of mutually disjoint sets with the same union so that we're in a place to capitalise on σ-additivity:

$$\mu(A_n) = \mu\Big(\bigsqcup_{k=1}^{n} B_k\Big) = \sum_{k=1}^{n}\mu(B_k)$$

Therefore, we conclude that:

$$\begin{aligned}
\lim_{n\to\infty}\mu(A_n) &= \lim_{n\to\infty}\sum_{k=1}^{n}\mu(B_k)\\
&= \sum_{k=1}^{\infty}\mu(B_k)\\
&= \mu\Big(\bigsqcup_{k} B_k\Big)\\
&= \mu\Big(\bigcup_{n} A_n\Big)\\
&= \mu\Big(\lim_{n\to\infty} A_n\Big)
\end{aligned}$$

■

Now recall the statement[2] of the theorem:

$$\mathcal{F} \supseteq \{X_n\}_{n\in\mathbb{N}} \uparrow X \implies X\in\mathcal{F},\ \lim_{n\to\infty}\int_\Omega X_n\,\mathrm{d}\mu = \int_\Omega \Big(\lim_{n\to\infty} X_n\Big)\,\mathrm{d}\mu.$$

[1] Non-decreasing doesn't apply to the individual X_n themselves . What is meant by a non-decreasing sequence is that for all $\omega\in\Omega$ and $n\in\mathbb{N}$:

$$X_n(\omega) \leqslant X_{n+1}(\omega)$$

i.e. for each $\omega\in\Omega$, $\{X_n(\omega)\}_{n\in\mathbb{N}}$ is a non-decreasing sequence in $\mathbb{R}$ which has a limit (possibly ∞). (Should this be $\overline{\mathbb{R}}$?) Therefore, the pointwise limit $X(\omega)$ exists for every ω.

[2] This is shorthand for X being $(\mathcal{F},\mathcal{B}_{\mathbb{R}})$-measurable.

Proof of the MCT. We'll begin by observing that $\forall a \in \mathbb{R}$:

$$\begin{aligned} X^{-1}((-\infty, a]) &= \{X \leqslant a\} \\ &= \left\{\left(\lim_{n\to\infty} X_n\right) < a\right\} \\ &= \left\{\left(\sup_n X_n\right) < a\right\} \\ &= \{\omega \in \Omega : \forall n \in \mathbb{N}, X_n(\omega) < a\} \\ &= \bigcap_{n=1}^{\infty} \{X_n < a\} \\ &\in \mathcal{F} \end{aligned}$$

In the 3rd equality, we noted that for each $\omega \in \Omega$, $\{X_n(\omega)\}_{n\in\mathbb{N}}$ is non-decreasing and so the limit is the supremum. Finally, we conclude that $X \in \mathcal{F}$ since the collection $\{(-\infty, a] \colon a \in \mathbb{R}\}$ is a generating set for $\mathcal{B}_{\mathbb{R}}$.

Now we prove the equality by demonstrating both inequalities.

$\leqslant$

The integral is a monotone operator and $X_n \uparrow X$ so for every $n \in \mathbb{N}$:

$$\int_\Omega X_n \,\mathrm{d}\mu \leqslant \int_\Omega X \,\mathrm{d}\mu =: \int_\Omega \left(\lim_{k\to\infty} X_k\right) \mathrm{d}\mu$$

and taking the limit of this inequality as $n \to \infty$ gives the desired result.

$\geqslant$

The outline of this direction is to define an arbitrary non-negative simple function $0 \leqslant s \leqslant X$ and consider a sequence of sets comparing the sequence of X_n to s so that we may compare the integral of the X_n to s. So fix $\varepsilon > 0$ (small), and let s-simple be s.t. $0 \leqslant s \leqslant X$. The sets $A_n = \{X_n \geqslant (1-\varepsilon)s\}$ will be useful and we will apply the previous lemma on them. First we must verify that $\{A_n\}_{n\in\mathbb{N}}$ is a non-decreasing sequence:

If $\omega \in A_n$, then $X_n(\omega) > (1-\varepsilon)s(\omega)$, but by monotonicity, $X_{n+1}(\omega) \geqslant X_n(\omega)$ so $X_{n+1}(\omega) \geqslant (1-\varepsilon)s(\omega)$ i.e. $\omega \in A_{n+1}$. Thus, $\{A_n\}_{n\in\mathbb{N}}$ is an non-decreasing sequence.

WHAT IS THE LIMIT OF THIS INCREASING SEQUENCE OF SETS?

Since X is the limit of the X_n, for each $\omega \in \Omega$ we know that $X_n(\omega) \uparrow X(\omega)$, and so there exists some $N \in \mathbb{N}$ s.t. $n \geqslant N \implies X_n(\omega) \geqslant (1-\varepsilon)X(\omega)$. Since $0 \leqslant s \leqslant X$, we have the inequality:

$$X_n(\omega) \geqslant (1-\varepsilon)X(\omega) \geqslant (1-\varepsilon)s(\omega)$$

i.e. $\omega \in A_n$. This means that $\Omega \subset A_n$ for large enough n. Thus, $A_n \uparrow \Omega$.

Now we want to compare the Lebesgue integral of X_n w.r.t. μ with the Lebesgue integral of s. By construction, we only know how to compare X_n and s on A_n.

$$\begin{aligned} \int_\Omega X_n \,\mathrm{d}\mu &\geqslant \int_\Omega X_n \mathbb{1}_{A_n} \,\mathrm{d}\mu \quad \text{by monotonicity} \\ &= \int_{A_n} X_n \,\mathrm{d}\mu \\ &\geqslant \int_{A_n} (1-\varepsilon)s \,\mathrm{d}\mu \\ &= (1-\varepsilon)\int_{A_n} s \,\mathrm{d}\mu \quad \text{by linearity} \end{aligned}$$

The next step is to understand the quantity in the final equality. We already have that $A_n \uparrow \Omega$. Let's introduce $\nu(A) \coloneqq \int_{A_n} s \,\mathrm{d}\mu$. Since s and the indicator functions of $A_n \in \mathcal{F}$ are non-negative,

it's clear that ν is non-negative. It's clear that if $s \equiv 1$, $\nu(A) = \mu(A)$. More generally, it's simple to conclude that for any non-negative simple function s, ν is a non-negative measure.

Now we're in the position to apply the previous lemma with ν on the non-decreasing sequence $A_n \uparrow \Omega$.

$$\int_\Omega X_n \,\mathrm{d}\mu \geqslant (1-\varepsilon) \int_{A_n} s \,\mathrm{d}\mu =: (1-\varepsilon)\nu(A_n)$$

By **Lemma C.1.2**, $\nu(A_n) \uparrow \nu(\Omega) = \int_\Omega s \,\mathrm{d}\mu$ and so taking the limit of the expression above as $n \to \infty$ gives

$$\lim_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu \geqslant (1-\varepsilon) \int_\Omega s \,\mathrm{d}\mu.$$

This expression is true for every $\varepsilon > 0$, and $0 \leqslant s \leqslant X$ is arbitrary, so we may take the limit as $\varepsilon \to 0^+$ and the supremum over such simple s to get:

$$\begin{aligned}
\lim_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu &\geqslant \lim_{\varepsilon\to 0^+} \sup_{0\leqslant s\leqslant X} (1-\varepsilon) \int_\Omega s \,\mathrm{d}\mu \\
&= \sup_{0\leqslant s\leqslant X} \int_\Omega s \,\mathrm{d}\mu \\
&=: \int_\Omega X \,\mathrm{d}\mu \\
&=: \int_\Omega \left(\lim_{n\to\infty} X_n\right) \mathrm{d}\mu
\end{aligned}$$

where the second equality is our definition of the Lebesgue integral of non-negative $X \in \mathcal{F}$. ■

C.2 Dominated Convergence Theorem

Theorem C.2.1 (Dominated Convergence Theorem) Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of measurable functions dominated by some integrable function and with pointwise limit X. Then X is integrable,

$$\lim_{n\to\infty} \int_\Omega |X_n - X| \,\mathrm{d}\mu = 0 \quad \text{and} \quad \lim_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu = \int_\Omega \left(\lim_{n\to\infty} X_n\right) \mathrm{d}\mu.$$

Remarks

- The phrase 'dominated by some integrable function' means that $\exists Y \in \mathcal{F}$ which is integrable (i.e. $\int_\Omega |Y| \,\mathrm{d}\mu < \infty$) s.t. $\forall n \in \mathbb{N}\colon\ |X_n| \leqslant Y$ is the domination condition.
- To contrast with the MCT, the DCT doesn't assume monotonicity of $\{X_n\}$. Thus, there's no guarantee of convergence (pointwise) to a limit. To remedy this, we add a pointwise limit to our assumptions.

  $$\text{i.e. the MCT's } \begin{pmatrix}\text{non-negative}\\ \text{and monotone}\end{pmatrix} \text{ is replaced by } \begin{pmatrix}\text{dominating function}\\ \exists\text{pointwise limit}\end{pmatrix}$$

  This new set of assumptions offers

  - a slightly stronger conclusion that X is integrable, and
  - a stronger type of convergence in the L^1 sense.

The first step is to prove Fatou's lemma:

C.2.1 Fatou's Lemma

Theorem C.2.2 (Fatou's Lemma) Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of non-negative measurable functions. Then

$$\int_\Omega \left(\liminf_{n\to\infty} X_n\right) \mathrm{d}\mu \leqslant \liminf_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu.$$

Remarks C.2.3

- The conditions of Fatou's lemma don't guarantee any convergence but we can still define a lim inf and the result is an inequality.

Proof. Note that

$$\liminf_{n\to\infty} X_n \coloneqq \lim_{n\to\infty} \underbrace{\left(\inf_{k\geqslant n} X_k\right)}_{=: Z_n}$$

and Z_n is non-decreasing by definition.[3] Thus, $Z_n \uparrow \liminf_{n\to\infty} X_n$.

Now we can apply the monotone convergence theorem to the non-decreasing sequence of non-negative positive measurable functions Z_n:

$$\begin{aligned}\lim_{n\to\infty} \int_\Omega Z_n \,\mathrm{d}\mu &= \int_\Omega \lim_{n\to\infty} Z_n \,\mathrm{d}\mu \\ &= \int_\Omega \liminf_{n\to\infty} X_n \,\mathrm{d}\mu\end{aligned}$$

Now for the statement of the theorem:

$$\begin{aligned}\int_\Omega \left(\liminf_{n\to\infty} X_n\right) \mathrm{d}\mu &\coloneqq \int_\Omega \lim_{n\to\infty} \left(\inf_{k\geqslant n} X_n\right) \mathrm{d}\mu \\ &=: \int_\Omega \lim_{n\to\infty} Z_n \,\mathrm{d}\mu \\ &= \lim_{n\to\infty} \int_\Omega Z_n \,\mathrm{d}\mu \quad \text{by the MCT} \\ &= \liminf_{n\to\infty} \int_\Omega Z_n \,\mathrm{d}\mu \\ &\leqslant \liminf_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mu\end{aligned}$$

The penultimate equality is because if the limit exists, then it certainly agrees with the lim inf (and the lim sup). The final inequality follows from monotonicity of the integral. Since $Z_n = \inf_{k\geqslant n} X_k$, then certainly $Z_n \leqslant X_n$ because X_n is included in the collection of measurable functions we're taking the inf of. ■

Proof of the DCT.

1. - X is by definition the pointwise limit of the X_n, and $|\cdot|$ is a continuous function so:

$$|X| = \left|\lim_{n\to\infty} X_n\right| = \lim_{n\to\infty} |X_n|.$$

   - Combining the dominating assumption $\forall n \in \mathbb{N}\colon |X_n| \leqslant Y$ with the point above implies that

$$|X| = \lim_{n\to\infty} |X_n| \leqslant |Y|$$

     i.e. $|X| \leqslant |Y|$.
   - Finally, by monotonicity of the integral, we conclude that

$$\int_\Omega |X| \,\mathrm{d}\mu \leqslant \int_\Omega |Y| \,\mathrm{d}\mu < \infty.$$

2. Apply Fatou's lemma to the sequence of functions $Z_n = 2Y - |X_n - X|$.

[3] Informally, looking at smaller and smaller sets as you push the tail along means the infimum is non-decreasing.

Check the assumptions:

- Z_n-measurable? Sums and differences of measurable functions are measurable.
- $Z_n \geqslant 0$? Triangle inequality.

$$\begin{aligned}\liminf_{n\to\infty} Z_n &= \liminf_{n\to\infty}(2Y - |X_n - X|) \\ &= \left(\liminf_{n\to\infty} 2Y\right) - \underbrace{\left(\limsup_{n\to\infty} |X_n - X|\right)}_{=0} \\ &= 2Y\end{aligned}$$

By Fatou's lemma:

$$\begin{aligned}\int_\Omega 2Y \,\mathrm{d}\mu &= \int \liminf_{n\to\infty} Z_n \,\mathrm{d}\mu \\ &\leqslant \liminf_{n\to\infty} \int_\Omega Z_n \,\mathrm{d}\mu \quad \text{Fatou} \\ &= \liminf_{n\to\infty} \int_\Omega (2Y - |X_n - X|) \,\mathrm{d}\mu \\ &= \int_\Omega 2Y \,\mathrm{d}\mu - \underbrace{\limsup_{n\to\infty} \underbrace{|X_n - X|}_{\geqslant 0} \,\mathrm{d}\mu}_{\geqslant 0} \quad \text{by linearity} \\ &\leqslant \int_\Omega 2Y \,\mathrm{d}\mu\end{aligned}$$

and we are done because we have an expression

$$\int_\Omega 2Y \,\mathrm{d}\mu \leqslant \ldots \leqslant \ldots \leqslant \int_\Omega 2Y \,\mathrm{d}\mu$$

which forces us to conclude the inequalities are equalities and thus

$$\limsup_{n\to\infty} \int_\Omega |X_n - X| \,\mathrm{d}\mu = 0.$$

Now note that since $\int_\Omega |X_n - X| \,\mathrm{d}\mu$ is non-negative, the lim inf must be non-negative, and lim inf $\leqslant$ lim sup so they are both equal to 0, and so the limit

$$\lim_{n\to\infty} |X_n - X| \,\mathrm{d}\mu = 0.$$

3. Now consider the difference:

$$\begin{aligned}\left|\int_\Omega X_n \,\mathrm{d}\mu - \int_\Omega X \,\mathrm{d}\mu\right| &= \left|\int_\Omega (X_n - X) \,\mathrm{d}\mu\right| \quad \text{by linearity} \\ &\leqslant \left|\int_\Omega |X_n - X| \,\mathrm{d}\mu\right| \quad \begin{array}{l}\text{by monotonicity of the integral} \\ \text{since } (X_n - X) \leqslant |X_n - X|\end{array} \\ &= \int_\Omega |X_n - X| \,\mathrm{d}\mu \overset{n\to\infty}{\longrightarrow} 0 \quad \text{by step 2, concluding the DCT.}\end{aligned}$$

■

C.3 Non-Example for MCT/DCT

Let $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \lambda)$. Consider the piecewise linear functions X_n illustrated below. Each X_n is supported[4] on its respective half-open interval e.g. X_1 is supported on $(1/2, 1]$, X_2 on $(1/4, 1/2]$ etc.

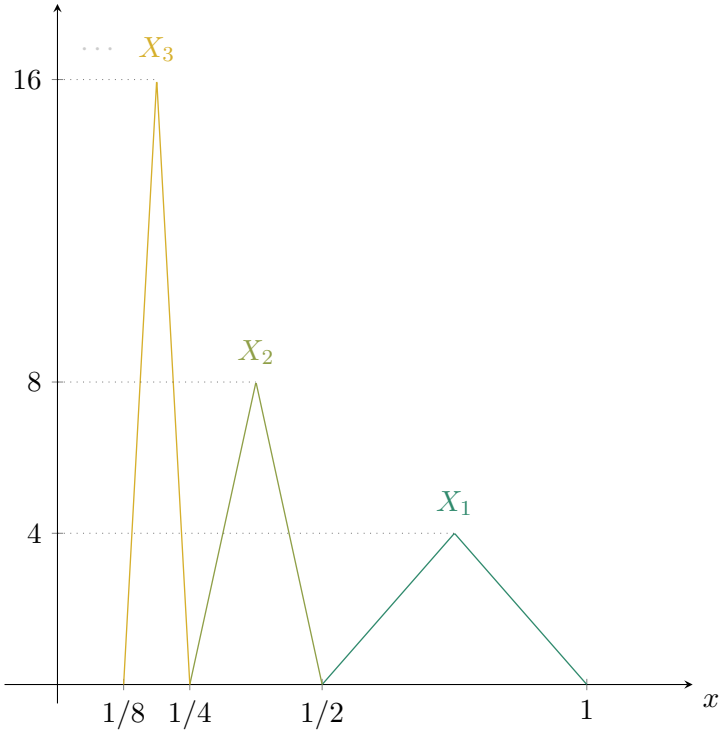


Figure C.1: Plots of X_1, X_2, and X_3 on their respective supports.

Each X_n, supported on $(a, b]$ with highest point (c, h) where $c = (a+b)/2$ is the midpoint of $(a, b]$, and h is the height of the triangle, is of the form

$$X_n(x) = \begin{cases} h - \dfrac{2h}{b-a}\left|x - \dfrac{a+b}{2}\right|, & x \in (a, b] \\ 0, & \text{otherwise.} \end{cases}$$

Now let $a = (1/2)^{n+1}$, $b = (1/2)^n$, $h = 2^{n+1}$, and so $c = 3/(2^{n+1})$ to obtain X_n.

For each $n \in \mathbb{N}$, X_n takes the form

$$X_n(x) = \begin{cases} 2^{n+1} - 2^{2n+3}\left|x - \dfrac{3}{2^{n+1}}\right|, & x \in \left((1/2)^{n+1}, (1/2)^n\right] \\ 0, & \text{otherwise.} \end{cases}$$

The supports of the X_n are disjoint so there's no monotonicity of the sequence on their supports. The smallest measurable function Y that dominates all the X_n is their sum; it's not integrable.

- $\displaystyle\int_{\mathbb{R}} X_1 \, \mathrm{d}\lambda = \text{area(triangle)} = \tfrac{1}{2}(1 - \tfrac{1}{2})(4) = 1$
- $\displaystyle\int_{\mathbb{R}} X_2 \, \mathrm{d}\lambda = \text{area(triangle)} = \tfrac{1}{2}(\tfrac{1}{2} - \tfrac{1}{4})(8) = 1$

⋮ Proceeding inductively, each step halves the base length and doubles the height of the triangle.

[4] This is the traditional (non-probabilistic) definition of support of a function; the set of points on which the function is non-zero.

- $\int_{\mathbb{R}} X_n \, \mathrm{d}\lambda = 1$ for every $n \in \mathbb{N}$

$$\therefore \lim_{n\to\infty} \int_{\mathbb{R}} X_n \, \mathrm{d}\lambda = 1.$$

Therefore, the Lebesgue integral of Y with respect to λ is equal to $\sum_{n=1}^{\infty} 1 = \infty$. The domination condition is not satisfied.

Now we focus on the limit of the X_n as $n \to \infty$. The support of every X_n is a subset of $\mathbb{R}_{\geqslant 0}$ i.e. $X_n(x) = 0$ for all $x \leqslant 0$. Take $x > 0$. Then we can always find an $N \in \mathbb{N}$ large enough s.t. $(1/2)^N < x$. This means that X_{n+1} has support $((1/2)^{N+1}, (1/2)^N]$ i.e. $X_{n+1}(x) = 0$. The same applies to $X_{n+2}(x) = 0$, $X_{n+3}(x) = 0$ and so on. This means that eventually, the limit of $X_n(x)$ for any $x > 0$ is zero.

The only value missing is $x = 0$ itself. However, with the way the X_n were defined, $X_n(0) = 0$ for all n. Therefore, $\lim_{n\to\infty} X_n = 0$. Clearly, the Lebesgue integral of the limit of the X_n w.r.t λ is zero.

Thus, we've constructed a sequence of measurable functions with

$$\infty = \lim_{n\to\infty} \int_{\mathbb{R}} X_n \, \mathrm{d}\lambda \neq \int_{\mathbb{R}} \lim_{n\to\infty} X_n \, \mathrm{d}\lambda = 0,$$

so the conditions of monotonicity and the assumption of domination in the MCT and DCT, respectively, are both important.

C.4 Fubini's Theorem

Theorem C.4.1 (Fubini's Theorem) Let $(\Omega_S, \mathcal{F}_S, \mu_S)$ and $(\Omega_T, \mathcal{F}_T, \mu_T)$ be two σ-finite measure spaces. Define:

- $\Omega = \Omega_S \times \Omega_T$
- $\mathcal{F} = \sigma(\{A \times B \colon A \in \mathcal{F}_S,\, B \in \mathcal{F}_T\})$
- The unique product measure μ on $\mathcal{F}$ satisfying $\mu(A \times B) = \mu_S(A)\mu_T(B)$

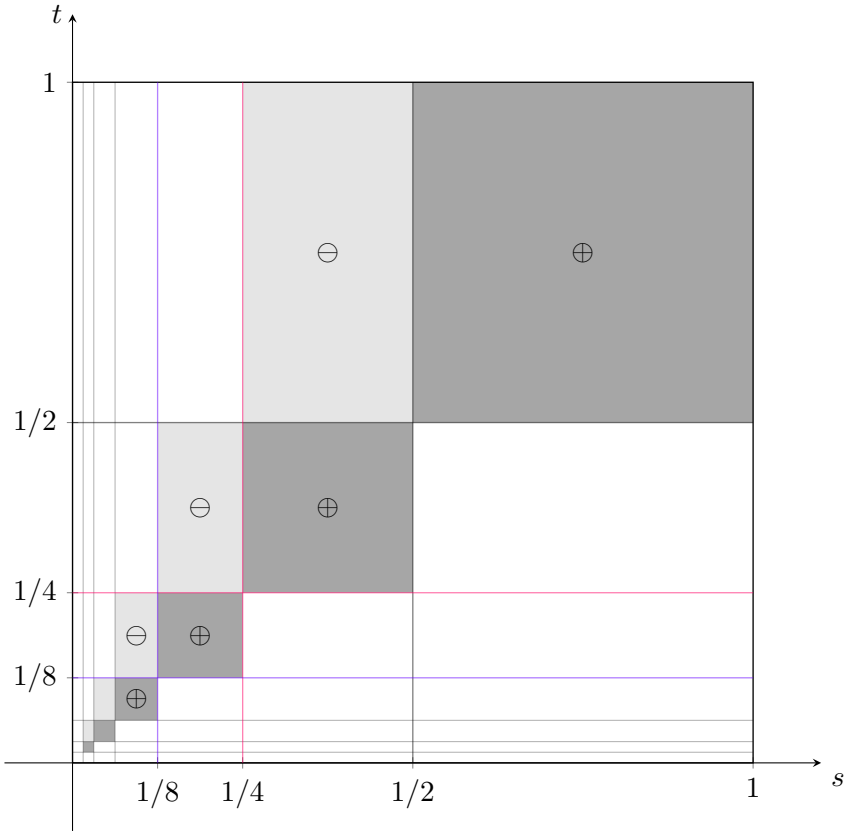
For every measurable function $X \colon \Omega \to \mathbb{R}$ that is either non-negative or integrable on a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$,

$$\int_{\Omega} X \, \mathrm{d}\mu = \int_{\Omega_S} \int_{\Omega_T} X \, \mathrm{d}\mu_T \, \mathrm{d}\mu_S = \int_{\Omega_T} \int_{\Omega_S} X \, \mathrm{d}\mu_S \, \mathrm{d}\mu_T.$$

C.5 Non-Example for Fubini's Theorem

Let $\Omega = [0,1] \times [0,1]$, $\mu_S = \mu_T = \lambda$.

- Subdivide the unit square as follows (black, red, blue, and so on...)
- and label $\oplus$ for each square on the diagonal, and $\ominus$ above each $\oplus$.



We're going to use the counterexample from the MCT/DCT section to construct a function X of two variables (s, t) in the following way:

- On each square with a $\oplus$, starting from the top-right $\left[\frac{1}{2}, 1\right] \times \left[\frac{1}{2}, 1\right]$, we define $X_1(s)X_1(t)$.
    - and proceed inductively down the main diagonal of $\oplus$ squares i.e. on $\left[\frac{1}{4}, \frac{1}{2}\right] \times \left[\frac{1}{4}, \frac{1}{2}\right]$ define $X_2(s)X_2(t)$ etc.
- Now take the highest $\ominus$ in the diagram $\left[\frac{1}{4}, \frac{1}{2}\right] \times \left[\frac{1}{2}, 1\right]$ and define $-X_2(s)X_1(t)$ on it.
- ... follow the pattern.

Assume the function is zero elsewhere (everywhere that isn't shaded). Then $X > 0$ on the interiors of the $\oplus$ squares, and $X < 0$ on the interiors of the $\ominus$ rectangles. The negative sign is so that our function doesn't satisfy the non-negative requirement of Fubini's theorem.
Visually, our function X looks like:

- square-based pyramids of increasing height on the $\oplus$ squares as you get closer to the origin,
- and on the $\ominus$ rectangles, they face downwards but the magnitude of the height increases as the squares get closer to the origin.

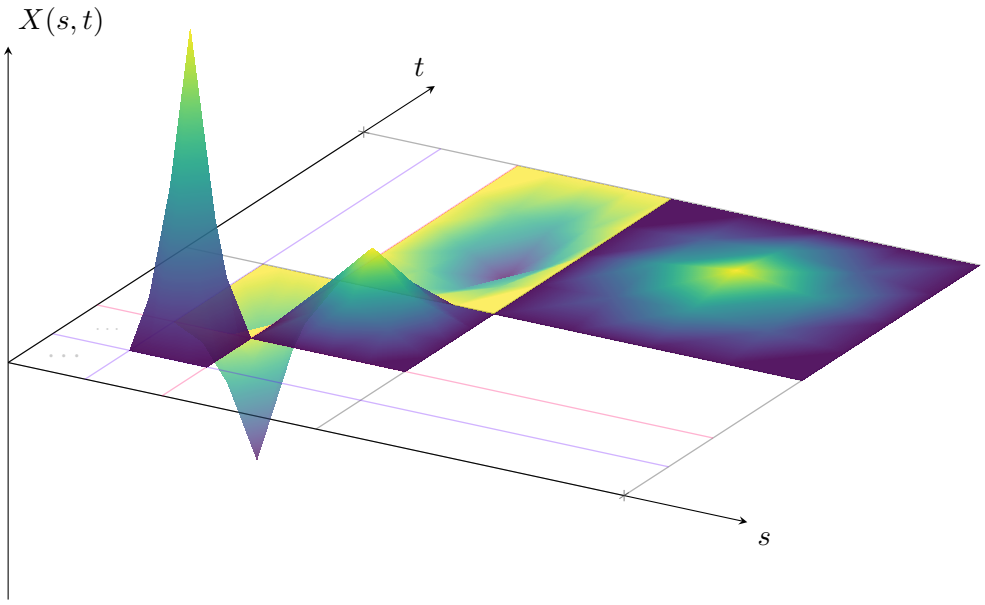


Figure C.2: A plot of the constructed $X(s,t)$.

Now for the calculations:

- Fix t first and consider:

$$\begin{aligned}\int_0^1 X(s,t)\,\mathrm{d}s &= \int_0^1 X_n(s)X_n(t)\,\mathrm{d}s - \int_0^1 X_{n+1}(s)X_n(t)\,\mathrm{d}s \quad \text{for some } n\\ &= X_n(t)\Big(\underbrace{\int_0^1 X_n(s)\,\mathrm{d}s}_{=1} - \underbrace{\int_0^1 X_{n+1}(s)\,\mathrm{d}s}_{=1}\Big) = 0\end{aligned}$$

  Therefore, $\displaystyle\int_0^1\int_0^1 X(s,t)\,\mathrm{d}s\,\mathrm{d}t = 0$.

- Fix $s < 1/2$ first and consider the integral across a vertical line:

$$\begin{aligned}\int_0^1 X(s,t)\,\mathrm{d}t &= \int_0^1 X_n(s)X_n(t)\,\mathrm{d}t - \int_0^1 X_n(s)X_{n-1}(t)\,\mathrm{d}t \quad \text{for } n>1\\ &= X_n(s)\Big(\int_0^1 X_n(t)\,\mathrm{d}t - \int_0^1 X_{n-1}(t)\,\mathrm{d}t\Big) = 0.\end{aligned}$$

- Fix $s > 1/2$. No $\ominus$ rectangle is crossed but the first square is:

$$\int_0^1 X(s,t)\,\mathrm{d}t = \int_0^1 X_1(s)X_1(t)\,\mathrm{d}t = X_1(s).$$

  Therefore, $\displaystyle\int_0^1\int_0^1 X(s,t)\,\mathrm{d}t\,\mathrm{d}s = \int_0^1 X_1(s)\,\mathrm{d}s = 1$.

Also, note that $|X|$ is a bunch of positive pyramids increasing in height (as we get closer to the origin) so the integral of $|X|$ is infinite i.e. X is not integrable. So the assumptions of Fubini's theorem are not satisfied for our constructed X. The order of integration matters.

CHAPTER D

Conditional Independence

This chapter also contains results from [3] but I'm figuring out where to put it.

D.1 Relating Conditional Expectation and Independence

Then a theorem that characterises independence in terms of conditional expectations:

Theorem D.1.1 Two sub-σ-algebras $\mathcal{G}_1$ and $\mathcal{G}_2$ of $\mathcal{F}$ are independent iff $\forall B \in \mathcal{G}_2$, we have $\mathbb{E}[\mathbb{1}_B \,|\, \mathcal{G}_1] = \mathbb{P}(B)$. Furthermore, if $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent, we have that $\forall X \geqslant 0$ s.t. $X \in \mathcal{G}_2$ (or every $X \in L^1(\Omega, \mathcal{G}_2, \mathbb{P})$):

$$\mathbb{E}[X \,|\, \mathcal{G}_1] = \mathbb{E}(X).$$

Proof.

$\Longrightarrow$ Suppose that $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent, and let $X \geqslant 0$ be $\mathcal{G}_2$-measurable. Then, for any non-negative $\mathcal{G}_1$ measurable random variable Z, we have that X and Z are independent and thus

$$\mathbb{E}(ZX) = \mathbb{E}(Z)\mathbb{E}(X) = \mathbb{E}(Z\mathbb{E}(X))$$

and so the constant random variable $\mathbb{E}(X)$ satisfies the characteristic property of $\mathbb{E}[X \,|\, \mathcal{G}_1]$. This implies that $\mathbb{P}$-a.s.

$$\mathbb{E}[X \,|\, \mathcal{G}_1] = \mathbb{E}(X).$$

Then we conclude by noting that in particular,

$$\mathbb{E}[\mathbb{1}_B \,|\, \mathcal{G}_1] = \mathbb{E}(\mathbb{1}_B) = \mathbb{P}(B).$$

The same result holds for integrable X by considering $X = X^+ - X^-$.

$\Longleftarrow$ Conversely, suppose that $\forall B \in \mathcal{G}_2 \colon\ \mathbb{E}[\mathbb{1}_B \,|\, \mathcal{G}_1] = \mathbb{P}(B)$. Then for every $A \in \mathcal{G}_1$, and for every $B \in \mathcal{G}_2$:

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{E}(\mathbb{1}_{A\cap B}) \\
&= \mathbb{E}(\mathbb{1}_A \mathbb{1}_B) \\
&= \mathbb{E}(\mathbb{1}_A \mathbb{E}[\mathbb{1}_B \,|\, \mathcal{G}_1]) \quad \text{by the averaging property of } \mathbb{E}(\mathbb{1}_B \,|\, \mathcal{G}_1) \\
&= \mathbb{E}(\mathbb{1}_A \mathbb{P}(B)) \\
&= \mathbb{P}(A)\mathbb{P}(B)
\end{aligned}$$

i.e. $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent σ-algebras.

■

Remarks D.1.2 Let X and Y be two real random variables. The random variables that are $\sigma(X)$-measurable are exactly the measurable functions of X. If we let $\mathcal{G}_1 = \sigma(Y)$ and $\mathcal{G}_2 = \sigma(X)$, the preceding theorem tells us that $X \in \mathcal{G}_2$ and $Y \in \mathcal{G}_1$ are independent iff

$$\mathbb{E}[h(X) \,|\, Y] = \mathbb{E}(h(X))$$

for every Borel-measurable function $h \colon \mathbb{R} \to \mathbb{R}$ s.t. $\mathbb{E}(|h(X)|) < \infty$ (i.e. $h(X) \in L^1$).

The next theorem states informally that, given a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$, if you have two random variables X, Y such that:

- X is independent from $\mathcal{G}$
- Y is $\mathcal{G}$-measurable

then conditioning a function $g(X, Y)$ on $\mathcal{G}$ means that Y behaves like a constant, so the best approximation of $g(X, Y)$ when knowing $\mathcal{G}$ is given by integrating $g(\,\cdot\,, Y)$ with respect to the law of X.

Theorem D.1.3 Let $(E, \mathcal{E})$ and $(S, \mathcal{S})$ be two measurable spaces, and let X and Y be two random variables taking values in E and S respectively. Assume that X is independent of $\mathcal{G}$ and that Y is $\mathcal{G}$-measurable. Then for every $(\mathcal{E} \otimes \mathcal{S})$-measurable function $g\colon E \times S \to [0, +\infty)$:

$$\mathbb{E}[g(X, Y) \,|\, \mathcal{G}] = \int_\Omega g(x, Y)\, \mathrm{d}\mathbb{P}_X,$$

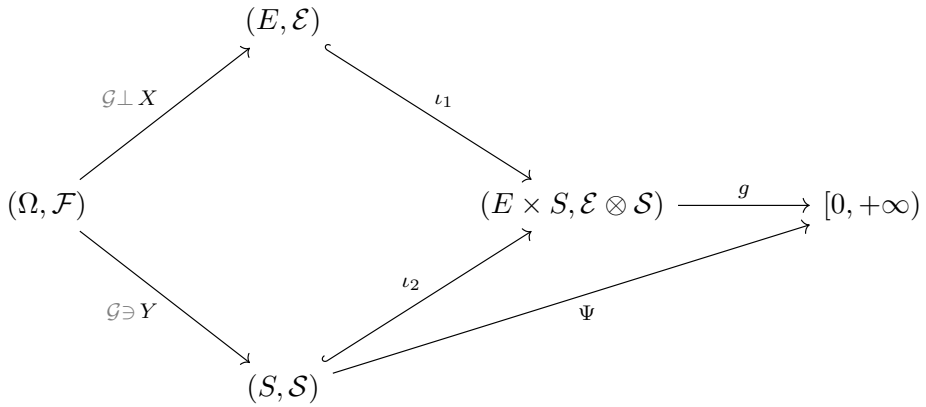
where the RHS is the composition of the random variable Y with the mapping $\Psi\colon S \to [0, +\infty)$ defined by

$$\Psi(y) = \int_\Omega g(x, y)\, \mathrm{d}\mathbb{P}_X.$$

Proof. We need to show that for any $Z \geqslant 0$ that is $\mathcal{G}$-measurable, we have

$$\mathbb{E}(g(X, Y)Z) = \mathbb{E}(\Psi(Y)Z).$$

The picture is:



and the conclusion is that $\Psi \circ Y = \mathbb{E}[g(X, Y) \,|\, \mathcal{G}]$.

Write $\mathbb{P}_{(X,Y,Z)}$ for the law of the triple (X, Y, Z), which is a probability measure on $\mathcal{E} \times \mathcal{S} \times \mathcal{B}_{[0,+\infty)}$. Since X is independent of $\mathcal{G}$, X is independent of the pair (Y, Z), and therefore

$$\mathbb{P}_{(X,Y,Z)} = \mathbb{P}_X \otimes \mathbb{P}_{(Y,Z)}.$$

Then,

$$\begin{aligned}
\mathbb{E}(g(X, Y)Z) &= \int_{E \times S \times [0,+\infty)} gZ\, \mathrm{d}\mathbb{P}_{(X,Y,Z)} \\
&= \int_{E \times S \times [0,+\infty)} gZ\, \mathrm{d}\mathbb{P}_X \otimes \mathbb{P}_{(Y,Z)} \\
&= \int_{S \times [0,+\infty)} z \left(\int_E g(x, y)\, \mathrm{d}\mathbb{P}_X \right) \mathrm{d}\mathbb{P}_{(Y,Z)} \quad \text{by Fubini's Theorem} \\
&= \int_{S \times [0,+\infty)} z\Psi(y)\, \mathrm{d}\mathbb{P}_{(Y,Z)} \\
&= \mathbb{E}(\Psi(Y)Z)
\end{aligned}$$

■

Then a proposition about $\mathcal{G}_1 \vee \mathcal{G}_2$ (the smallest σ-algebra that contains both $\mathcal{G}_1$ and $\mathcal{G}_2$ — sub-σ-algebras of $\mathcal{F}$).

Proposition D.1.4 Let Z be a random variable in L^1, and let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two sub-σ-algebras of $\mathcal{F}$. Assume that $\mathcal{G}_2$ is independent of $\sigma(Z) \vee \mathcal{G}_1$. Then

$$\mathbb{E}[Z \,|\, \mathcal{G}_1 \vee \mathcal{G}_2] = \mathbb{E}[Z \,|\, \mathcal{G}_1].$$

Proof. It suffices to prove the equality

$$\mathbb{E}(\mathbb{1}_A Z) = \mathbb{E}(\mathbb{1}_A \mathbb{E}[Z \,|\, \mathcal{G}_1])$$

holds for every $A \in \mathcal{G}_1 \vee \mathcal{G}_2$. Consider the case where $A = B \cap C$ with $B \in \mathcal{G}_1$, $C \in \mathcal{G}_2$. Then we have

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_A Z) &= \mathbb{E}(\mathbb{1}_B \mathbb{1}_C Z) \\
&= \mathbb{E}(\mathbb{1}_C \mathbb{1}_B Z) \\
&= \mathbb{E}(\mathbb{1}_C)\mathbb{E}(\mathbb{1}_B Z) \quad \text{since } \mathcal{G}_2 \perp \sigma(Z) \vee \mathcal{G}_1 \\
&= \mathbb{P}(C)\mathbb{E}(\mathbb{1}_B Z) \\
&= \mathbb{E}(\mathbb{P}(C)\mathbb{1}_B Z) \\
&= \mathbb{E}(\mathbb{P}(C)\mathbb{1}_B \mathbb{E}[Z \,|\, \mathcal{G}_1]) \quad \text{by the averaging property of } \mathbb{E}[Z \,|\, \mathcal{G}_1] \text{ since } \mathbb{P}(C)\mathbb{1}_B \in \mathcal{G} \\
&= \mathbb{E}(\mathbb{1}_C \mathbb{1}_B \mathbb{E}[Z \,|\, \mathcal{G}_1]) \\
&= \mathbb{E}(\mathbb{1}_A \mathbb{E}[Z \,|\, \mathcal{G}_1])
\end{aligned}$$

Replacing the constant $\mathbb{P}(C)$ with $\mathbb{1}_C$ **within** the expectation is legitimate because $\mathbb{1}_C$ is independent of $\mathbb{1}_B$ and $\mathbb{E}[Z \,|\, \mathcal{G}_1]$.

Thus, the class of all sets $A \in \mathcal{G}_1 \vee \mathcal{G}_2$ that satisfy

$$\mathbb{E}(\mathbb{1}_A Z) = \mathbb{E}(\mathbb{1}_A \mathbb{E}[Z \,|\, \mathcal{G}_1]) \qquad (\dagger)$$

contains a class closed under finite intersections (a π-system) that generate the σ-field $\mathcal{G}_1 \vee \mathcal{G}_2$. An application of the π-λ theorem shows that ($\dagger$) holds for **every** $A \in \mathcal{G}_1 \vee \mathcal{G}_2$. ■

Chapter E

Future Topics

E.1 Hypothesis Tests

Hypothesis testing is a general term for assessing whether sample data is consistent or otherwise with *statements made about the population.* [19]

Definition E.1.1 A ***statistical hypothesis*** is a hypothesis concerning the parameter or form of the probability distribution for a designated population or populations, or, more generally, of a probabilistic mechanism which is supposed to generate the observations. [4]

A null hypothesis is typically the 'no difference' or 'no association' hypothesis to be tested (usually by means of a *significance test*) against an alternative hypothesis that postulates non-zero difference or association.

Definition E.1.2 An effect is said to be **significant** if the value of the statistic used to test it lies outside acceptable limits, so that there's strong evidence against the hypothesis that the effect is not present. A ***test of significance*** is one which, by use of a test statistic, purports to provide a test of the hypothesis that the effect is absent. By extension, the critical values of the statistics are themselves called significant. [4]

Bibliography

[1] George Casella and Roger Berger. *Statistical Inference*. CRC Press, 2nd edition, 2024.

[2] Nicolas Lanchier. *Stochastic Modeling*. Springer, 2017.

[3] Jean-François Le Gall. *Measure Theory, Probability, and Stochastic Processes*. Springer, 2022.

[4] Yadolah Dodge. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, USA, 2003.

[5] Mark John Schervish. *Theory of Statistics*. Springer Science & Business Media, 2012.

[6] Dennis David Wackerly, William Mendenhall III, and Richard Lewis Scheaffer. *Mathematical Statistics with Applications*. Duxbury Press, 6th edition, 2001.

[7] Dennis David Wackerly, William Mendenhall III, and Richard Lewis Scheaffer. *Mathematical Statistics with Applications*. Thomson Brooks/Cole, 7th edition, 2008.

[8] Gerald Budge Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2nd edition, 1999.

[9] Kai Lai Chung. *Elementary Probability Theory with Stochastic Processes*. Springer, 3rd edition, 1979.

[10] Jun Shao. *Mathematical Statistics*. Springer, 2nd edition, 1999.

[11] Yiping Cheng. *A Mathematically Sensible Explanation of the Concept of Statistical Population*. arXiv Pre-Print, arXiv:1704.01732, 2017.

[12] Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

[13] Krishna Balasundaram Athreya and Soumendra Nath Lahiri. *Measure Theory and Probability Theory*. Springer, 2006.

[14] Vladimir Igorevich Bogachev. *Measure Theory Vol II*. Springer, 2006.

[15] Malempati Madhusudana Rao and Randall J Swift. *Probability Theory with Applications*. Springer, 2nd edition, 2006.

[16] Michel Loève. *Probability Theory II*. Springer, 4th edition, 1978.

[17] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 3rd edition, 1995.

[18] Aleksandr Alekseevich Borovkov. *Mathematical Statistics*. CRC Press, 1999.

[19] Anders Skrondal and Brian Sidney Everitt. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 4th edition, 2010.