Probability for Statistical Science
Draft Version, September 2023

Joseph K. Blitzstein and Carl N. Morris
Department of Statistics, Harvard University

1

Introduction: The Power of Probability

Probability theory is nothing but common sense reduced to calculation.... It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
– Pierre Simon Laplace (translated from *Théorie analytique des probabilités*, 1812)

I graduated from Douglass College without distinction. I was in the top 98% of my class and damn glad to be there. I slept in the library and daydreamed my way through history lecture. I failed math twice, never fully grasping probability theory. I mean, first off, who cares if you pick a black ball or a white ball out of the bag? And second, if you're bent over about the color, don't leave it to chance. Look in the damn bag and pick the color you want.
– Stephanie Plum (character in the novel *Hard Eight* by Janet Evanovich)

1.1 Why Study Probability?

For those of us who do not ordinarily spend our days picking balls from bags or urns, or indulging in games of chance, why study probability? Randomness and uncertainty surround us, yet humans are notoriously inaccurate in their intuitions about chance and coincidences. Probability gives precise language and tools for quantifying uncertainty, making it indispensable in a vast number of fields. To list a baker's dozen of these fields:

1. Physics: Einstein asserted that "God does not play dice with the universe," but randomness is fundamental to the modern understanding (or lack thereof) of quantum mechanics. Meanwhile, statistical mechanics relies on probability to describe the behavior of large collections of moving particles.
2. Genetics and bioinformatics: the mechanism by which genes are inherited is probabilistic; probability is also crucial in studying the evolution of DNA sequences, and patterns ("motifs") in the sequences.
3. Computer science: in recent years, many computational problems have been found to have fast solutions via *randomized algorithms*, which artificially introduce randomization into deterministic problems. Even for a deterministic algorithm, we often would like to know how well it performs "on average"; making this notion precise and computing the average running time requires probability.
4. Information theory and coding: probability plays an essential role in defining *entropy* and quantifying *information*, as developed by Claude Shannon. A famous result of Shannon showed that it is possible to transmit information with nearly perfect reliability, even

through noisy channels, using random codes; only much later were deterministic codes constructed coming close to the random codes in performance.

5. Mathematics: probability is thriving mathematically, having overcome early prejudices of it being merely measure theory with the whole space having measure 1. Moreover, deep connections with other areas of mathematics have been found. For example, Paul Erdős developed a method for proving the *existence* of an object with desired properties by showing that the *probability* is positive of having those properties, in a suitably defined space. This may sound bizarre at first, since it would seem that computing the probability of the properties holding is a harder problem than checking whether it is possible for the properties to hold; but only a bound showing the probability is positive is needed, and it is often easier to give such a bound than to give an explicit construction.
6. Meteorology and forecasting: what does it mean when the weather forecaster says there is a 30% chance of rain tomorrow? Such statements are routinely made on the evening news, but how was the 30% arrived at? Can the claim ever be verified or falsified, considering that tomorrow it either will or won't rain?
7. Sociology: much of the sociology literature is deterministic despite how difficult it is to predict the behavior of human beings. Yet probability is increasingly being needed in problems on all scales, such as modeling measurement errors and modeling the probability of two people knowing each other in a social network.
8. Economics: probability is pervasive in economics, especially in econometrics (such as in working with random error terms) and in game theory (where one often needs to randomize between several strategies to obtain a so-called *Nash equilibrium*).
9. Public Health: probability is becoming increasingly crucial in many applications to public health. For example, in survival analysis we need probability distributions to model how long someone will live; in epidemiology we need probability to predict the spread of a disease; and randomized controlled experiments, which were one of the most important breakthroughs in scientific medicine in the 20th century, rely essentially on randomization at the design stage and an understanding of probability at the analysis stage.
10. Law: what could be more fundamental to a criminal trial than the question of what is the probability that the defendant is guilty, given all the evidence?
11. Gambling: games of chance provide both a historically grounded and a practical reason for studying probability.
12. Finance: probability is extremely widely-used in quantitative finance (though some would consider mentioning "finance" to be redundant, in view of the previous item on the list).
13. Statistics: please read the rest of this book!

1.2 Language and Tools for Statistical Science

This book develops probability as a language and tool for *statistical science*, construed broadly: probability gives a framework needed for modeling and exploring the variability in data. This is in contrast with probability texts of the "probability for probabilists" flavor, with heavy doses of measure-theoretic details, and emphasis on mathematical concerns rather than applicability. Certain fundamental tools for statisticians often slip through the cracks, by being too theoretically difficult for the "elementary" books and ignored as "trivial" by most graduate level books. On the other hand, neglecting the elegant and powerful tools of probability hampers the statistician's ability to understand distributions, create estimators, perform simulations, and build models.

David Williams [15] light-heartedly but aptly notes that "Probability and Statistics used to be married; then they separated; then they got divorced; now they hardly ever see each other... this book is a move towards much-needed reconciliation."

We continue in this spirit of reconciliation, showing the power of probability while still emphasizing statistical sense and common sense. Probability and statistical inference are two sides of the same coin. Given an unknown parameter θ, probability allows us to assess how likely various outcomes are for the data y; given the data y, statistical inference allows us to estimate the parameter θ and quantify its uncertainty.

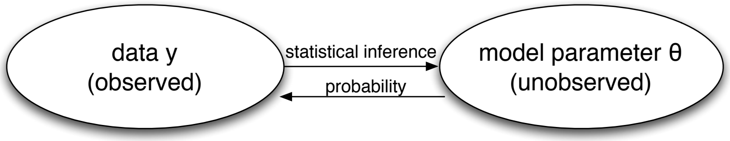


Figure 1.1 Probability and statistical inference as two sides of the same coin: the former says what outcomes are "average" and how likely various outcomes are, given θ, while the latter quantifies the uncertainty about θ, given observed data y.

This is illustrated in Figure 1.1. This simple picture can be extended to allow for multiple models, missing data, and future observations. In particular, the two directions are linked by the idea of *prediction*: given the data y, what can we say about a future observation y'? To answer this question, we need to go "there and back again": from y to estimates of θ and then from θ to a prediction for y', taking into account both the uncertainty about θ and the additional uncertainty inherent with each new observation.

Probability provides the language and tools needed for statistical inference and prediction, for quantifying uncertainty, for making sense of statements like "the outcome was consistent with being the result of pure chance," for understanding both random data generating processes and randomization artificially imposed (e.g., in clinical trials), and for simulations needed for complex models (e.g., via Markov chain Monte Carlo).

Our approach emphasizes *scientific reasoning via conditioning* to build models by breaking complicated problems into smaller, simpler problems, the *fundamental distributions* needed for a great many parametric models, and the *object-oriented* notions of *stories*, *representations*, and *characterizations* to connect these distributions to each other and to the world. We briefly describe these ideas below, along with our approach to measure theory; the following chapters develop the ideas in detail.

1.3 Object-Oriented Approach, via Stories, Representations, and Characterizations

Object-oriented programming is an approach to programming which emphasizes designing reusable, self-sufficient modules called *classes*, which can be thought of as blueprints for specific instantiations known as *objects*. For example, a class for a (simulated) cat would define characteristics such as color, length of fur, and food preferences, together with methods (functions) such as the ability to eat, sleep, and purr; a cat object is a specific realization of the cat class. Object-oriented programming centers around interactions between objects, in contrast with "procedural" programming, where a program is thought of as a list of commands to carry out.

In this book we take an approach that we call *object-oriented probability*, where distributions play the role of classes and random variables play the role of objects. Confusing a random variable with its distribution is akin to confusing a house with the blueprint for the house. We build a toolkit of useful distributions, which serve as reusable building blocks for many useful models and allow us to create many useful random variables. The connections between distributions allow for efficient arguments and give structure to the menagerie of possible distributions. Two of the main ideas in object-oriented programming are *inheritance* and *modularity*; these are mirrored in our approach via *representation*, *characterization*, and *stories*.

Inheritance is the ability of a class to produce offshoots which are related to the "parent class", thus allowing for code reuse and a well-organized collection of classes. We show how the most important distributions are connected by representation (described more below), which often gives simpler proofs, avoids tedious calculations, and allows for a form of code reuse in *shared distributions*, i.e., simulations and computations for one distribution can easily be extended to various other distributions.

More precisely, a *representation* is an expression for a distribution in terms of random variables, rather than density (or other) formulas. As a simple example for readers familiar with the Cauchy distribution (if you have not yet studied this distribution, stay tuned for Chapter 3), contrast two ways of specifying that a random variable C has the Cauchy distribution: by saying that C can be represented as the ratio of two independent standard Normal random variables, or by saying that C has density $f(c) = \frac{1}{\pi(1+c^2)}$. The representation definition is easier to remember, makes it easier to see why this distribution might arise, and makes it immediate to see that $1/C$ is also Cauchy without needing any calculus. Along these lines, we can also readily see that $(1+C)/(1-C)$ is Cauchy, using the fact (discussed in Chapter 3) that the sum and difference of independent standard Normal random variables are independent. With this approach we can obtain many results with basic algebra instead of messy calculus.

This approach leads to a logical progression for introducing the major distributions,

Class: Exponential distribution $\mu \cdot \text{Expo}$

Parameter: scale and mean parameter μ (rate parameter $\lambda = 1/\mu$)
Story and characterization: waiting memorylessly in continuous time
Representation: $X \sim -\mu \log U$ with $U \sim \text{Unif}$
Generalizations and offshoots: Gamma, Weibull

Figure 1.2 The Exponential distribution as a "class" in the object-oriented sense; the class proclaims its stories, representations, and characterizations, and can often be used without worrying about the implementation details given by the density function. Exponential random variables are objects instantiated from this class.

showing how they are all related rather than presenting a disjointed collection of densities. Many proofs become more elementary, relying on algebra rather than calculus. This in turn helps one to *anticipate* new results, since the proofs are more transparent.

Modularity is the idea of separating the "public" functionality of code (made available through some interface) from the implementation. The user can then employ the tool without worrying about messy implementation details or whether the implementation will be changed in the future. A closely related notion is *encapsulation*, where an object is bundled up with its internal details hidden from view: as long as it does its job correctly, as advertised in the interface to the outside world, it is entitled to its privacy. Analogously, we emphasize *characterizations*: many of the most important distributions can be defined by special properties they have, rather than with a seemingly unmotivated density formula.

For example, readers may remember that the memoryless property characterizes the Exponential distribution. Figure 1.2 shows a possible class diagram for the Exponential distribution: the properties that make the Exponential distribution special are separated from the "implementation details" of its density. Of course, we often need to use the implementation details too, but we can understand a lot about the Exponential just through its stories, representations, and characterizations, and without a compelling story there would be much less reason to consider the distribution in the first place.

As another example of a characterization, we will later prove the sublimely beautiful fact that if X and Y are independent, identically distributed random variables such that $X + Y$ and $X - Y$ are independent, then X and Y are both Normal. This result is not only elegant, but also it helps explain the nice properties of the t-test in statistics, and it even has applications to the "cocktail party problem" of separating a signal (such as the sound of several simultaneous conversations) into the original source signals. The Normal also enjoys many other characterizations, such as having maximum entropy for a given mean and variance; together, these characterizations help justify the Normal distribution's special status in statistics.

Relatedly, we emphasize *stories*: each important distribution comes with a story, which may be a *generative story* explaining why data may follow this distribution, a *characterization story* explaining why if certain properties are desired this should be the choice, or a *representation story* explaining why the distribution arises naturally from working with other important distributions. A story is more than just an example or special case; it is an application which puts a theorem or concept into focus without losing generality, making the theorem or concept easier to remember and understand.

A reader could easily find after reading several graduate probability texts that he or she has only the foggiest understanding of the most important distributions in statistics, such as the Gamma, Beta, t, LogNormal, and Multivariate Normal distributions. These distributions are ubiquitous in statistical modeling, and are teeming with characterizations, representations, and stories. We explore these (and other) fundamental distributions and show how they are connected.

In many fields, the difference between a novice and an expert has been described as follows: the novice scrambles to memorize a large number of seemingly disconnected facts and formulas, whereas the expert sees a unified structure in which a few principles and ideas connect these facts coherently. The object-oriented approach to probability facilitates developing this kind of coherent picture, by revealing the connections between distributions, and between distributions and the world.

1.4 Measure in good measure

Measure theory is a branch of mathematical analysis which provides probability theory with various foundational results. How much measure theory statistics graduate students "should" know is a controversial issue. Heavy doses often involve a level of abstraction beyond anything they are used to; more importantly, they may be unhappy at not seeing any practical relevance. After all, in practice one never encounters a nonmeasurable set (with respect to Lebesgue measure, which is introduced in the next chapter; in fact, it is impossible to write down explicitly a set of real numbers which is not Lebesgue measurable; the existence of such sets relies on the Axiom of Choice). So why expend so much energy on proving that various sets and functions are measurable? Moreover, a heavy emphasis on measure-theoretic niceties necessarily squeezes out other topics due to time constraints.

The other extreme – neglecting measure theory entirely – is also ill-advised. If nothing else, some basic comfort with the language of measure theory is a necessary part of "statistical culture." Furthermore, measure theory gives a unified notation and language for distributions and provides a firm foundation for probability, as well as providing useful physical analogies (e.g., thinking of probability as mass). Here, we introduce the language, concepts, and a few key results of measure theory, but do not dwell on abstractions and technicalities.

1.5 Scientific reasoning through conditioning

Science proceeds iteratively from special cases to the general, and back. We encourage such reasoning by emphasizing looking for and studying "SNoTEs" (simple non-trivial examples) whenever possible, and by discussing how to break complicated problems into smaller, simpler problems. A fundamental tool for this – and throughout statistics – is *conditioning.* We discuss why *all* probabilities are conditional probabilities, and that how to choose what to condition is a crucial and powerful skill. In most subjects, wishing you knew something is just an idle fantasy. In probability, if you wish you knew something you *can* act as though you know that thing, by conditioning on it! In an important sense, we would even claim that:

Conditioning is the soul of statistics.

Dennis Lindley emphasized the fundamental importance of conditioning by coining the phrase "Law of the Extended Conversation," which we abbreviate to LotEC. Lindley chose such a chatty name to make the point that in science we start by conversing about some event A, but to make progress we often have to "extend the conversation" by discussing what happens for A under each possibility for other event B, but then to study B it may help to further condition on some event C, and so on. In return for having more cases to deal with, the conditional probabilities we need may be much simpler to think about and compute than the original probability of interest. This dovetails with the object-oriented approach, by using the strategy of solving a problem by breaking it into smaller, simpler building blocks.

1.6 Tools of the Trade

A relatively small number of problem-solving strategies, which we call tools, tricks, and devices, are useful again and again in probability and statistics. Seemingly very different proofs are often variations on the same theme, employing one or a combination of these tools. We discuss such tools as they arise, and name and collect them in an appendix. In general, proofs are presented to emphasize the problem-solving strategies and tactics underlying the proof, rather than trying for ultra-terse proofs that obscure the essential ideas. This approach of decomposing proofs into simpler reusable tools linked by a clear strategy is again in keeping with the object-oriented perspective. The reader is strongly encouraged to look for further examples of these tools, and to build his or her own tools! See Appendix D for a list of the major tricks, tools, and devices discussed in this book.

1.7 How to Read This Book: Active Learning

Pencil Problems (✎). It is nearly impossible to understand probability without **doing** probability, by solving many problems and thinking hard about what the various theorems really mean (especially by building simple examples and counterexamples). Richard Feynman's blackboard displayed the following quotation: "What I cannot create, I do not understand." Create probability rather than passively reading about it!

To encourage the reader to be as active as possible in the learning process, we intersperse problems ("pencils," denoted by ✎) throughout. The pencil icon is meant to encourage the reader to pick up a pencil and try the problem *now*, rather than separating the reading process from the problem-solving process by relegating the problems to the end of the chapter. Of course there are additional problems at the ends of chapters, for further practice, but we strongly recommend solving pencil problems while reading rather than as an afterthought.

Biohazards (☣). We also point out and emphasize common mistakes and misconceptions ("biohazards," denoted by ☣). Understanding why an idea is wrong can be just as important as understanding why an idea is right, and a mistake that is obviously a mistake *after* the material has been grasped may seem very natural initially.

1.8 A Note on Capitalization

We capitalize distribution names in this book, e.g., writing Normal instead of normal for that distribution. Most books and papers follow the reverse convention, although the American Statistical Association has recently given approval for authors to choose either convention. Why then capitalize?

Proper names mark the special distributions as important, while clarifying and simplifying our prose. Do normal distributions arise normally? For rare events, is the Poisson distribution the normal distribution? These questions are ambiguous when Normal is not capitalized. As adjectives, normal and geometric need "distribution." Normal and Geometric can stand alone.

Capitalization puts Binomials on a par with Bernoullis, and Normals with Gaussians. It distinguishes the Beta and Gamma distributions from the beta and gamma functions, and the geometric mean from the mean of a Geometric. For clarity and as a token of appreciation for our named distributions, we capitalize them and encourage others to do likewise.

2

Meaning of Measure

I said that there were never any surprises – that the mathematicians only prove things that are obvious. Topology was not at all obvious to the mathematicians. There were all kinds of weird possibilities that were "counterintuitive." Then I got an idea. I challenged them: "I bet there isn't a single theorem that you can tell me – what the assumptions are and what the theorem is in terms I can understand – where I can't tell you right away whether it's true or false."

It often went like this: they would explain to me, "You've got an orange, OK? Now you cut the orange into a finite number of pieces, put it back together, and it's as big as the sun. True or false?"

"No holes?"

"No holes."

"Impossible! There ain't no such thing."

"Ha! We got him! Everybody gather around! It's So-and-so's theorem of immeasurable measure!"

Just when they think they've got me, I remind them, "But you said an orange: You can't cut the orange peel any thinner than the atoms."

"But we have the condition of continuity: We can keep on cutting!"

"No, you said an orange, so I *assumed* that you meant a *real orange*."

– Richard P. Feynman

2.1 Introduction

Measure theory is the branch of real analysis that studies, in a very general setting, ideas such as length, volume, and mass. Once the notion of measure has been precisely defined, a probability measure is defined to be a measure P on a sample space Ω such that $P(\Omega) = 1$. Thus, some take the point of view that probability is merely a special case of measure theory ("special" in that the whole space has measure 1, an easier situation to handle than spaces of infinite measure).

However, probability has a flavor of its own, motivated by understanding and quantifying *variation*, *uncertainty*, and *information*. The ideas relevant for studying uncertainty and building statistical models (such as dependence, correlation, and conditioning) are often different from the ideas arising naturally in pure measure theory.

For decades, controversy has continued over how much measure theory should be taught in a graduate-level probability course. Two main advantages of introducing measure theory are:

1. *Unified notation and language.* The language of measure theory allows us to prove results for all random variables in a unified way, in contrast to elementary courses (which typically handle discrete and continuous cases separately, and ignore random variables that are neither purely discrete nor purely continuous). The statistical literature often uses this language, so it is useful to be comfortable with, and even friends with, the basic ideas.

   Also, the language of measure theory suggests useful analogies with concepts such as mass and volume. For example, the center of mass in physics corresponds to expected value; the moment of inertia corresponds to variance. Thinking of probability distributions as distributions of mass often gives a helpful mental picture.

2. *Mathematical foundation.* Measure theory puts the study of probability into a firm mathematical framework. Andrey Kolmogorov, who pioneered the measure-theoretic approach to probability, wrote that "The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as geometry and algebra." Done poorly, this can result in spending so much effort on technical details that one misses the key ideas. Done well, this adds clarity and sheds light on various paradoxes that we will encounter later.

On the other hand, it would be easy to get lost in abstraction, or bogged down in proving measurability. In the real world, you will never run into a nonmeasurable set (with respect to the standard *Lebesgue measure*). Measuring instruments have finite precision, so any experiment has only finitely many possible outcomes. Even with infinitely many possible outcomes, it is impossible to construct a nonmeasurable set without resorting to the Axiom of Choice (which means that no explicit formula can be given for one). Stephen Senn quipped that "a theoretical statistician knows all about measure theory but has never seen a measurement, whereas the actual use of measure theory by the applied statistician is a set of measure zero." We would distinguish here between *theoretical statistics* and *mathematical statistics* (and think that Senn meant the latter). Theoretical statistics attempts to capture the essential structure of a real problem, providing useful frameworks, tools, bounds, etc.; the math may (or may not) be easy. For there to be a divide between theory and practice is unfortunate: theory without practice is hollow, while practice without theory is ad hoc.

> Theory and practice are not mutually exclusive; they are intimately connected. They live together and support each other. This has always been the main credo of my professional life. I have always tried to develop theories that shed light on the practical things I do, and I've always tried to do a variety of practical things so that I have a better chance of discovering rich and interesting theories. . . The best theory is inspired by practice. The best practice is inspired by theory.
> – Donald Knuth

In this book we seek a happy medium, introducing some of the main language and concepts of measure theory while avoiding getting bogged down in measurability proofs. This allows time to go into greater depth in other areas of probability. Meanwhile, readers who wish to pursue measure theory at a level beyond the scope of this book can approach the subject with a richer collection of motivating ideas and examples than they otherwise would have.

2.2 σ-Algebras

Start with any nonempty set Ω, the *sample space* (the set of all possible outcomes of some experiment). The crucial idea needed in giving probability a firm mathematical footing is to *interpret events as being subsets of* Ω. Saying that event A occurred is the same as saying that the outcome of the experiment is in the set A. Probability will be defined as a function P which assigns a number between 0 and 1 to each event, and satisfying a couple natural conditions. This correspondence between events and subsets is detailed in the table below.

Probability	**Sets**
sample space	Ω
an outcome ω	$\omega \in \Omega$
event A	$A \subseteq \Omega$
A occurs	$\omega \in A$
A or B	$A \cup B$
A xor B (exclusive or)	$A \bigtriangleup B \equiv (A \cap B^c) \cup (A^c \cap B)$
A and B	$A \cap B$
not A	A^c (complement)
A and B are mutually exclusive	$A \cap B = \emptyset$ (empty set)
A implies B	$A \subseteq B$
probability of A	$P(A)$
A and B are independent	$P(A \cap B) = P(A)P(B)$

What does Ω look like, and how is it specified in practice? In the stereotypical discrete examples (such as flipping coins, rolling dice, or drawing balls from bags) there is not much ambiguity: Ω is just the list of possible outcomes of the "experiment", encoded in some way (e.g., $\{H, T\}$ for a coin flip). When the experiment is more complicated (such as the results of a complicated chemical reaction over time, or the results of an international multistage survey), it may become unwieldy or unclear how to specify Ω.

To some extent, this is mitigated by the fact that random variables are described by distributions over the real line (as discussed below), regardless of how complicated Ω is. Nevertheless, it is conceptually helpful to think about what Ω is for a given problem, as a way to clarify the structure of the problem. In practice, one often receives noisy, imprecise observations as well as missing data.

For example, consider a survey measuring support for a certain political candidate, on a scale of 1 to 5. It is often useful to assume that each person has a continuous underlying "true" level of support, which might be rounded to an integer between 1 and 5: there is a limit to how precisely the measurement can be made. Some people may lie or misunderstand the question, creating noise. Some may refuse to respond to some questions, creating missing data. Conceptually, Ω consists of *all possible* outcomes, with full precision. This includes *both* the "scientific" aspect – people's true beliefs – and the "sampling" aspect – the randomness associated with who is chosen for the survey, and how they respond. If, for example, certain types of people are more likely to refuse to respond, then the relevant variables should be built into Ω and reflected in P.

As an idealization, imagine a videotape documenting in perfect precision every detail that could ever be relevant for the experiment, throughout its time span. Then we may think of Ω as the collection of all possible videotapes that might have been generated, of which some specific one ω is realized. The statistician generally has only partial information about ω, but seeks to make valid probabilistic statements about the unknowns of interest, given what he or she does know. This may involve computing certain random variables – functions of the videotape – to provide numerical summaries of some features of the tape. The researcher may have access only to a grainy version of the tape, and will report estimates accordingly.

Thus, we wish to define the probability of A as $P(A)$, where A is an "event" considered to be a subset of Ω. Are all subsets events, so that we can take the domain of P to be 2^Ω (the collection of all subsets)? If Ω is finite or countably infinite, that approach is viable. For uncountable Ω, such as $\Omega = \mathbb{R}$, technical difficulties arise.

A dramatic illustration of this is the *Banach-Tarski Paradox*, hinted at in Feynman's story about the orange. This states that it is possible to decompose a solid ball (of constant density) into five disjoint pieces, rotate and translate the pieces in space, and then reassemble them to have two copies of the original ball. In addition to sounding intuitively impossible (as well as an alchemist's dream), this seems mathematically impossible as it seems to imply that the mass of the ball is twice its own mass.

The loophole here is that some of the pieces must be *nonmeasurable* (in the Lebesgue sense, defined later), which means that they are so bizarrely complicated that mass is meaningless. Carefully describing which subsets can be assigned probabilities leads to the notion of a *σ-algebra*, defined next.

Definition 2.2.1 A *σ-algebra* on Ω is a collection $\mathcal{F}$ of subsets of Ω such that

1. $\emptyset \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $A_1, A_2, \cdots \in \mathcal{F}$, then $\cup_{j=1}^{\infty} A_j \in \mathcal{F}$.

That is, $\mathcal{F}$ contains $\emptyset$ and is closed under complements and countable unions.

The idea is to start with some simple events that clearly should be measurable, and use them to construct other events that should also be measurable. If it makes sense to talk about the probability of A occurring, it also makes sense to talk about the probability of A *not* occurring; this corresponds to $\mathcal{F}$ being closed under complements. If we are interested in a sequence of events $A_1, A_2, \ldots$, then we may also be interested in knowing whether at least one of the events occurs; this corresponds to $\mathcal{F}$ being closed under countable unions.

One may question whether such a notion is really needed, since in practice one could treat Ω as finite (albeit huge) with no risk of nonmeasurability. However, it is often mathematically more convenient to work with a continuum (e.g., integrals are often easier than sums). Furthermore, a σ-algebra can be thought of as a way to encode *information.* In the videotape analogy, if one observes more data over time and progressively gets a more vibrant, detailed view of the tape, then this can be represented conveniently as an increasing sequence (or indexed collection) of σ-algebras. Such a collection of σ-algebras, indexed by time, is known as a *filtration* (that is, at each time t we have a σ-algebra $\mathcal{F}_t \subseteq \mathcal{F}$, with $\mathcal{F}_t \subseteq \mathcal{F}_{t'}$ for $t \leq t'$). Here $\mathcal{F}_t$ can be thought of as the information available at time t: events therein correspond to questions that can be answered at time t.

Note on terminology. σ-algebras are also known as "σ-fields". For historical reasons, the σ (which suggests summation) indicates that we can take countable unions; an *algebra* satisfies the same properties as a σ-algebra, except that it is only required to be closed under finite unions.

☣ 2.2.2 Be sure to keep clearly in mind where each kind of object "lives". For example, a σ-algebra $\mathcal{F}$ is a collection of sets, and an element of $\mathcal{F}$ is a subset of Ω. An intersection of σ-algebras is very different from an intersection of events.

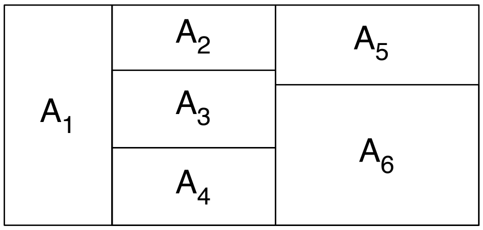


Figure 2.1 A partition of Ω into 6 blocks.

Example 2.2.3 Trivially, the smallest possible σ-algebra is $\{\emptyset, \Omega\}$. More interestingly, chop Ω up into finitely many disjoint pieces $A_1, A_2, \ldots, A_n$ (so $A_i \cap A_j = \emptyset$ for $i \neq j$ and $\cup_{i=1}^n A_i = \Omega$), as illustrated in Figure 2.1 for $n = 6$ (of course, the blocks don't need to be rectangular, and in fact it may be hard to define a notion of "shape" in Ω anyway). This is a *partition* of Ω into *blocks*, and gives rise to a σ-algebra consisting of all possible unions $\cup_{i \in I} A_i$ of A_i's (where I is any subset of $\{1, \ldots, n\}$).

In this example, a natural way to interpret the blocks A_i is in terms of *partial information* about the outcome of the experiment. Specifically, we may know which A_i contains the actual outcome ω, without knowing exactly what ω is. A coarse partition, with a small number of big blocks, corresponds to not knowing much; a refined partition, with many small blocks, corresponds to having detailed information (the extreme case being when each A_i is a singleton).

Remark. In fact, *any* finite σ-algebra is obtained in this way (see below).

✎ 2.2.4 (a) Explain why the σ-algebra described in Example 2.2.3 is indeed a σ-algebra.

(b) Conversely, show that any finite σ-algebra is induced in this way by some partition of Ω. That is, let $\mathcal{F}$ be a σ-algebra of subsets of Ω, containing only finitely many sets. Then there is a partition of Ω that gives rise to $\mathcal{F}$. In particular, conclude that the number of sets in $\mathcal{F}$ is a power of 2.

Hint for (b): for each $\omega \in \Omega$, consider the smallest element of $\mathcal{F}$ containing ω.

✎ 2.2.5 Show that a σ-algebra is automatically closed under countable intersections, i.e., if $A_1, A_2, \ldots$ are in a σ-algebra $\mathcal{F}$, then $\cap_{j=1}^{\infty} A_j \in \mathcal{F}$. Then show that the requirement $\emptyset \in \mathcal{F}$ can be replaced by the requirement $\mathcal{F} \neq \emptyset$.

✎ 2.2.6 Check that an intersection of σ-algebras, even uncountably many, is a σ-algebra. Give a simple example (based on an extreme case) showing that a union of σ-algebras may not be a σ-algebra.

Definition 2.2.7 Given any collection of sets $\{A_i : i \in I\}$, by the above there is a unique smallest σ-algebra containing all the A_i. This is called the σ-algebra *generated by* the A_i, sometimes written as $\sigma(\{A_i : i \in I\})$.

2.3 Borel sets

People often call certain subsets of $\mathbb{R}$ *nice* when wishing to avoid both pathology and technical detail. What niceties make a set nice? To give a precise meaning to niceness, we start with sets so simple and fundamental that their niceness is indisputable: intervals. Then, we use the intervals to generate a σ-algebra, giving the *Borel sets.*

Definition 2.3.1 (Borel sets) The *Borel σ-algebra $\mathcal{B}$* on $\mathbb{R}$ is defined to be the σ-algebra generated by all open intervals (a, b) with $a, b \in \mathbb{R}$. A *Borel set* is a set in the Borel σ-algebra. Analogously, we define the Borel σ-algebra on $\mathbb{R}^2$ to be the σ-algebra generated by open rectangles and, more generally, the Borel σ-algebra on $\mathbb{R}^n$ to be the σ-algebra generated by *open boxes* $B = \{(x_1, \ldots, x_n) : a_1 < x_1 < b_1, \ldots, a_n < x_n < b_n\}$.

✎ 2.3.2 Show that the Borel σ-algebra is the same if we use closed intervals $[a, b]$ instead of open intervals to generate it, and that it is also the same if we use “semi-infinite” intervals $(-\infty, b]$ to generate it. Moreover, show that we can take b to be rational without changing the resulting σ-algebra.

Example 2.3.3 (Some Borel sets) Let us look at a few sets in the Borel σ-algebra $\mathcal{B}$. Start with some intervals, say, $(0, 1), (2, 3), (4, 5), (6, 7), \ldots$. To get a more complicated Borel set, we can take their union: $(0, 1) \cup (2, 3) \cup (4, 5) \ldots$. We can then take the complement of this, which is $(-\infty, 0] \cup [1, 2] \cup [3, 4] \cup \ldots$. But of course we picked the endpoints here just for the sake of example, and in the same way we could take countably many sets of this type, then take their union, take complements, etc. It is clear that we would soon have dizzyingly complicated Borel sets, and it is difficult to explicitly write down a subset of $\mathbb{R}$ which is *not* Borel.

To go into more detail on how complicated a Borel set can be, first note that any open set S is in $\mathcal{B}$ (a set S is *open* in $\mathbb{R}$ if for each point x in the set, there is an open interval containing x and contained within S). To see this, let I_x be an open interval with $x \in I_x \subseteq S$ for each x, so that S is the union of the I_x. This may be a union of uncountably many intervals though, whereas we only know that *countable* unions of open intervals are in $\mathcal{B}$. To get around this, use the fact that the rational numbers are *dense*: for any real number x and any positive number ϵ, we can always find a rational number r with $|x - r| < \epsilon$. So the intervals I_x can be chosen to have *rational endpoints*, and there are only countably many such intervals.

So all open sets are Borel, and it follows that all closed sets are Borel (a closed set is the complement of an open set). But since we are working in a σ-algebra, we are free to take complements, countable unions, and countable intersections. Using these operations iteratively, we can build more and more complicated sets.

More precisely, the Borel sets can be constructed as follows.

1. Let $\mathcal{G}_0$ be the set of all open sets in $\mathbb{R}$.
2. Let $\mathcal{F}_0$ be the set of all closed sets in $\mathbb{R}$.
3. For any integer $n \geq 0$, let $\mathcal{G}_{n+1}$ be the set of all countable unions of sets from $\mathcal{F}_n$, and $\mathcal{F}_{n+1}$ be the set of all countable intersections of sets from $\mathcal{G}_n$.

After even a few steps this process exhausts the human mind's ability to visualize such a set, but it does *not* exhaust the Borel sets: every set in $\mathcal{G}_n$ or $\mathcal{F}_n$ is Borel, but there are Borel sets that are not in any of the $\mathcal{G}_n$'s or $\mathcal{F}_n$'s. We can continue the process by letting

$$\mathcal{G}_\gamma = \bigcup_{n=1}^{\infty} \mathcal{G}_n, \text{ and } \mathcal{F}_\gamma = \bigcup_{n=1}^{\infty} \mathcal{F}_n.$$

This *still* does not exhaust the Borel sets: we get new Borel sets by letting $\mathcal{G}_{\gamma+1}$ be the set of all countable unions of sets from $\mathcal{F}_\gamma$, $\mathcal{F}_{\gamma+1}$ be the set of all countable intersections of sets from $\mathcal{G}_\gamma$, and so on. If you are familiar with the notion of *ordinal numbers* (a concept not needed for the rest of this book), it turns out that $\mathcal{B}$ is the union of $\mathcal{G}_\alpha$ over all countable ordinal numbers α.

Despite the complexity of the hierarchy described above, it turns out that $\mathcal{B}$ and $\mathbb{R}$ have the same cardinality! So we have a very rich system of subsets of $\mathbb{R}$, while maintaining some control over the complexity of the sets. The Borel sets will be our default choice for a σ-algebra on $\mathbb{R}$, and if we say a set of real numbers is measurable then we mean it is Borel, unless otherwise specified. It is possible but difficult to explicitly write down a non-Borel set, and such sets are concocted for the sake of pathology, not for the sake of statistics.

In addition to measurable *sets*, we will need the notion of a measurable *function.*

Definition 2.3.4 A function $g : \mathbb{R} \to \mathbb{R}$ is called *Borel-measurable* if the preimage $g^{-1}(B) \equiv \{x : g(x) \in B\}$ is Borel for every Borel set B. We may abbreviate "Borel-measurable" to "measurable", or even abbreviate "measurable" to "", when the meaning is clear from the context.

2.4 Lebesgue Measure

The most commonly used measure on $\mathbb{R}$ is known as *Lebesgue measure*. We will not go through the details of the construction of this measure here, but will just mention a few basic facts about it: Lebesgue measure m is an extension of the notion of length for an interval, so any interval (a, b) (with $a < b$) is Lebesgue-measurable with $m((a, b)) = b - a$. Naturally, for a countable union of disjoint intervals, the Lebesgue measure is the total length: $m(\cup_{j=1}^{\infty}(a_j, b_j)) = \sum_{j=1}^{\infty}(b_j - a_j)$, if the intervals (a_j, b_j) don't overlap. Also, Lebesgue measure has the intuitively appealing property of being *translation invariant*: if A is Lebesgue-measurable and $A_c \equiv \{x + c : x \in A\}$ is a shifted version of A, then $m(A_c) = m(A)$. Similarly, there is a notion of Lebesgue measure in $\mathbb{R}^n$ for any n.

It is technically convenient to define Lebesgue measure on a certain σ-algebra which is larger than $\mathcal{B}$, so that "completeness" holds: if $m(A) = 0$, then $m(A_0) = 0$ for any subset A_0 of A. It is not only difficult but *impossible* to explicitly write down a non-Lebesgue-measurable set.

Any Lebesgue-measurable set is "almost" a Borel set, in the following sense: if A is Lebesgue-measurable, then there is a Borel set B such that the symmetric difference $A \triangle B$ has Lebesgue measure 0 (so A and B "agree" on which points to include, except for a set of measure 0). This further supports working with Borel sets for any practical purpose, though sometimes it is technically convenient to have the completeness property of the Lebesgue measurable sets.

2.5 Axioms of Probability

How dare we speak of the laws of chance? Is not chance the antithesis of all law? – Bertrand Russell

The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as geometry and algebra. – Andrey Kolmogorov

A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ with Ω a sample space, $\mathcal{F}$ a σ-algebra on Ω, and P a *probability measure* (i.e., a function satisfying the axioms below) defined on $\mathcal{F}$.

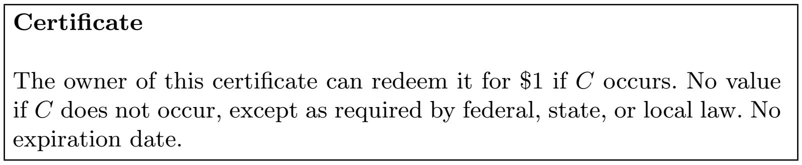
A probability measure P is a function on $\mathcal{F}$, taking values between 0 and 1 (although some physicists have attempted to make sense of negative probabilities, while trying to make sense of quantum mechanics). By convention, impossible events have probability 0 and events that are certain to occur have probability 1. So the first axiom of probability is that:

$$P(\emptyset) = 0, P(\Omega) = 1.$$

Amazingly, only one other axiom is needed for probability:

$$P(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j), \text{ if the } A_j \text{ are disjoint events.}$$

This condition, *countable additivity*, says that if $A_1, A_2, \ldots$ don't overlap each other, then the probability that at least one of them happens is the sum of their probabilities.



Certificate

The owner of this certificate can redeem it for \$1 if C occurs. No value if C does not occur, except as required by federal, state, or local law. No expiration date.

Figure 2.2 Certificate used in a Dutch Book argument.

✎ 2.5.1 Show that in the first axiom, the condition $P(\emptyset) = 0$ can be eliminated. (We include it anyway so that the extremes of $\emptyset$ and Ω are considered together.)

The axioms of probability are natural extensions of the notions of area, volume, and mass. An alternative justification is through what is known as a *Dutch book argument.* The idea is that if probabilities are interpreted as personal degrees of belief, and if a person's probabilities violate the axioms, then we can attempt to construct a sequence of bets which would individually seem fair to that person, yet which together would *guarantee* that the person would lose money (this is also known as an *arbitrage opportunity*). For example, consider a certificate of the form shown in Figure 2.2, which we will call a *C-certificate.*

Assume that someone named Arby is willing to buy or sell C-certificates at a price of $P_{\text{Arby}}(C)$, where P_{Arby} is Arby's personal probability function, for any event C (this is the "fair price" in the sense of making the expected value 0; expected values are formally introduced in Chapter 4). Suppose that there are disjoint events A and B such that $P_{\text{Arby}}(A \cup B) < P_{\text{Arby}}(A) + P_{\text{Arby}}(B)$.Then Arby is willing to pay $P_{\text{Arby}}(A) + P_{\text{Arby}}(B)$ to buy an A-certificate and a B-certificate, and is willing to sell an $(A \cup B)$-certificate for $P_{\text{Arby}}(A \cup B)$. In those transactions, Arby loses $P_{\text{Arby}}(A) + P_{\text{Arby}}(B) - P_{\text{Arby}}(A \cup B)$, and will not recoup any of that loss when it is known whether A occurred and whether B occurred, since A and B are disjoint.

✎ 2.5.2 Give an analogous Dutch Book argument for the case that there are events A, B (not necessarily disjoint) with $P_{\text{Arby}}(A \cup B) > P_{\text{Arby}}(A) + P_{\text{Arby}}(B)$. Also give Dutch Book arguments showing what would happen if P_{Arby} ever fails to be in $[0, 1]$.

Such arguments show that the axioms of probability are not arbitrary; rather, they are essential in order to have coherent degrees of belief (at least with finite additivity rather than countable additivity). Countable additivity vs. finite additivity, and the general question of how important it is to be coherent in this sense remain somewhat controversial questions. In this book, we are happy to accept countable additivity as a natural and convenient form of continuity; Section 2.8 defines a precise sense in which probability is continuous.

There are deep philosophical questions about the meaning of probability. To some extent these questions can be bypassed by taking a measure theoretic point of view (probability is anything that satisfies the axioms of probability), but ultimately we wish to use probability as a tool for learning about the real world, and then philosophical and foundational questions can be important for interpreting and choosing statistical models and methods.

Example 2.5.3 Let $\Omega = \{a_0, a_1, a_2, \dots\}$ be countable, and let $p_0, p_1, p_2, \dots$ be nonnegative numbers summing to 1. If we interpret the p_i as the probabilities of the outcomes a_i, it is natural to define

$$P(A) = \sum_{i:a_i \in A} p_i$$

for all $A \subseteq \Omega$. The sum is well-defined because the p_i are nonnegative and the sum is bounded above by 1, and here we can safely take $\mathcal{F}$ to consist of all subsets of Ω. A different choice of p_i's would yield a different probability measure on the same sample space. This construction does not work for uncountable sample spaces such as the real line, because for an uncountable sum of nonnegative numbers to converge, all but countably many of the terms must be 0 (reducing the problem back to the discrete case).

Note the drastic simplification that the above gives, compared with having to specify $P(A)$ explicitly for all A! For example, if Ω is finite with 100 elements, then we need to specify 99 probabilities $p_0, \dots, p_{98}$, rather than writing down probabilities for all $2^{100} \approx 10^{30}$ subsets of Ω, and making sure that they are compatible with each other!

2.6 Random Variables

Random variables appear throughout probability (and this book). Indeed, the main subject matter is *random variables and their distributions*. The formal definition of a random variable is simultaneously simple and subtle.

Definition 2.6.1 A *random variable* is a measurable function X from Ω into $\mathbb{R}$, where "measurable" means that the preimage $X^{-1}(B) \equiv \{\omega \in \Omega : X(\omega) \in B\}$ is in $\mathcal{F}$ for all Borel sets B. Note that typically X^{-1} does not exist in the sense of an invertible function; rather, X^{-1} is the preimage (also known as the *inverse image*), and is defined on *sets* of real numbers. We often abbreviate "random variable" to "r.v."; in fact, random variables are so central to everything discussed here that even a two-letter abbreviation doesn't seem short enough!

The measurability condition is in place since we wish to ask questions like "What is the probability that X is in B?" and thus we need $P(\{\omega : X(\omega) \in B\})$ to be defined. Note that this definition is relative to $\mathcal{F}$: with respect to a smaller σ-algebra in place of $\mathcal{F}$, X may lose its status as a random variable. We will generally assume that $\mathcal{F}$ is rich enough that the functions we are interested in will indeed be random variables.

✎ 2.6.2 Show that if X is a random variable, then so is $g(X)$ for any measurable function g.

Remark 2.6.3 We usually use capital letters such as X, Y, Z to denote random variables, and corresponding lowercase letters x, y, z for the corresponding values. This is a convenient mnemonic, linking a r.v. to its values while maintaining the distinction, but it should not be followed fanatically since notation is sometimes much simpler using one letter for both a r.v. and its values (when it is obvious from the context which is meant). If the r.v. is denoted by a Greek letter (which is extremely common in Bayesian statistics), it is even more convenient to use the same letter for both a r.v. and its values.

As for which letters to use, an old joke asks "How can you tell a probabilist from a statistician?" and answers "The probabilist uses X for a random variable, while the statistician uses Y." We will often use both, in keeping with the spirit of reconciliation.

The definition of a random variable may seem mysterious at first: if X is a deterministic function, mapping Ω to $\mathbb{R}$, then where does the randomness come from? It seems that X is just a function, and neither random nor a variable!

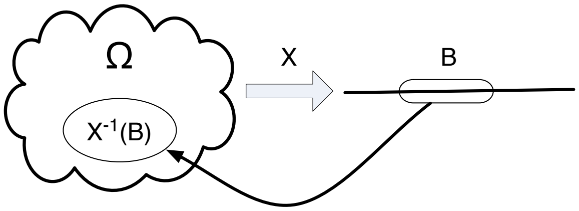


Figure 2.3 An r.v. maps Ω into the real line. This induces a probability measure on the real line, using preimages to "carry" a set B back to Ω.

A closer look resolves this objection. We imagine an outcome $\omega \in \Omega$ of a random experiment, where P governs the randomness. Then $X(\omega)$ is a chosen *numerical* summary of some aspect of the experiment. The fact that X is real-valued often allows us to work with the real line rather than explicitly having to deal with Ω, which may be an incredibly complicated space. That is, any random variable *induces* a probability measure on $\mathbb{R}$: to find the measure of a Borel set B in $\mathbb{R}$, we "transport" B back to Ω by finding the preimage, and then compute the probability of that preimage.

Definition 2.6.4 The *distribution* (also known as the *law*) of a random variable X is the probability measure it induces on $\mathbb{R}$, given by $P_X(B) \equiv P(X \in B)$ for all Borel sets B. We denote this measure by $\mathcal{L}(X)$. The *cumulative distribution function (CDF)* of X is obtained by taking B to be of the form $(-\infty, x]$: take $F(x) \equiv P(X \leq x)$.[1]

Part of the beauty of distributions comes from the fact that two statisticians, who are in different parts of the world working on completely different problems with completely different probability spaces, may both find the same distribution extremely useful; they can even discuss this distribution together in the common language of probability measures on $\mathbb{R}$, without worrying about whether they have completely different Ω's.

The remaining chapters are devoted in large part to the study of random variables and their distributions, so for now we give only a few quick examples.

Example 2.6.5 Let X be the height of a randomly chosen person in some country. View

[1] Other conventions are possible, such as working with open intervals $(-\infty, x)$ and the left continuous function $P(X < x)$ instead of the right continuous function $F(x) = P(X \leq x)$.

Ω as the set of all people in the country (or as a set of databases or "videotapes", with one for each individual). Then it makes sense to think of X as a function: each person has a height. Note that the randomness comes from the selection of a person, not from the mapping X itself.

Example 2.6.6 (Uniform distribution) Let $\Omega = [0, 1]$, with $\mathcal{F}$ the Borel σ-algebra. Define the probability of an interval to be its length. This extends uniquely to a measure on $\mathcal{F}$ (which in fact is the earlier-alluded-to Lebesgue measure, restricted to $[0, 1]$). Let U be the identity function, i.e., $U(\omega) = \omega$ for all $\omega \in [0, 1]$.

Then U has the *Uniform* distribution on $[0, 1]$. It may seem convoluted to use the identity function here, for something so natural as choosing a random number between 0 and 1. This is possible here since the elements of Ω are numerical, and the choice of measure on Ω automatically ensured that intervals of equal length would have equal probability. Writing U in this way maintains the distinction between the raw outcome of the experiment and the numerical summary provided by a r.v. Note that many other r.v.s can be defined on the same Ω, such as U^2 (or any other (measurable) function of U).

Example 2.6.7 (Indicator Random Variables) A simple but extremely useful type of random variable is an *indicator random variable* I_A, defined by letting $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ otherwise, where A is some fixed event. Such an r.v. has a *Bernoulli distribution*, with parameter $p = P(A)$. The CDF is then a step function, jumping up from 0 to $1 - p$ at $x = 0$ and from $1 - p$ to 1 at $x = 1$. Indicator r.v.s are a key ingredient in defining expected values, and provide a critical link between averages and probabilities. Note that I_A to any positive power is I_A, $I_{A^c} = 1 - I_A$, and $I_A I_B = I_{A \cap B}$ for any events A, B; such facts make indicators extremely convenient tools.

It turns out that since intervals of the form $(-\infty, x)$ generate all the Borel sets, specifying the CDF for all x determines the entire distribution (this follows from the $\pi - \lambda$ Theorem, which is discussed in Section 2.10). This is a great simplification: instead of dealing with a complicated Ω, or having to figure out $P(X \in B)$ for all Borel B, we merely need specify a CDF F.

Proposition 2.6.8 *A function $F : \mathbb{R} \to [0, 1]$ is a CDF iff F is increasing, right continuous, and $F(x) \to 0$ as $x \to -\infty$, $F(x) \to 1$ as $x \to \infty$.*

That any CDF satisfies these properties follows from continuity of probability. The converse is more interesting: given any function F satisfying these properties, we can construct a random variable whose CDF is F. The beautiful idea for doing so is the *probability integral transform*, which enables one to obtain *any* distribution starting from a Uniformly distributed random variable. This is carried out in detail in Chapter 3.

Definition 2.6.9 If F is a CDF, we write $X \sim F$ to indicate that X is a r.v. with CDF F. Extending this notation, we also write $X_1 \sim X_2$ to indicate that X_1 and X_2 have the same distribution.

✎ 2.6.10 Give an example of two random variables X_1 and X_2 on the same space, with the same distribution, such that X_1 *never* equals X_2.

✎ 2.6.11 Find three random variables X_1, X_2, X_3 on the same space such that $P(X_1 > X_2) \geq 0.5, P(X_2 > X_3) \geq 0.5,$ and $P(X_3 > X_1) \geq 0.5$. (Such intransitivity can be counterintuitive at first. A famous example is *Efron's dice*, which consists of 4 dice labeled so that there is a cycle where each die beats the next with probability exactly 2/3.)

Any random variable has a corresponding σ-algebra, encoding what it means to know the random variable.

Definition 2.6.12 For any r.v. X, define $\sigma(X)$ to be the smallest $\sigma-$algebra on Ω containing the events $\{X \in B\}$ for all Borel sets B. Here, $\{X \in B\} \equiv X^{-1}(B) \equiv \{\omega \in \Omega : X(\omega) \in B\}$ is the preimage of B. Intuitively, this σ-algebra corresponds to knowing and being able to answer questions about X. For a collection of r.v.s $X_j, j \in J$, we define $\sigma(X_j, j \in J)$ to be the smallest σ-algebra containing all of the $\sigma(X_j)$'s.

2.7 Random Vectors

A random elephant is a function from Ω into a suitable space of elephants. – John Kingman

Analogously to the definition of random variable, we define a random vector in $\mathbb{R}^n$ as follows.

Definition 2.7.1 (Random Vectors) A *random vector* in $\mathbb{R}^n$ is a measurable function $\mathbf{X}$ from Ω into $\mathbb{R}^n$, where "measurable" means that the preimage $\mathbf{X}^{-1}(B) \equiv \{\omega \in \Omega : \mathbf{X}(\omega) \in B\}$ is in $\mathcal{F}$ for all Borel sets B in $\mathbb{R}^n$. We will often denote random vectors with bold capital letters. The *distribution* of $\mathbf{X}$ is the function $P(\mathbf{X} \in B)$, as a function of Borel $B \subseteq \mathbb{R}^n$.

Many ideas and results about random variables extend easily to random vectors. For simplicity, we will sometimes state results in terms of random variables even when we could have stated similar results for random vectors. It is often useful to think about whether certain results for random variables can be generalized to random vectors. More generally still, we can define other kinds of random objects by considering mappings from Ω to a space of possible objects, as long as there is a notion of which sets of these objects are measurable.

We define the *joint distribution* and *joint CDF* of a collection of random variables analogously to the 1-dimensional case. For example, the joint CDF of two random variables X_1, X_2 is the function $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$, defined on all of $\mathbb{R}^2$.

✎ 2.7.2 Let $F(x, y)$ be a bivariate CDF (i.e., the joint CDF of two r.v.s). Prove that for all $a_1 \leq b_1$ and $a_2 \leq b_2$,

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0,$$

and explain why this is called the *rectangle inequality.*

It follows from the $\pi - \lambda$ Theorem that the joint CDF of a random vector determines the joint distribution completely.

2.8 Limits of Events

Recall that a function is *continuous* if it preserves limits: if $x_n \to x_\infty$ as $n \to \infty$, then $f(x_n) \to f(x_\infty)$ as $n \to \infty$. This is particularly useful for approximations, as by making x' close to x we can ensure that $f(x')$ is close to $f(x)$. We would like to extend this idea to probabilities: if two events are "close," then their probabilities should be close; but what does it mean for a sequence of events to have a limit?

To answer this, let's start with the simplest kind of sequence of events: a nested increasing sequence $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$. This sequence ensnares more and more points, expanding towards the union $\cup_{j=1}^{\infty} A_j$. So here we define the limit to be the union.

Dually, we can take the complement of each set to obtain a decreasing sequence $A_1^c \supseteq A_2^c \supseteq A_3^c \supseteq \dots$, for which we define the limit to be the intersection of the A_j^c.

In the case of the increasing sequence, we envision more and more points being added; once a point is added, it stays in the set forever. In contrast, for the decreasing sequence, we envision more and more points being jettisoned; once a point is jettisoned, it is outside of the set forever.

Definition 2.8.1 (Limits of sequences of events) For a general sequence of events $A_1, A_2, \dots$, we say that the limit exists if for each $\omega \in \Omega$, the sequence eventually *makes up its mind* whether or not to include ω. That is, there is a number $n(\omega)$ such that either $\omega \in A_n$ for all $n \geq n(\omega)$, or $\omega \notin A_n$ for all $n \geq n(\omega)$. If the limit exists, the limit consists of all ω that the sequence decides in favor of, and we denote it by $\lim_{n\to\infty} A_n$.

✎ 2.8.2 Check that the above definition is consistent with the definition of limits for increasing and decreasing sequences. Give an example of a sequence A_n of sets that converges, and is neither eventually increasing (from some stage onward) nor eventually decreasing.

✎ 2.8.3 Let $A_n = \{1/n\}$. Does the sequence of sets $A_1, A_2, \dots$ converge to a limit and if so, what is the limiting set?

Limits of sequences of events can be described equivalently in terms of the corresponding indicator r.v.s.

Proposition 2.8.4 *Let $A_1, A_2, \dots$ be events, and let I_A be the indicator r.v. for any event A. Then $\lim_{n\to\infty} A_n$ exists iff $\lim_{n\to\infty} I_{A_n}$ exists pointwise (i.e., for each $\omega \in \Omega$, $I_{A_n}(\omega)$ converges to some value, which must of course be 0 or 1). If they do exist, then*

$$I_{\lim_{n\to\infty} A_n} = \lim_{n\to\infty} I_{A_n}.$$

✎ 2.8.5 Prove the above proposition.

The above proposition is an example of the extremely close relationship between events and their indicators (which we call the *fundamental bridge*, as discussed in Chapter 4). Bruno de Finetti and others have argued for taking this a step further, by using the *same* notation for an event and its indicator r.v. Pollard [11] elegantly advocates this approach, pointing out that the above proposition becomes $\lim_{n\to\infty} A_n = \lim_{n\to\infty} A_n$ in de Finetti notation, "a fact that is quite easy to remember".

We can also define limsup and liminf of sequences of events, in a manner analogous to how limsup and liminf of sequences of numbers are defined.

Definition 2.8.6 (Limsup and liminf of events) The limsup of events $A_1, A_2, \ldots$ is defined as

$$\limsup_{n\to\infty} A_n \equiv \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k;$$

in words, the limsup of the A_n is the event "infinitely many of the A_n occur." The liminf is defined as

$$\liminf_{n\to\infty} A_n \equiv \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k;$$

in words, the liminf is the event "from some stage onward, *all* of the A_n occur."

✎ 2.8.7 Show that the limit of A_n exists and equals A if and only if $\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = A$.

☣ 2.8.8 This definition may look cryptic at first, but it expresses a simple idea. Think of $\cap$ as corresponding to "for all" and $\cup$ as corresponding to "there exists", keeping in mind that order is very important in such expressions (saying that for any y we can find x to solve $x + y = 5$ is quite different from saying we can find an x that solves $x + y = 5$ for all y).

To show continuity of probability, we first need a handy trick for obtaining disjoint events from any sequence of events.

✎ 2.8.9 (Disjointification) Let $A_1, A_2, \ldots$ be events. Construct *disjoint* events $B_1, B_2, \ldots$ such that $\cup_{k=1}^{n} A_k = \cup_{k=1}^{n} B_k$ for all n.

Using the above, we can easily show continuity of probability for an increasing sequence of events.

Lemma 2.8.10 *Let $A_1 \subseteq A_2 \subseteq A_3 \ldots$ be events. Then*

$$P(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

Proof Let $B_1, B_2, \ldots$ be disjoint as in ✎ 2.8.9. Then

$$P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(B_n) = \lim_{n\to\infty} \sum_{k=1}^{n} P(B_k) = \lim_{n\to\infty} P(\cup_{k=1}^{n} B_k) = \lim_{n\to\infty} P(A_n).$$

□

This gives inequalities relating the liminf of events to the liminf of probabilities:

Theorem 2.8.11 *For any sequence of events $A_1, A_2, \ldots,$*

$$P(\liminf_{n\to\infty} A_n) \le \liminf_{n\to\infty} P(A_n) \le \limsup_{n\to\infty} P(A_n) \le P(\limsup_{n\to\infty} A_n).$$

Proof The first inequality follows from the fact that $\tilde{A}_n \equiv \cap_{k=n}^{\infty} A_k$ is an increasing sequence of events, the second is true for all sequences of numbers, and the third is dual to the first. □

Corollary 2.8.12 *If $A_1, A_2, \ldots$ is a convergent sequence of events, then*

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n).$$

This follows immediately from Theorem 2.8.11 since the lefthand side and righthand side of the inequality are equal. Thus, probability is continuous.

2.9 Independence

Much of statistical modeling deals with the question of when events (or random variables) are associated with each other, and when they are independent. Most models crucially depend on certain independence or conditional independence assumptions. Independence is a very natural notion in probability and a rather unnatural notion in measure theory, which is one reason why probability has its own character rather than being "just" measure theory. We now introduce independence, starting with the usual definition of independence of two events A, B.

Definition 2.9.1 Two events A, B are *independent* if

$$P(A \cap B) = P(A)P(B).$$

☣ 2.9.2 This definition is relative to P, so if several different probability measures are being considered then it needs to be specified which probability the events are independent with respect to.

☣ 2.9.3 A common mistake, especially at the beginning level, is to confuse independence and disjointness. These concepts could hardly be further apart! If A and B are disjoint, then B occurring automatically implies that A did not occur, whereas independence would say that knowing that B occurred gives no information about whether A occurred. Note that the only way that A and B can be both independent and disjoint is if $P(A) = 0$ or $P(B) = 0$.

✎ 2.9.4 Show that if A and B are independent, then the complements A^c and B^c are independent.

Next we define independence for any number of events, even infinitely many.

Definition 2.9.5 (Independence of Events) Similarly, we define independence of any finite number of events $A_1, \ldots, A_n$ by

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

for any distinct indices $i_1, \ldots, i_k$. We also would like to define independence of infinitely many events, and of σ-algebras. Does this require getting embroiled in infinite products?

Fortunately, matters are simpler than that. The events in an arbitrary collection are *independent* if the events in any selection of *finitely many* of the events are independent. (Restricting to finite subsets avoids infinite products of probabilities.)

✎ 2.9.6 Give a simple example, e.g., based on 3 coin flips, of three events that are pairwise independent but not independent.

✎ 2.9.7 Give a simple example to show that $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ is *not* sufficient to imply that A_1, A_2, A_3 are independent.

Having defined independence of events, we can easily define independence of r.v.s.

Definition 2.9.8 (Independence of Random Variables) Two random variables X_1, X_2 are *independent* if the event $X_1 \in B_1$ is independent of the event $X_2 \in B_2$ for all Borel B_1, B_2. We denote this by $X_1 \perp\!\!\!\perp X_2$, by analogy with the orthogonality symbol. Similarly, r.v.s $X_j, j \in J$ (where J is some index set) are independent if for any integer $n \geq 1$, indices $j_1, \dots, j_n \in J$, and Borel sets $B_1, \dots, B_n$, the events $X_{j_1} \in B_1, \dots, X_{j_n} \in B_n$ are independent.

✎ 2.9.9 Show that events $A_1, A_2, \dots, A_n$ are independent if and only if their indicator r.v.s are independent.

Fortunately, there is a much simpler way to describe independence of r.v.s, using their joint CDFs. We state this here for any finite list of r.v.s, but for infinitely many we can use the joint CDFs for each finite subset, as with events.

Proposition 2.9.10 *Let $X_1, \dots, X_n$ be random variables, with F_j the CDF of X_j. The random variables are independent iff the joint CDF factors as*

$$F(x_1, \dots, x_n) = F_1(x_1) \dots F_n(x_n),$$

for all real $x_1, \dots, x_n$.

The above result follows from the $\pi - \lambda$ Theorem (see Section 2.10). Essentially, it is not necessary to check that the joint distribution factors for all choices of Borel sets; rather, showing the factorization for a rich enough collection suffices, and intervals $(-\infty, x]$ serve this purpose.

A lemma is a path. Remember a dilemma? That's two paths, presumably hard to choose between.
– Jay Kadane

In later chapters, the following independence lemma will be a useful tool.

Lemma 2.9.11 *For random variables X, Y and measurable functions g, h, suppose that $X \perp\!\!\!\perp Y$ and that $g(X) \perp\!\!\!\perp h(X)$. Then, the random variables $\{g(X), h(X), Y\}$ are fully independent, that is,*

$$P(g(X) \in A, h(X) \in B, Y \in C) = P(g(X) \in A)P(h(X) \in B)P(Y \in C)$$

for all $A, B, C \in \mathcal{B}$. This result also holds for random vectors $\mathbf{X} = (X_1, \dots, X_k)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$, with essentially the same proof.

Note that, in contrast to showing independence of events, here it follows at once that $\{g(X), h(X), Y\}$ are pairwise independent, since we may take any of A, B, or C equal to $\mathbb{R}$ in the above.

Proof Using preimages ($f^{-1}(B) \equiv \{a : f(a) \in B\}$), we have

$$\begin{aligned} P(g(X) \in A, h(X) \in B, Y \in C) &= P(X \in g^{-1}(A) \cap h^{-1}(B), Y \in C) \\ &= P(X \in g^{-1}(A) \cap h^{-1}(B))P(Y \in C) \\ &= P(g(X) \in A, h(X) \in B)P(Y \in C) \\ &= P(g(X) \in A)P(h(X) \in B)P(Y \in C). \end{aligned}$$

□

Remark 2.9.12 The above proof is a good example illustrating that it can be helpful and simple to think directly about distributions and preimages: no unnecessary mess is introduced. In contrast, it would be a nightmare to prove the above result using CDFs, especially as g and h need not be monotone or invertible. Preimages are extremely "respectful" of set operations: if f is a function from a set S into a set T, and $B_\alpha \subseteq T$ for all α in some index set A, then we have the following convenient properties.

1. $f^{-1}(\cup_\alpha B_\alpha) = \cup_\alpha f^{-1}(B_\alpha)$;

2. $f^{-1}(\cap_\alpha B_\alpha) = \cap_\alpha f^{-1}(B_\alpha)$;

3. $f^{-1}(B_\alpha^C) = (f^{-1}(B_\alpha))^C$.

✎ 2.9.13 Verify the above properties of preimages.

The following is a closely related independence lemma, which has a nice sequential interpretation in terms of observing random vectors one at a time, with each new random vector independent of all the previous ones.

Lemma 2.9.14 *If* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ *are random vectors (possibly in different dimensions) with* $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$ *and* $(\mathbf{X}_1, \mathbf{X}_2) \perp\!\!\!\perp \mathbf{X}_3$*, then* $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ *are independent.*

✎ 2.9.15 Prove Lemma 2.9.14.

2.10 Uniqueness and $\pi - \lambda$

In this section, we give the $\pi - \lambda$ Theorem, a tool for proving uniqueness results in probability. In other words, it gives conditions under which we can specify P on some simple events, without fearing the ambiguity of having many possible ways to extend the definition to all of $\mathcal{F}$. This section only treats *uniqueness*; another issue is the need for an *extension theorem* to ensure that there is a way to extend the definition of P to all of $\mathcal{F}$.

A collection $\mathcal{A}$ of subsets of Ω is called a *π-system* if $A \cap B \in \mathcal{A}$ for all $A, B \in \mathcal{A}$

(closure under finite intersections). For example, the set of all open intervals (a, b) in $\mathbb{R}$ is a π-system.

The collection $\mathcal{L}$ of subsets of Ω is called a λ-*system* if the following conditions hold:

1. $\Omega \in \mathcal{L}$;
2. if $A, B \in \mathcal{L}$ with $A \subseteq B$, then $B \setminus A \in \mathcal{L}$ (closure under proper differences);
3. if $A_1 \subseteq A_2 \subseteq \ldots$ are in $\mathcal{L}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{L}$ (closure under increasing unions).

Remark 2.10.1 The "π" in π-system is meant to suggest intersections; more cryptically, the "λ" is apparently meant to suggest limits: a λ-system is closed under limits of increasing sequences. One might picture a λ-system by thinking of a sequence of growing concentric discs: both the union of the discs, and the "rings" formed from differences of the discs, must be in the system. A mnemonic for remembering the 3 properties of a λ-system is lot, lop, limit ("lot" refers to Ω, the full set; "lop" refers to chopping off a proper subset; "limit" refers to the increasing limits).

✎ 2.10.2 Show that if $\mathcal{A}$ is both a π-system and a λ-system, then it is a σ-algebra.

Theorem 2.10.3 (Dynkin $\pi - \lambda$) *If $\mathcal{S}$ is a π-system contained in a λ-system $\mathcal{L}$, then $\sigma(\mathcal{S})$ is also contained in $\mathcal{L}$.*

Proof WELoG (Without Essential Loss of Generality), take $\mathcal{L}$ to be the smallest λ-system containing $\mathcal{S}$. It then turns out that $\mathcal{L}$ is a σ-algebra. To prove this, we just need to show that $\mathcal{L}$ is a π-system. We do this by bootstrapping ourselves up in stages: first show that if $A, B \in \mathcal{S}$, then $A \cap B \in \mathcal{S}$ (this is immediate!); then show that if $A \in \mathcal{S}, B \in \mathcal{L}$, then $A \cap B \in \mathcal{L}$; and then show that if $A \in \mathcal{L}, B \in \mathcal{L}$, then $A \cap B \in \mathcal{L}$.

For any $A_0 \in \mathcal{L}$, define

$$\mathcal{L}(A_0) \equiv \{B \in \mathcal{L} : A_0 \cap B \in \mathcal{L}\}.$$

✎ 2.10.4 Check that $\mathcal{L}(A_0)$ is a λ-system.

Also, if $A_0 \in \mathcal{S}$, we have that $\mathcal{L}(A_0)$ contains $\mathcal{S}$. So $\mathcal{L}(A_0) = \mathcal{L}$ for $A_0 \in \mathcal{S}$.

It follows that $A_0 \cap B \in \mathcal{L}$ for all $A_0 \in \mathcal{S}, B \in \mathcal{L}$. But this implies that $\mathcal{L}(A_0) = \mathcal{L}$ for all $A_0 \in \mathcal{L}$, not just for $A_0 \in \mathcal{S}$! (This is because $\mathcal{L}(A_0)$ is a λ-system containing $\mathcal{S}$, and $\mathcal{L}$ is the smallest such.) So $\mathcal{L}$ is a π-system and a λ-system, and hence is a σ-algebra. We then have $\sigma(\mathcal{S}) = \mathcal{L}$. □

By far the most important application of the $\pi - \lambda$ Theorem for our purposes is through the following uniqueness result.

Corollary 2.10.5 *Let $\mathcal{S}$ be a π-system and let P and Q be probability measures on $\sigma(\mathcal{S})$ with $P(A) = Q(A)$ for all $A \in \mathcal{S}$. Then $P = Q$.*

Proof Note that the set of events on which P and Q agree is a λ-system. □

Example 2.10.6 To see what is going on in a finite case, consider two overlapping events A and B in a sample space Ω, and let P and Q be probabilities on the σ-algebra generated by A and B. Suppose that $P(A) = Q(A)$ and $P(B) = Q(B)$. Does it follow that $P = Q$?

No, since so far no information has been given about how P shares its mass between $A \cap B$ and $A \cap B^c$, and likewise for Q. For example, Figure 2.4 illustrates that we could have

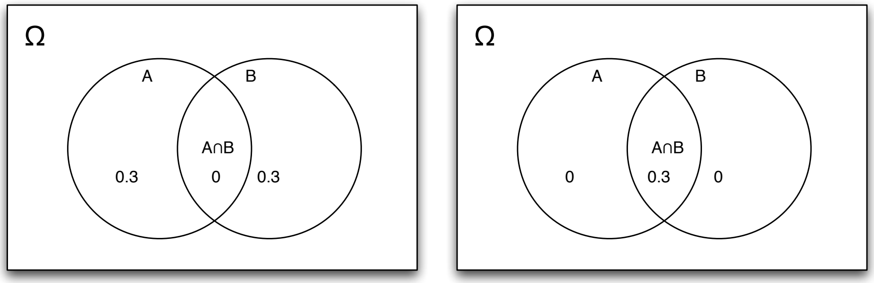


Figure 2.4 A finite illustration of why a π-system is desired for $\pi - \lambda$; the numbers on the left are probabilities for P and those on the right are probabilities for Q, but we would need agreement on $A \cap B$ in addition to A and B in order to apply $\pi - \lambda$.

$P(A) = P(B) = Q(A) = Q(B) = 0.3$ but $P(A \cap B) = 0, Q(A \cap B) = 0.3$. This stems from the fact that $\{A, B\}$ is not a π-system. On the other hand, $\{A, B, A \cap B\}$ *is* a π-system, and if P and Q agree on these sets it is straightforward to see that they must agree on all 16 sets in $\sigma(A, B)$.

Example 2.10.7 To illustrate $\pi - \lambda$, we show how to apply it to prove Proposition 2.9.10. Let X and Y be random variables with CDFs $F(x)$ and $G(y)$ respectively, such that $P(X \leq x, Y \leq y) = F(x)G(y)$. Why does it follow that X and Y are independent? Without $\pi - \lambda$, we might try to extend the equality from intervals $(-\infty, a]$ to more general intervals, then to open sets, then to intersections of open sets, etc., eventually trying to handle all Borel sets. Using $\pi - \lambda$ makes it much easier to bridge this gap; the only obstacle is that $\pi - \lambda$ pertains to subsets of Ω, whereas we need to show

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for all *pairs* (A, B) of Borel sets.

To handle this, start by *fixing* an interval $B = (-\infty, b]$ with $P(Y \in B) \neq 0$ (the equality is obvious if $P(Y \in B) = 0$). By assumption, we have

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for all A of the form $(-\infty, a]$. But the collection of all Borel sets A where this equality holds is a λ-system, so the equality holds for all Borel sets A. Now fix A rather than fixing B, and apply $\pi - \lambda$ again!

✎ 2.10.8 Prove that the CDF of a random variable completely determines its distribution, i.e., knowing $F(y) \equiv P(Y \leq y)$ for all $y \in \mathbb{R}$ determines $P(Y \in B)$ for all Borel sets $B \subseteq \mathbb{R}$.

Exercises

2.1 (**Preserving** $\sim$) Let $Y_1 \sim Y_2$ (i.e., they have the same distribution). Give a short proof that for every (Borel-measurable) function $g : \mathbb{R} \to \mathbb{R}$, we have $g(Y_1) \sim g(Y_2)$.

2.2 (**Introducing an independent r.v.**) Show that if $X \sim Y$, $X \perp\!\!\!\perp Z$, and $Y \perp\!\!\!\perp Z$, then $(X, Z) \sim (Y, Z)$.

2.3 (**Sums of r.v.s**) Show that if X and Y are r.v.s on the same probability space, then $X + Y$ is an r.v. on the same space.

2.4 (**Translation noninvariance**) (a) Show that it is impossible to have $X \sim (X + 1)$.

(b) On the other hand, show that for any $p < 1$ it is possible to have $X_1 \sim X_2$ with $P(X_2 = X_1+1) \geq p$.

Hint: in (a), consider a CDF; in (b), consider a clock.

2.5 (**CDF Conventions**) Let X be a r.v. whose possible values are $0, 1, 2, \ldots$, with CDF F. Rather than using a CDF, an alternative convention is to use the function G defined by $G(x) = P(X < x)$ to specify a distribution. Give simple formulas (possibly in terms of ceiling or floor functions) for how to convert from F to G and back again.

2.6 (**Functions of independent r.v.s**) Show that $X \perp\!\!\!\perp Y$ implies $g(X) \perp\!\!\!\perp h(Y)$ for any functions g, h.

2.7 (**Independent functions of the same r.v.**) Find an example of a random variable X and nontrivial functions g, h such that $g(X) \perp\!\!\!\perp h(X)$.

2.8 (**Total variation distance**) The *total variation distance* between two probability measures P and Q is the maximum discrepancy between them, $||P - Q||_{TV} \equiv \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$. Assume that Ω is finite and that $\mathcal{F} = 2^{\Omega}$. Show that

$$||P - Q||_{TV} = \frac{1}{2} \sum_{\omega \in \Omega} |P(\{\omega\}) - Q(\{\omega\})|$$

(so that, aside from the constant factor, the total variation distance is L_1 distance).
Hint: consider the set $B \equiv \{\omega : P(\{\omega\}) \geq Q(\{\omega\})\}$, and note that $P(B) - Q(B) = Q(B^c) - P(B^c)$.

2.9 (**Rational endpoints**) Suppose that $P(X \in I, Y \in J) = P(X \in I)P(Y \in J)$ for all finite open intervals I and J *with rational endpoints*. Does it follow that X and Y are independent?

2.10 (**Northeast rectangles**) Let Ω be the unit square $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, and let P and Q be probability measures on Ω (defined for all Borel sets within the square).

(a) Show that if $P(A) = Q(A)$ for all "Northeast" rectangles A, where a Northeast rectangle is defined to be a rectangle of the form $\{(x, y) \in \Omega : x \geq x_0, y \geq y_0\}$ for some $(x_0, y_0) \in \Omega$, then $P = Q$.

(b) Show by example that it is possible to have P and Q agree on all horizontal strips $\{(x, y) \in \Omega : c \leq y \leq d\}$ and on all vertical strips $\{(x, y) \in \Omega : a \leq x \leq b\}$, yet have $P \neq Q$.

2.11 (**Coupling Bound**) A common and elegant technique for bounding rates of convergence of certain stochastic processes is to use a *coupling argument*. Suppose that $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ are random variables which "coalesce" at some random time T, in the sense that $X_n = Y_n$ for all $n \geq T$, with T a random variable taking positive integer values. Show that the total variation distance between the distributions at time n (see Problem 2.8) is bounded by the probability that the X's and Y's haven't coalesced yet:

$$||\mathcal{L}(X_n) - \mathcal{L}(Y_n)||_{TV} \leq P(T > n).$$

Hint: Bound $P(X_n \in B) - P(Y_n \in B)$ by considering the cases $T \leq n$ and $T > n$, and noting that $\{X_n \in B, T \leq n\} = \{Y_n \in B, T \leq n\}$ as events.

3

Reasoning by Representation and the Named Distributions

Our forefathers made one mistake. What they should have fought for was representation without taxation.
– Fletcher Knebel

‘Welcome!’ said Griffin. ‘I hope you will find this forest of interest.’
‘I have come a long way,’ said Craig, ‘and I am very curious to know what birds you have here.’
‘A starling and a kestrel,’ replied Griffin.
‘That’s all?’ asked Craig.
‘And all birds derivable from them,’ replied Griffin.
‘Oh, that’s different! Are many birds derivable from just the starling and the kestrel?’
‘Very many indeed!’ replied Griffin, with a subtle and rather mysterious smile.
– Raymond Smullyan, in *To Mock a Mockingbird*

3.1 Introduction

This chapter introduces many of the named families of distributions, where by “named” we mean that they are famous through their many applications, and fundamental as building blocks for models and more complicated distributions. There are many ways to define a specific distribution. Most commonly in the literature, a density is provided. While useful, this is a “calculus” approach that can easily obscure motivation and properties provided by a more “elementary” approach. Many other ways to specify a distribution are also possible, based on various transforms such as the moment generating function (a form of the Laplace transform), characteristic function (more commonly called the *Fourier transform* outside of probability), ... All of these transforms are computationally useful, but again they are often unilluminating. Instead of plowing ahead with brute force density calculations, we introduce and study distributions via *representation.*

A representation is an expression for a distribution in terms of random variables with already-known distributions. We use representation definitions and arguments whenever possible because this often gives simpler, more intuitive proofs and shows how many important distributions are closely connected. Furthermore, the approach makes it more transparent how and when to handle location, scale, transformations, mixtures, and conditional distributions in models.

Defining distributions via representation leads to a logical progression for introducing the main distributions. Knowing the representation also immediately suggests *simulation* techniques, useful for efficiently drawing random samples from the distributions.

But is reasoning by representation rigorous? The answer is *yes*, provided that the essence of the problem at hand is about distributions rather than specific r.v.s realizing those distributions. This is typically the case in probability since as soon as we decide to consider the probability of our r.v. behaving in some way, we are looking at properties of the distribution, rather than properties of the specific r.v. Given this, the CYF (Choose Your Favorite) principle takes effect: if the answer depends only on the distribution, choose your favorite construction of a r.v. with that distribution.

The reader is encouraged to use representation as the first line of approach to the problems in this chapter and beyond, unless otherwise noted.

Of course, care is required since equality in distribution is not the same thing as equality of random variables; we must be especially careful not to alter the dependence structure of a problem without justification. Reasoning by representation is helped greatly by Exercise 2.1, which is so useful that we include a short proof here.

Lemma 3.1.1 *If $X \sim Y$ and g is a measurable function, then $g(X) \sim g(Y)$.*

Proof Use preimages:

$$P(g(X) \in B) = P(X \in g^{-1}(B)) = P(Y \in g^{-1}(B)) = P(g(Y) \in B).$$

□

Thus, we can "do the same deterministic thing to both sides" of any representation. We can also add or multiply both sides by a r.v., as long as we are careful about independence.

Lemma 3.1.2 *If $X \sim Y$ and $Z \perp\!\!\!\perp X, Z \perp\!\!\!\perp Y$, then $X + Z \sim Y + Z$ and $XZ \sim YZ$. More generally, $g(X, Z) \sim g(Y, Z)$ for any measurable $g \colon \mathbb{R}^2 \to \mathbb{R}$.*

Proof This is immediate from the structure of the problem: $X + Z$ and $Y + Z$ are both distributed as the sum of a r.v. with distribution $\mathcal{L}(X)$ and an independent r.v. with distribution $\mathcal{L}(Z)$. Similarly, we have $(X, Z) \sim (Y, Z)$, so $g(X, Z) \sim g(Y, Z)$. □

☣ 3.1.3 The assumption that Z is independent of X and independent of Y is crucial in the above lemma. Also, note that $X_1 Y \sim X_2 Y$ does not necessarily imply $X_1 \sim X_2$, even if X_1, X_2, Y are independent and never zero. For a simple example, take X_1 and Y to be independent random signs (i.e., 1 or -1 with probability 1/2 each), and take $X_2 = 1$. We will see in a later chapter that if X_1, X_2, Y are independent and *positive* r.v.s, then this cancellation property does hold.

3.2 Bernoulli and Binomial

Met a new distribution today! Hi, what's your name, what's your story, who represents you?
– Stat 210 student

We start with sheer simplicity: a coin flip. Let Y be an indicator random variable for Heads: 1 if the coin lands Heads, and 0 if it lands Tails. Of course, what matters here is the *distribution* of Y, not whether it resulted from a coin flip.

Definition 3.2.1 (Bernoulli) A random variable Y has the *Bernoulli distribution* with parameter p, denoted by $Y \sim \text{Bern}(p)$, if $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$.

Definition 3.2.2 (Symmetric Bernoulli) A random variable S has the *Symmetric Bernoulli* distribution, and is called a *random sign*, if $S = 2Y - 1$ for $Y \sim \text{Bern}(1/2)$.

The Binomial distribution is then defined by representation as the sum of i.i.d. Bern(p) random variables.

Definition 3.2.3 (Binomial) Let $Y = Y_1 + \cdots + Y_n$, with the Y_j's i.i.d. Bern(p). Then we say that Y is Binomial with sample size n and probability of success p, and we write $Y \sim \text{Bin}(n, p)$.

Definition 3.2.4 (Convolution) The *convolution* of two distributions F, G is the distribution obtained from a sum $X + Y$ of independent r.v.s $X \sim F, Y \sim G$. For example, the above definition states that the Binomial is the convolution of n i.i.d. Bernoullis; here n is the *convolution parameter.*

☣ 3.2.5 The CDF of $X + Y$ is not $F + G$. Of course, $F + G$ is not even a CDF, but it is a common mistake to confuse a r.v. with its CDF.

The density of $Y \sim \text{Bin}(n, p)$ is $P(Y = k) = \binom{n}{k} p^k q^{n-k}$ for $k \in \{0, 1, \ldots, n\}$ (with $q \equiv 1-p$), since the probability of any specific sequence of values for $Y_1, \ldots, Y_n$ with k 1's is $p^k q^{n-k}$, and the binomial coefficient counts the number of ways to choose where the k 1's go. Many results about the Binomial are much easier to anticipate and prove using the definition by representation rather than with density calculations.

Example 3.2.6 If $X \sim \text{Bin}(m, p), Y \sim \text{Bin}(n, p)$ independently, then $X + Y \sim \text{Bin}(m + n, p)$. This follows immediately by representation, as we can view $X + Y$ as the sum of $m + n$ i.i.d. Bern(p)'s. To use densities instead would require using a famous combinatorial formula, *Vandermonde's identity*:

$$\sum_{j=0}^{k} \binom{m}{j} \binom{n}{k-j} = \binom{m+n}{k}.$$

The representation argument is not only simpler, but also it gives an easy proof of Vandermonde's identity.

☣ 3.2.7 Note that both X and Y need to have "the same p" in the above example; the distribution does not take a simple form if the success probabilities differ.

Story 3.2.8 Suppose that $X|N \sim \text{Bin}(N, p_1)$, where $N \sim \text{Bin}(n, p_2)$ (so the sample size of X is itself random). What is the marginal (unconditional) distribution of X? A brute force approach would be to start calculating sums; a representational approach makes it easy to show that $X \sim \text{Bin}(n, p_1 p_2)$.

To see this, it helps to have a *story* in mind, i.e., a concrete application which does not lose generality. For example, suppose there are n eggs, each of which hatches with probability p_2. For any egg that hatches a chick, assume that the chick survives with probability p_1. Then we can interpret N as the number of eggs that hatch and X as the number of hatched chicks that survive.

Thus, $X \sim \text{Bin}(n, p_1 p_2)$ as each of the original n eggs has a "success" if it hatches and the chick survives. More formally, we can represent X as $\sum_{j=1}^{n} X_j N_j$ where $X_j \sim \text{Bern}(p_1)$ and $N_j \sim \text{Bern}(p_2)$, independently, with $N = \sum_{j=1}^{n} N_j$.

3.3 Uniform Distribution

Armed with an infinite sequence of coin tosses, we define the *Uniform distribution.*[1]

Definition 3.3.1 (Uniform) Let $Y_1, Y_2, \ldots$ be i.i.d. Bern(1/2). Then we say that $U \equiv \sum_{j=1}^{\infty} \frac{Y_j}{2^j}$ has the *Uniform distribution*, and we write $U \sim \text{Unif}$. This is the *dyadic representation* of a real number in $[0, 1]$.

It is more common to define the Uniform distribution by its having constant density on $[0, 1]$, but we shall see below that this is equivalent to the representation in terms of Bernoullis. It is also more common to use notation such as Unif(0,1) to indicate the support of the distribution, but we prefer to write $U \sim \text{Unif}$ both for simplicity and to emphasize that *any two Uniform distributions are related by a change of location and scale.* That is, for any real $a < b$, if $U \sim \text{Unif}$ then $a + (b - a)U$ is Uniform between a and b. Writing $a + (b - a)U$ rather than $\text{Unif}(a, b)$ serves as a reminder that the distribution is obtained from the standard Uniform by shifting by a and scaling by $b - a$.

✎ 3.3.2 Show that if $U \sim \text{Unif}$, then $1 - U \sim \text{Unif}$. Also, show that $2U - 1 \sim SU$ for S a random sign independent of U (so $2U - 1$ is symmetric about 0, while U is symmetric about 1/2; see Section 3.10).

☣ 3.3.3 It is *not* true in general that sums or nonlinear functions of Uniforms are Uniform. Certain functions of a Uniform are discussed in the next section.

✎ 3.3.4 Let $U \sim \text{Unif}$. Construct two independent Uniforms U_1, U_2 as functions of U. More generally, construct i.i.d. Unif $U_1, U_2, \ldots$, in terms of just U.
Hint: why do the rational numbers have the same cardinality as the integers?

✎ 3.3.5 For continuous Uniform distributions, we saw above that it is easy to convert one to another by a change in location and scale. For discrete Uniforms, more work is required. Write $Y \sim \text{DUnif}(k)$ if Y is discrete Uniform over $1, 2, \ldots, k$ (so $P(Y = j) = 1/k$ for $j \in \{1, \ldots, k\}$). Given $Y_3 \sim \text{DUnif}(3)$, is it possible to construct $Y_9 \sim \text{DUnif}(9)$ in

[1] In practice, we have only finitely many Bernoullis, but we can only generate a Uniform to finite precision anyway. In theory, one may wonder how to construct infinitely many i.i.d. random variables. It turns out that nature is bountiful: given any distribution, it is possible to construct as many i.i.d. copies as one could ever desire; this is most easily seen using the Probability Integral Transform of Section 3.4.

terms of Y_3? What about given two independent DUnif(3)'s? What if we need to generate DUnif(10)'s rather than DUnif(9)'s, but we have access only to a stream of DUnif(3)'s?

We now show that the definition in terms of a sequence of coin flips agrees with the usual definition based on the CDF or density.

Theorem 3.3.6 *Let* $U \sim$ Unif. *Then the CDF of* U *is* $F(u) = u$ *for* $0 \leq u \leq 1$ *(and* $F(u) = 0$ *for* $u < 0$, $F(u) = 1$ *for* $u > 1$*), and the density is 1 on* $[0,1]$ *and 0 elsewhere.*

Proof Write $U = \sum_{j=1}^{\infty} \frac{Y_j}{2^j}$, and interpret this as a binary (dyadic) expansion $U = 0.Y_1Y_2\dots$. Fix a number $u = 0.u_1u_2 \cdots \in [0,1]$, also expanded in binary. To compute $P(U < u)$, note that in comparing two numbers, it suffices to look at the first position in which they differ (here there is a mildly annoying technicality in that some numbers have two binary expansions, e.g., $0.0111 \cdots = 0.1000\dots$; there are only countably many such numbers though, so their combined probability is 0. Also, the event $U = u$ has 0 probability and so will be ignored.) Let J be the first position where U and u differ. Conditioning on J, we have

$$P(U < u) = \sum_{j=1}^{\infty} P(U < u | J = j) P(J = j) = \sum_{j=1}^{\infty} \frac{u_j}{2^j} = u.$$

Thus, the CDF and density are as claimed. □

3.4 Probability Integral Transform

In this section, we prove the beautiful and useful fact that *any* distribution can be represented as a function of a Uniform r.v. The method goes by various names in the literature, but is most commonly known as the *Probability Integral Transform (PIT)*. We also call this method and property the *Universality of the Uniform*, since it asserts that a Uniform r.v. can, in principle, be converted to an r.v. with *any* desired distribution.

Definition 3.4.1 The *quantile function* of a distribution with CDF F is the function

$$F^{-1}(p) = \min\{x : F(x) \geq p\},$$

defined on $(0,1)$. Note that if F is continuous and strictly increasing, F^{-1} is indeed the inverse of F. Of course, F may have jumps or regions where it is flat, in which case F^{-1} serves as a surrogate for an inverse.

✎ 3.4.2 Sketch the CDF F and the quantile function F^{-1} of the Bin(4,1/2) distribution. Note that jumps in F correspond to flat regions in F^{-1}, while flat regions in F correspond to jumps in F^{-1}, and that F is right continuous, while F^{-1} is left continuous.

✎ 3.4.3 Show that the minimum in the definition of the quantile function is achieved (which justifies writing "min" rather than "inf").

Definition 3.4.4 A number m is a *median* of a distribution F if for $X \sim F$, we have $P(X \leq m) \geq 1/2$ and $P(X \geq m) \geq 1/2$. One choice of m is $F^{-1}(1/2)$, which we denote by med(X) for $X \sim F$, but in general there may be many medians.

Remark 3.4.5 Other definitions of the quantile function are sometimes used, such as $\inf\{x : F(x) > p\}$. This need not trouble us much here, though in practice it sometimes has significant consequences, e.g., legislation is sometimes written using terms such as "median" ambiguously, possibly with significant repercussions.

Theorem 3.4.6 (Probability Integral Transform Sampling) *Let F be any CDF, with quantile function F^{-1}, and let $U \sim$ Unif. Then $Y \equiv F^{-1}(U) \sim F$.*

Going in the other direction, given any continuous CDF (here we mean that the CDF is continuous as a function, not in the sense of "continuous random variable"), we can generate a Uniform r.v. in a beautifully self-referential way: *put a random variable as the argument of its own CDF*, letting $U = F(Y)$.

Theorem 3.4.7 (Probability Integral Transform Pivoting) *Let F be a CDF which is continuous as a function from $\mathbb{R}$ to $[0, 1]$, and let $Y \sim F$. Then $U \equiv F(Y) \sim$ Unif.*

✎ 3.4.8 Sketch the graph of a CDF, and explain how the PIT uses random quantiles to create a correspondence between U on the vertical axis and Y on the horizontal axis.

✎ 3.4.9 Explain why the assumption that F is continuous as a function is needed for the second form of the PIT, but not for the first.

The first form of the Probability Integral Transform (PIT), Theorem 3.4.6, shows how to simulate any distribution using a Uniform. Generating pseudorandom numbers that "look" Unif can be done by almost any computational software, and then using the Unifs we can sample from other distributions. How useful this is in practice depends of course on F, since the quantile function may be very difficult to compute. The PIT is also useful in many proofs; for example, it plays a prominent role in Skorohod's Theorem (see Chapter 10).

The second form of the PIT, Theorem 3.4.7, is also often useful. Specifically, suppose that $Y \sim F_\theta$ has a distribution depending on an unknown parameter θ (where F_θ is a continuous function). Then $F_\theta(Y) \sim$ Unif is a *pivotal quantity*, i.e., its distribution does not depend on any unknown parameters. Pivotal quantities are often useful tools, e.g., in constructing confidence intervals.

✎ 3.4.10 Prove Theorem 3.4.6 for the case that F is continuous and strictly increasing, by directly computing the CDF of Y.

We now prove that the PIT Theorem 3.4.6 holds in general, for all CDFs.

Proof of Theorem 3.4.6 We first show that $u \leq F(y)$ is equivalent to $F^{-1}(u) \leq y$, for all $u \in (0, 1), y \in \mathbb{R}$. By definition of the quantile function, $u \leq F(y)$ implies $F^{-1}(u) \leq y$. Conversely, if $F^{-1}(u) \leq y$ then $u \leq F(F^{-1}(u)) \leq F(y)$. So the two events $U \leq F(y)$ and $F^{-1}(U) \leq y$ are in fact the same event. Thus, the CDF of $F^{-1}(U)$ is $P(U \leq F(y)) = F(y)$.

The other form of the PIT follows readily.

Proof of Theorem 3.4.7 Let $Y \sim F$, a CDF which is a continuous function. Reasoning by representation, we can take $Y = F^{-1}(U)$, with $U \sim$ Unif. Then $F(Y) = F(F^{-1}(U)) = U$ because F takes on every value in $(0, 1)$.

Example 3.4.11 (Logistic) The *Logistic distribution* is defined by representation as the distribution of $\log(U/(1-U))$, where $U \sim$ Unif. Note that this representation makes it easy to sample random draws from this distribution, given Uniform draws. The name stems from the *logit function* (log-odds): $\text{logit}(p) = \log(p/(1-p))$, mapping $(0,1)$ onto the entire real line. The inverse function, $\text{logit}^{-1}(y) = e^y/(1+e^y)$, is known as the *logistic function.* Logit and its inverse are widely used in statistical and mathematical modeling (e.g., in logistic regression and population dynamics). The PIT makes it obvious that the Logistic distribution has logit^{-1} as its CDF.

✎ 3.4.12 Let Y_1 and Y_2 be r.v.s (possibly defined on different Ω's) with CDFs F_1 and F_2 respectively. A commonly used partial order on distributions (and thus on r.v.s), *stochastic domination*, is defined by the relation $Y_1 \preceq Y_2$ iff $F_1(y) \geq F_2(y)$ for all $y \in \mathbb{R}$.
Find an example of Y_1 and Y_2 on the same space with $Y_1 \preceq Y_2$, but $P(Y_1 > Y_2) \geq 0.95$.

Hint: clockwork.

✎ 3.4.13 Show that $Y_1 \preceq Y_2$ if and only if there exist Y_1^*, Y_2^* on the same space Ω^* as each other (not necessarily the same space or spaces on which Y_1, Y_2 are defined) such that $Y_1^* \sim Y_1, Y_2^* \sim Y_2$, and $Y_1^* \leq Y_2^*$. That is, stochastic domination can be replaced by ordinary inequality of r.v.s, after replacing Y_1 and Y_2 by suitably chosen copies.

Hint: use the PIT with the *same* Uniform for both r.v.s. This is an example of a *coupling argument*, bringing together several r.v.s to a common space.

3.5 Exponential and Gamma Distributions

Definition 3.5.1 (Expo and Gamma) The Exponential distribution is defined by representation as the distribution of $X = -\log U$, where $U \sim$ Unif. This is denoted by $X \sim$ Expo. When r is a positive integer, we define Gamma(r) by representation, as the distribution of $G_r = X_1 + \cdots + X_r$, with the X_j's i.i.d. Expo. Here r is called the *convolution parameter.*

Remark 3.5.2 Most authors write the Exponential with a parameter, as in Expo(θ); but there is no agreement on whether θ is a scale parameter or a rate parameter (reciprocal to the scale parameter). We write the scale explicitly: if $X \sim$ Expo, we can let $Y = \mu X$ to scale by a positive constant μ. Then we write $Y \sim \mu$Expo.

Similarly, Gamma is generally introduced as a 2-parameter family, Gamma(a, b). In addition to the scale vs. rate issue, there is a lack of agreement on the order in which to list the parameters. Again we prefer to write the scale explicitly: $G \sim$ Gamma(r) has convolution parameter r and scale parameter 1, and we can rescale G if needed, e.g., letting $Y \sim \lambda^{-1}G$.

Story 3.5.3 (Memoryless Property) A crucial property of the Exponential distribution, which in fact characterizes it, is the *memoryless property.* Intuitively, if we have, for example, that $T \sim$ Expo is the lifetime of some product, the memoryless property says that the product is always good as new, not remembering or caring how much time has elapsed: regardless of how much time has passed, if the product is still working at a certain time, then the additional lifespan is still Expo. Formally, let $X \sim$ Expo. Then X has the memoryless property, which is defined to mean that the conditional distribution of $X - a$ given $X > a$ is also Expo; equivalently, this says that $P(X > a+b) = P(X > a)P(X > b)$ for all $a > 0, b > 0$. Conversely, if X is a continuous, memoryless r.v. with support $(0, \infty)$, then $X \sim \mu$Expo for some $\mu > 0$.

✎ 3.5.4 Prove the above characterization.

Hint: Let $G(x) = P(X > x)$ be the *survival function*. The memoryless property amounts to the functional equation $G(x + y) = G(x)G(y)$. Consider various choices of x and y to find all continuous, decreasing functions G satisfying this equation.

Returning to the the Gamma distribution, to go further we need its namesake: the *gamma function*, which is one of the most important "special functions" in all of mathematics. The gamma function is defined by

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

which is finite for any $\alpha > 0$. (In complex analysis, it is shown that the domain can be extended to all complex numbers except for 0 and negative integers.) Integration by parts quickly yields the identity

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

for all $\alpha > 0$. In particular, for n a positive integer we have $\Gamma(n) = (n-1)!$, so the gamma function is a generalization of the factorial function. We show below that

$$\Gamma(1/2) = \sqrt{\pi},$$

from which all the other half-integer values can be obtained: $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2) = \sqrt{\pi}/2$, $\Gamma(5/2) = 3\sqrt{\pi}/4$, etc.

Proposition 3.5.5 *For any positive integer r, the Gamma(r) density is given by*

$$f(x)dx = \Gamma(r)^{-1} e^{-x} x^r \frac{dx}{x}$$

for $x > 0$, and 0 otherwise.

The above density makes sense for all real $r > 0$, so we extend the Gamma family to allow the convolution parameter r to be any positive real number. We write the above density with respect to $\frac{dx}{x}$ rather than "simplifying" x^r/x to x^{r-1} since this avoids the need for Jacobians if we make a scale change $x = \lambda y$: the dx/x is *invariant* (formally, this is known as working with *Haar measure*).

Of course, we have Gamma(1) $\sim$ Expo.

✎ 3.5.6 Prove Proposition 3.5.5 by induction.

To show that $\Gamma(1/2) = \sqrt{\pi}$, a famous and beautiful trick can be used. First, make a change of variables to obtain $\Gamma(1/2) = \int_{-\infty}^{\infty} e^{-x^2} dx$. When faced with a hard integral, rarely does it help to write down the integral a second time beside the first; usually, writing down the same problem repeatedly is more likely to be a sign of frustration than a brilliant tactic. Here however, this allows for a neat conversion to polar coordinates.

$$\begin{aligned}\Gamma^2(1/2) &= \int_{-\infty}^{\infty} e^{-x^2}dx \int_{-\infty}^{\infty} e^{-x^2}dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}dxdy \\ &= \int_0^{\infty}\int_0^{2\pi} e^{-r^2} rdrd\theta \\ &= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} e^{-u}dud\theta \\ &= \pi,\end{aligned}$$

where the Jacobian r from converting to polar coordinates comes to the rescue.

The *Laplace distribution* is a symmetrized Exponential distribution. (See Section 3.10 for discussion of symmetry.)

Definition 3.5.7 (Laplace) A *Laplace* random variable is obtained by multiplying an Expo by a random sign: $L \sim$ Laplace if $L \sim SX$, where S is a random sign and $X \sim$ Expo are independent.

Next we define the *Weibull distribution*, which is widely used in survival and reliability analysis (e.g., to model the lifetime of a product) because of its versatility and convenient *hazard function* (which gives the density for the product failing at a particular point in time, given that it has survived up until that time).

Definition 3.5.8 (Weibull) The *Weibull distribution* is given by powers of an Exponential: letting $X \sim$ Expo, the power $W = X^\beta$ is Weibull with shape parameter $\beta > 0$, denoted by $W \sim \text{Wei}(\beta)$. As usual, a scale parameter is often introduced.

✎ 3.5.9 Find the CDF and density of $\text{Wei}(\beta)$.

✎ 3.5.10 Let $W_1 \sim \text{Wei}(\beta_1), W_2 \sim \text{Wei}(\beta_2)$. Show that $(W_1|W_1 \geq 1) \preceq (W_2|W_2 \geq 1)$ iff $\beta_1 \leq \beta_2$.

3.6 Normal Distribution

The Normal distribution is well-deserving of its central role in probability and statistics. Much of this stems from the Central Limit Theorem (see Chapter 13) and the richness of the Multivariate Normal distribution (see Chapter 8). But the Normal enjoys many other special properties. For example, there are many characterizations of the Normal, such as:

1. Maxentropic: the Normal maximizes entropy, for prescribed mean and variance.

2. Invariance under rotations: if Y_1, Y_2 are i.i.d. such that $Y_1 + Y_2, Y_1 - Y_2$ are also independent, then Y_1 and Y_2 are Normal. There are other *spherically symmetric distributions* (i.e., where rotating (Y_1, Y_2) by some fixed angle in the plane does not change the distribution), but if Y_1, Y_2 are i.i.d. and a rotated version still has independence between the components, then Normality holds.
3. Independence of sample mean and sample variance: If $Y_1, \ldots, Y_n$ are i.i.d. with the sample mean $\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$ independent of the sample variance $S^2 = \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$, then the Y_j are Normal.
4. Stable law with finite variance: the Normal is the only stable law with a finite variance (see Example 3.6.15 for discussion of stable laws).

There are other characterizations too, stemming from the Normal family (with mean parameter μ and σ^2 fixed) being both a location family and a natural exponential family (these concepts are discussed later). There are thus many equivalent ways in which one can define the Normal. Here we first define the Chi-Square distribution, which usually is defined through sums of squares of i.i.d. Normals, and then define the standard Normal distribution as a symmetrized χ_1, and then define any Normal in terms of the standard Normal via location and scale.

Chi-Square distributions appear throughout statistics, and (conveniently) are special cases of Gammas.

Definition 3.6.1 (Chi and Chi-Square) Write $G \sim \chi_n^2$ if $G \sim 2\text{Gamma}(n/2)$, for all $n \in \{1, 2, \ldots, \}$. This defines the *Chi-Square distribution* with n *degrees of freedom.* The "Square" in "Chi-Square" suggests that there should be a Chi distribution too. Naturally enough, we define the Chi as the square root of a Chi-Square, and write $W \sim \chi_n$ if $W^2 \sim \chi_n^2$ and $W \geq 0$.

Definition 3.6.2 (Normal) The *standard Normal distribution* $\mathcal{N}(0, 1)$ is the distribution of $Z \equiv SX$, where S is a random sign independent of $X \sim \chi_1$. The *Normal distribution* $\mathcal{N}(\mu, \sigma^2)$ with parameters μ (the *mean*) and σ^2 (the *variance*) is defined to be the distribution of $Y = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0, 1)$. We call μ a *location parameter* and σ a *scale parameter*, and assume $\sigma \geq 0$ (the case $\sigma = 0$ is degenerate, and sometimes is implicitly excluded). Collectively, the distributions $\mathcal{N}(\mu, \sigma^2)$ form the *Normal family* of distributions.

In the next chapter, mean and variance will be defined formally. For now, they can just be viewed as the two parameters which convert from a standard Normal to any Normal: $Y = \mu + \sigma Z$ shows how to create a family of distributions from one distribution, by rescaling (by multiplying by σ, which widens or narrows the spread of the distribution) and shifting to a new location (by adding μ). This is an example of a *location-scale family.*

The fact that Chi-Square is given by a sum of squares of i.i.d. Normals, which is usually taken as the definition of Chi-Square, follows readily.

✎ 3.6.3 Show that the sum of squares $G \equiv Z_1^2 + \cdots + Z_n^2$, where $Z_j \sim \mathcal{N}(0, 1)$ are i.i.d., satisfies $G \sim \chi_n^2$.

Proposition 3.6.4 *The standard Normal density is*

$$f(z)dz = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)dz.$$

✎ 3.6.5 Prove the above proposition, using the fact that $\Gamma(1/2) = \sqrt{\pi}$ to show that the density integrates to 1.

Remarkably, it has been shown that the standard Normal CDF cannot be obtained in closed form in terms of "elementary" functions. As consolation, there is a standard name for this function.

Definition 3.6.6 (Standard Normal CDF) The CDF of the standard Normal distribution is denoted by Φ.

✎ 3.6.7 Show that $\Phi(-z) = 1 - \Phi(z)$.

A handy approximation for Φ is given below.

Approximation 3.6.8

$$\operatorname{logit}(\Phi(z)) \approx 1.6z\sqrt{1 + \frac{z^2}{10}},$$

where $\operatorname{logit}(p) \equiv \log(\frac{p}{1-p})$.

The Box-Muller representation generates two independent Normals from two independent Uniforms. Somewhat paradoxically, it is about as easy to generate two Normals as to generate one! The representation is also striking in that at first glance, Z_1 and Z_2 hardly look independent; but they are.

Theorem 3.6.9 (Box-Muller) *Let U_1, U_2 be i.i.d. Uniform r.v.s. Then*

$$Z_1 \equiv \sqrt{-2\ln U_2}\cos(2\pi U_1)$$

$$Z_2 \equiv \sqrt{-2\ln U_2}\sin(2\pi U_1)$$

are i.i.d. $\sim \mathcal{N}(0,1)$. *Note that here* $-\ln U_2 \sim \text{Expo}$, *so* $\sqrt{-2\ln U_2} \sim \chi_2$.

Proof To prove Box-Muller, we start with the i.i.d. Normals Z_1, Z_2, and show how the representation arises naturally. Consider (Z_1, Z_2) as a point in the plane. In polar coordinates, let $R^2 \equiv Z_1^2 + Z_2^2 \sim \chi_2^2$ be the radius squared and let $\theta \in [0, 2\pi)$ be the angle. Note that by symmetry, R^2 is independent of θ, and $\theta \sim 2\pi\text{Unif}$. Formally, this follows from the fact that the joint density of Z_1, Z_2 depends only on the distance from (Z_1, Z_2) to the origin, and conditioning on this distance then gives a Uniform distribution. Representing $R \sim \chi_2$ in terms of a Unif, we have written Z_1, Z_2 in the desired form. The argument was "backwards" in the sense that we started with Normals and derived Uniforms, but the result follows since the expressions involving U_1, U_2 have a uniquely determined joint distribution, so we are free to choose any construction that yields these expressions. □

Example 3.6.10 A useful fact is that the sum and difference of i.i.d. standard Normal r.v.s are themselves i.i.d. Normals (a more general version of this is given in the Multivariate Normal chapter). Let Z_1, Z_2 be i.i.d. $\mathcal{N}(0,1)$, and show that $Z_1 + Z_2$ and $Z_1 - Z_2$ are independent and distributed as $\mathcal{N}(0,2)$. To show this by representation, use the Box-Muller representation to write $Z_1 = W\cos(2\pi U), Z_2 = W\sin(2\pi U)$ with $W \sim \chi_2, U \sim$ Unif. Then

$$Z_1 + Z_2 = W(\cos(2\pi U) + \sin(2\pi U)) = \sqrt{2}W\sin(2\pi U + \pi/4),$$

$$Z_1 - Z_2 = W(\cos(2\pi U) - \sin(2\pi U)) = \sqrt{2}W\cos(2\pi U + \pi/4).$$

Then by a slight variation of Box-Muller (where the angle is taken to be Uniform over $(\pi/4, \pi/4 + 2\pi)$ rather than over $(0, 2\pi)$), $Z_1 + Z_2$ and $Z_1 - Z_2$ are i.i.d. $\sim \mathcal{N}(0,2)$.

Box-Muller allows us to reduce messy calculus to basic algebra and trigonometry for some problems. Sometimes we wind up with an angle that is no longer in $[0, 2\pi)$, and then it is useful to know that "going around the circle k times" gives the same distribution as going around once, for any positive integer k.

Lemma 3.6.11 (Winding Lemma) *Let k be a positive integer and $U \sim$ Unif. Then* $\sin(2\pi kU) \sim \sin(2\pi U)$.

A short proof of the Winding Lemma is given in Chapter 5.

Example 3.6.12 Suppose that we want to find the distribution of $\frac{2XY}{\sqrt{X^2+Y^2}}$, where $X, Y \sim \mathcal{N}(0,1)$ are i.i.d. By Box-Muller, we can choose $X = R\cos\theta$, $Y = R\sin\theta$, with $R^2 \sim \chi_2^2$ independent of $\theta \sim 2\pi$Unif. Then by Box-Muller and the Winding Lemma,

$$\frac{2XY}{\sqrt{X^2+Y^2}} = \frac{2R^2\sin\theta\cos\theta}{R} = R\sin(2\theta) \sim \mathcal{N}(0,1).$$

Definition 3.6.13 (Student-t and Cauchy) The *Student-t distribution* with n degrees of freedom, denoted by t_n, is defined to be the distribution of

$$T = \frac{Z}{\sqrt{V_n/n}},$$

where $Z \sim \mathcal{N}(0,1)$ is independent of $V_n \sim \chi_n^2$. We denote this by $T \sim t_n$. The special case $n = 1$ is the *Cauchy distribution.* Note that by symmetry, the Cauchy can also be represented as the ratio Z_1/Z_2 of i.i.d. $\mathcal{N}(0,1)$ r.v.s.

Example 3.6.14 Let $C \sim$ Cauchy. What is the distribution of $(1+C)/(1-C)$? Rather than compute Jacobians, we use the representation $C = Z_1/Z_2$ with Z_1, Z_2 i.i.d. $\sim \mathcal{N}(0,1)$. Then

$$\frac{1+C}{1-C} = \frac{Z_2+Z_1}{Z_2-Z_1} \sim \frac{\sqrt{2}Z_1}{\sqrt{2}Z_2} \sim C,$$

using the fact from Example 3.6.10 that $Z_2 + Z_1$ and $Z_2 - Z_1$ are i.i.d. $\sim \mathcal{N}(0,2)$.

Example 3.6.15 (Stable Laws and the Inverted Chi-Square(1)) Above, we saw that the sum of two i.i.d. Normals is Normal (with a new scale), which is a very convenient property. We will see later that this extends to sums of any number of Normals: if $X_1, \dots, X_n$ are i.i.d. with $X_j \sim \mathcal{N}(\mu, \sigma^2)$, then $X_1 + X_2 + \cdots + X_n \sim \mathcal{N}(n\mu, n\sigma^2)$. Later, we will also see that sums of i.i.d. Cauchys are scaled Cauchys, which is a much less convenient property! Specifically, if $C_1, \dots, C_n \sim$ Cauchy are i.i.d., then $C_1 + \cdots + C_n \sim n$Cauchy. This says that taking the average of a billion Cauchys is no different (in distribution) from just taking one Cauchy. We prove this by representation in Chapter 5.

In general, a distribution F is called a *stable law* if for $X_1, X_2, \dots \sim F$ i.i.d., we have $X_1 + X_2 + \cdots + X_n \sim a_n X_1 + b_n$ for all n (with a_n, b_n constants). So the Normal and Cauchy distributions are stable.

Amazingly, only one other stable law is known to have a density which can be written in closed form: the Inverted Chi-Square(1) distribution, defined to be the distribution of the reciprocal of a χ_1^2 r.v. To show this for the sum of two of them, let Z_1, Z_2 be i.i.d. $\mathcal{N}(0,1)$. Then

$$\begin{aligned}
\frac{1}{Z_1^2} + \frac{1}{Z_2^2} &= \frac{1}{R^2}\left(\frac{1}{\sin^2\theta} + \frac{1}{\cos^2\theta}\right) \\
&= \frac{1}{R^2}\left(\frac{1}{\sin^2\theta\cos^2\theta}\right) \\
&= \frac{1}{R^2}\left(\frac{4}{\sin^2 2\theta}\right) \\
&\sim \frac{4}{Z_1^2}.
\end{aligned}$$

✎ 3.6.16 Let $C \sim$ Cauchy and $U \sim$ Unif. Show that

$$\tan(2\pi U) \sim C.$$

✎ 3.6.17 Let $C \sim$ Cauchy. Use the Box-Muller representation to show that $C - 1/C \sim 2C$.

The *Log-Normal distribution* is given by exponentiating a Normal. Some consider the name a misnomer, but the meaning is easy to remember since the log of a Normal is undefined; Log-Normal can be thought of as Log-Is-Normal. The Log-Normal distribution is widely used for modeling positive quantities such as heights, stock prices, and financial costs.

Definition 3.6.18 (Log-Normal) The *Log-Normal distribution* is defined by $Y = e^X$, where $X \sim \mathcal{N}(\mu, \sigma^2)$. We denote this by $Y \sim \mathcal{LN}(\mu, \sigma^2)$. That is, $Y \sim \mathcal{LN}(\mu, \sigma^2)$ if $Y = e^\mu e^{\sigma Z}$ with $Z \sim \mathcal{N}(0,1)$. Note that μ is a location parameter for X, but identifies a scale parameter for Y (through exponentiation).

3.7 Beta Distribution and Beta-Gamma Calculus

What's truer than truth? The story.

– Isabel Allende

The Beta family of distributions generalizes the Uniform distribution, allowing more flexibility in shape (e.g., skew, U-shaped densities, sharp peaks) while still being supported on the unit interval. Betas are particularly commonly used as a distribution for an unknown probability (e.g., in modeling a coin with an unknown bias p one might posit that p follows some Beta distribution). We will see below that Betas are very closely related both to Gammas and to Binomials, making them especially convenient to work with.

Definition 3.7.1 (Beta) The Beta(a, b) distribution is defined by representation as the distribution of

$$B = \frac{G_a}{G_a + G_b},$$

where $G_a \sim \text{Gamma}(a), G_b \sim \text{Gamma}(b)$ are independent (and $a > 0, b > 0$).

✎ 3.7.2 Show that if $B \sim \text{Beta}(a, b)$, then $1 - B \sim \text{Beta}(b, a)$.

Story 3.7.3 (Beta-Gamma) Suppose that one has waited G_1 minutes in line at the bank and G_2 minutes in line at the post office (independently). Is it true that the *total* time waiting, $G_1 + G_2$, is independent of the *proportion* of time one was waiting at the bank? For Gamma waiting times, this turns out to be true – and turns out to be an extremely useful fact to keep in mind when working with Gammas or Betas. Let $B = \frac{G_a}{G_a+G_b}$, with $G_a \sim \text{Gamma}(a) \perp\!\!\!\perp G_b \sim \text{Gamma}(b)$. Then the total $T \equiv G_a + G_b \sim \text{Gamma}(a + b)$ is independent of the proportion $B \sim \text{Beta}(a, b)$.

Remark 3.7.4 The Beta and Gamma are the *only* distributions where such an independence result holds: the above result *characterizes* the Gamma distribution. More precisely, Lukacs [6] showed that if X_1 and X_2 are independent, nondegenerate positive r.v.s such that $X_1/(X_1+X_2)$ is independent of X_1+X_2, then X_1 and X_2 follow Gamma distributions with the same scale.

Remark 3.7.5 A simple but useful fact that is sometimes overlooked is that x_1/x_2 is a function of $b = x_1/(x_1 + x_2)$ and vice versa, via

$$\frac{x_1}{x_2} = \frac{x_1/(x_1 + x_2)}{x_2/(x_1 + x_2)} = \frac{b}{1 - b}.$$

For example, it is easy to convert between a probability p and the corresponding odds $\frac{p}{1-p}$.

Example 3.7.6 Suppose that $Y_j \sim \frac{1}{\lambda_j}\text{Expo}$ are independent for $j \in \{1, 2\}$, and we wish to find $P(Y_1 < Y_2)$. This problem comes up in many settings, e.g., we can envision two people independently trying to solve a problem, at possibly different rates, and we may want to know the probability that the first person will finish the problem first. Rather than writing down a double integral, we can reason by representation as follows.

$$P(Y_1 < Y_2) = P\left(\frac{X_1}{X_2} < \frac{\lambda_1}{\lambda_2}\right),$$

with X_1, X_2 i.i.d. Expo. This becomes

$$P\left(\frac{X_1/X_2}{1+X_1/X_2} < \frac{\lambda_1/\lambda_2}{1+\lambda_1/\lambda_2}\right) = P\left(\frac{X_1}{X_1+X_2} < \frac{\lambda_1}{\lambda_1+\lambda_2}\right) = \frac{\lambda_1}{\lambda_1+\lambda_2},$$

since $X_1/(X_1+X_2) \sim$ Unif. This result makes sense intuitively since it is $1/2$ when $\lambda_1 = \lambda_2$ and it says that the probability of the first person finishing the problem first is proportional to his or her rate.

Before proving the Beta-Gamma result, we state the form of the Beta density.

Proposition 3.7.7 *If $B \sim \text{Beta}(a_1, a_2)$, then the density $f(b)$ of B is*

$$f(b)db = \frac{1}{\beta(a_1,a_2)} b^{a_1}(1-b)^{a_2} \frac{db}{b(1-b)}$$

on $(0,1)$, and 0 otherwise, where the normalizing constant $\beta(a_1,a_2)$ is the beta function, *given by*

$$\beta(a_1,a_2) = \frac{\Gamma(a_1)\Gamma(a_2)}{\Gamma(a_1+a_2)}.$$

✎ 3.7.8 Show that the Beta(a_1,a_2) density is unimodal if $a_1 > 1, a_2 > 1$, the Uniform density if $a_1 = a_2 = 1$, monotone if one of a_1, a_2 is less than 1 and the other is at least 1, and U-shaped if $a_1 < 1, a_2 < 1$.

By computing the joint density of the total and the proportion, both Theorem 3.7.3 and Proposition 3.7.7 follow in one fell swoop.

Proof Let

$$T = G_1 + G_2$$

$$B = \frac{G_1}{G_1+G_2} = \frac{G_1}{T},$$

where $G_j \sim \text{Gamma}(a_j)$. Then $G_1 = TB, G_2 = T(1-B)$.

The joint density of (T, B) is

$$g(t,b)dtdb = f(g_1,g_2)\left|\frac{d(g_1,g_2)}{d(t,b)}\right| dtdb$$

where the Jacobian is

$$J = \det(\frac{d(g_1,g_2)}{d(t,b)}) = \det\begin{pmatrix} \frac{\partial g_1}{\partial t} & \frac{\partial g_1}{\partial b} \\ \frac{\partial g_2}{\partial t} & \frac{\partial g_2}{\partial b} \end{pmatrix} = \det\begin{pmatrix} b & t \\ 1-b & -t \end{pmatrix} = -tb + tb - t = -t$$

Then the joint density of T and B factors neatly:

$$g(t,b)dtdb = \frac{t^{a_1+a_2-1}e^{-t}dt}{\Gamma(a_1+a_2)} \frac{b^{a_1-1}(1-b)^{a_2-1}\Gamma(a_1+a_2)db}{\Gamma(a_1)\Gamma(a_2)}$$

Since this separates into a function of T times a function of B, we have $T \perp\!\!\!\perp B$. Moreover, pulling the normalizing constant of $\text{Gamma}(a_1 + a_2)$ into its part of the expression immediately reveals the value of the Beta normalizing constant. □

With this fundamental Beta-Gamma relationship in place, we can build further extensions. For example, what happens to the relevant proportions with three Gammas instead of two?

Proposition 3.7.9 *Let G_1, G_2, G_3 be independent Gamma r.v.s. Then the random variables $B_1 \equiv \frac{G_1}{G_1+G_2}$, $B_2 \equiv \frac{G_1+G_2}{G_1+G_2+G_3}$, and $S \equiv G_1 + G_2 + G_3$ are fully independent.*

Proof Let $T \equiv G_1 + G_2$. Then

$$B_2 = \frac{T}{T + G_3} \perp\!\!\!\perp T + G_3 = G_1 + G_2 + G_3 = S.$$

The joint independence of B_1, B_2, and S then follows from the independence lemma 2.9.11 from Chapter 2, taking $X = (G_1 + G_2, G_3)$ and $Y = G_1/(G_1 + G_2)$ (which are independent by a second application of the lemma). We have $G_1 + G_2 + G_3 \perp\!\!\!\perp B_2$, both of which are functions of $(G_1 + G_2, G_3)$, and it follows that B_1, B_2, S are fully independent. □

Proposition 3.7.10 *We have*

$$\text{Beta}(a, b) \cdot \text{Beta}(a + b, d) \sim \text{Beta}(a, b + d),$$

if the Betas on the left are independent (for any positive numbers a, b, d).

✎ 3.7.11 Prove Proposition 3.7.10.

The Beta distribution arises very frequently as the *conjugate prior* for Binomial data, which means if we assume a Beta distribution for a parameter p before observing $\text{Bin}(n, p)$ data (this is the *prior distribution* for p), and then observe the data and update the distribution for p (this is the *posterior distribution* for p), we still have a Beta distribution: if $p \sim \text{Beta}(a, b)$ (the *prior*) and $X|p \sim \text{Bin}(n, p)$, then $p|X \sim \text{Beta}(a + X, b + n - X)$ (the *posterior*). (See Chapter 5 for an introduction to conditional distribution statements such as $X|p \sim \text{Bin}(n, p)$; meanwhile, think of this as saying what the distribution of X is if p is treated as a constant.)

Another Beta-Binomial connection comes through their CDFs as in the following result, whose proof is requested in Exercise 3.16. This is an example of a *shared distribution* result, connecting a discrete distribution to a continuous distribution (we will see another important example of this in the next section).

Proposition 3.7.12 *Let $B \sim \text{Beta}(j, n - j + 1)$ and $X \sim \text{Bin}(n, p)$, where j and n are positive integers with $j \leq n$. Then*

$$P(B \leq p) = P(X \geq j).$$

Story 3.7.13 (Bayes' Billiards) In the original paper in which Thomas Bayes proved Bayes' Theorem, he gave a remarkably beautiful derivation of a beta integral. This is a rare example where a complicated-looking integral can be computed just by telling a story, without doing any calculus! The integral in question is $\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp$, where $0 < p < 1$ and k, n are integers with $0 \leq k \leq n$. Suppose that we start with $n+1$ red billiard balls, positioned on the interval $[0,1]$ (treat the balls as points). Here are two possible generative stories:

Story 1: paint one of the balls green, and then randomly throw all the balls onto the interval $[0,1]$ (independently picking a Uniform position for each).

Story 2: randomly throw all the balls onto the interval $[0,1]$ (independently picking a Uniform position for each). Then, pick a random ball (with equal probabilities) and paint it green.

Let X be the number of red balls to the left of the green ball. Note that the distribution of X does not depend on whether Story 1 or Story 2 was used to generate the configuration of balls! For Story 1,

$$P(X = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp$$

by conditioning on p, the location of the green ball (see Chapter 5). For Story 2,

$$P(X = k) = \frac{1}{n+1},$$

since each of the balls is equally likely to be painted green. Thus,

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \frac{1}{n+1},$$

which agrees with the result for the Beta$(k+1, n-k+1)$ normalizing constant.

The Beta distribution is also closely related to the Uniform, beyond just being a generalization.

Proposition 3.7.14 *Let* $U \sim \text{Unif}$ *and* $\alpha > 0$. *Then*

$$U^{1/\alpha} \sim \text{Beta}(\alpha, 1),$$

and in particular, $U \sim \text{Beta}(1,1)$. *The Beta also arises trigonometrically:*

$$\sin^2(2\pi U) \sim \text{Beta}(1/2, 1/2).$$

✎ 3.7.15 Verify Proposition 3.7.14; one approach is to use the PIT for the first part and a geometric argument for the second part, considering a right triangle with legs of length Z_1, Z_2 which are i.i.d. $\mathcal{N}(0,1)$.

Definition 3.7.16 The Beta$(1/2, 1/2)$ distribution is famous enough in its own right to have its own name: the *Arcsine distribution.* This distribution often appears in engineering (e.g., in studying periodic signals), in statistics as a convenient U-shaped prior, and in random walks (e.g., if a fair coin is flipped $2n$ times with n large, Feller [2] shows that the distribution of the *last* time when the number of heads equals the number of tails is, after suitable scaling, approximately Arcsine).

✎ 3.7.17 Show that the CDF of the Arcsine distribution is

$$F(x) = \frac{2}{\pi}\text{asin}(\sqrt{x})$$

for $0 < x < 1$ (and is 0 for $x \leq 0$ and 1 for $x \geq 1$), thus explaining the name. Show that this also can be written as $F(p) = \frac{1}{\pi}\text{asin}(p - q) + \frac{1}{2}$, where $q \equiv 1 - p$; this expression is simpler to compute and makes the symmetry about $1/2$ clearer. Sketch the density function.

✎ 3.7.18 Show that if $C \sim$ Cauchy, S is a random sign, and $B \sim \text{Beta}(1/2, 1/2)$, then

$$C + \frac{1}{C} \sim 2S\sqrt{1 + C^2} \sim \frac{2S}{\sqrt{B}}.$$

3.8 Poisson Distribution

And sometimes remembering will lead to a story, which makes it forever. That's what stories are for.
– Tim O'Brien

We introduce the Poisson distribution via what is known as a *Poisson process*, which is defined in terms of a story involving Exponential r.v.s.

Story 3.8.1 Consider the following simple model for a sequence of events such as cars passing by a certain point, radioactive decay of an unstable compound, phone calls being received, or customers arriving at a store. Each occurrence is called an *arrival*, and we wish to study how the arrivals are distributed over time. Suppose that the times between successive arrivals are i.i.d. Exponentials with some rate λ (e.g., $\lambda = 5$ could correspond to 5 customers per hour arriving on average). Then λt is the expected number of events that will occur in time t, and $1/\lambda$ is the expected time from one event to the next. Such a process is known as a *Poisson process*, and reveals a deep connection between the Exponential distribution and the Poisson distribution (defined below).

Definition 3.8.2 (Poisson) Let $0 < T_1 < T_2 < \dots$ be "arrival times" such that the differences $T_1, T_2 - T_1, T_3 - T_2, \dots$ are i.i.d. $\sim \lambda^{-1}$Expo. Let $N_t \equiv \max\{n : T_n \leq t\}$ be the number of arrivals that have occurred up until time t, for every $t > 0$. Then N_t has the *Poisson* distribution with parameter λt, and we write $N_t \sim \text{Pois}(\lambda t)$.

In the above definition, we have two processes of interest: $N_t, t \geq 0$ consists of discrete random variables in continuous time, while $T_n, n \in \{1, 2, 3, \dots\}$ consists of continuous random variables in discrete time. There is an important duality between these two processes, which not only connects the two processes but also connects a fundamental continuous distribution (the Gamma) with a fundamental discrete distribution (the Poisson), via *shared distributions.* This duality is elegant, fundamental, and often-used, but does not seem to have a standard name; we call it the *Count-Time Duality.*

Theorem 3.8.3 (Count-Time Duality) *With notation as above, the following two events are identical:*

$$\{N_t \geq n\} = \{T_n \leq t\}.$$

✎ 3.8.4 Prove Theorem 3.8.3. Does it remain true if the inequalities are replaced by strict inequalities?

By definition, in the above we have $T_n \sim \lambda^{-1}\text{Gamma}(n)$. Using the Count-Time Duality and properties of the Gamma distribution, we can obtain the density of a Poisson r.v.

Proposition 3.8.5 *Let* $N \sim \text{Pois}(\lambda)$*. Then* $P(N = k) = e^{-\lambda}\lambda^k/k!$ *for* $k \in \{0, 1, 2, \dots\}$*.*

Proof For $t > 0$, let $N_t \sim \text{Pois}(t)$. By the Count-Time Duality with $\lambda = 1$, $N_t \geq k$ is equivalent to $T_k \leq t$, where $T_0 \equiv 0$, $T_k \sim \text{Gamma}(k)$ for $k \in \{1, 2, \dots\}$, and the T_k are increasing in k. Then

$$P(N_t = k) = P(T_k \leq t < T_{k+1}) = P(T_k \leq t) - P(T_{k+1} \leq t)$$

is the difference of two Gamma CDFs. This simplifies nicely using integration by parts on the term on the right, together with the fact that $\Gamma(k + 1) = k! = k\Gamma(k)$:

$$P(N_t = k) = \frac{1}{\Gamma(k)} \int_0^t e^{-x} x^k \frac{dx}{x} - \frac{1}{\Gamma(k+1)} \int_0^t e^{-x} x^{k+1} \frac{dx}{x} = e^{-t}\frac{t^k}{k!}.$$

□

3.9 Geometric and Negative Binomial

Just as the Exponential distribution is characterized by memorylessness among continuous distributions, the *Geometric distribution* is characterized by memorylessness among discrete distributions. The Geometric distribution is defined by representation as follows.

Definition 3.9.1 (Geometric) Let $G = \lfloor X \rfloor$, where $X \sim \lambda^{-1}\text{Expo}$. Then we say that G is *Geometric* with parameter $p \equiv 1 - e^{-\lambda}$, and write $G \sim \text{Geom}(p)$.

☣ 3.9.2 In the literature, several different definitions are used (having the values start at 1 rather than at 0, or reversing the roles of p and $1 - p$.

✎ 3.9.3 Show that $G \sim \text{Geom}(p)$ has PMF $P(G = k) = q^k p$, where $q \equiv 1 - p$. Thus, G can be interpreted as the number of failures before the first success in Bernoulli trials with p the probability of success.

Definition 3.9.4 (Negative Binomial) For r a positive integer, we define the *Negative Binomial* distribution with convolution parameter r and success probability p to be the distribution of $X = \sum_{j=1}^{r} G_j$, with the G_j i.i.d. Geom(p). We then write $X \sim \text{NBin}(r, p)$, and interpret X as the number of failures before r successes have been obtained in Bernoulli trials, with p the probability of success.

✎ 3.9.5 Show that the NBin(r, p) PMF is $P(X = x) = \binom{r+x-1}{x} p^r q^x$, where $q \equiv 1 - p$.

Remark 3.9.6 The name *Negative Binomial* can be explained by considering a binomial expansion of a negative power of a sum, such as $(a+b)^{-m}$.

✎ 3.9.7 The coefficient $\binom{r+x-1}{x}$ is the number of ways to distribute x indistinguishable particles into r boxes (this is known as a *Bose-Einstein* value). Explain this, showing how these two problems (balls in boxes and waiting for the rth success) are related.

The Negative Binomial distribution is *infinitely divisible*, which means that for any n, the distribution can be obtained as the sum (convolution) of n i.i.d. random variables. Infinite divisibility enables us to extend NBin to allow any real convolution parameter $r > 0$. In the above density, we can extend the definition of the coefficient by interpreting factorials via the gamma function.

3.10 Symmetry Representation

Definition 3.10.1 A random variable Y is *symmetric* (about 0) if $Y \sim -Y$.

Theorem 3.10.2 *Any symmetric random variable Y can be represented as $Y = SA$, with $A \geq 0$ and S a random sign independent of A.*

✎ 3.10.3 Show that if Y is symmetric, then so is YW for *any* r.v. $W \perp\!\!\!\perp Y$. In particular, A in Theorem 3.10.2 need not be required to be positive. Also show that any linear combination of independent symmetric r.v.s is symmetric.

Proposition 3.10.4 *Let Y_1, Y_2 be i.i.d. symmetric r.v.s. Then*

$$|\min(Y_1, Y_2)| \sim |Y_1| \sim |\max(Y_1, Y_2)|.$$

Proof Let $M \equiv \min(Y_1, Y_2)$, and condition on which of Y_1, Y_2 is smaller (this is slightly easier if Y_1 is continuous, so that $P(Y_1 = Y_2) = 0$, but it is not necessary to assume this). We have

$$\begin{aligned} P(|M| \leq y) &= P(|M| \leq y | Y_1 \leq Y_2) P(Y_1 \leq Y_2) + P(|M| \leq y | Y_1 > Y_2) P(Y_1 > Y_2) \\ &= P(|Y_1| \leq y | Y_1 \leq Y_2) P(Y_1 \leq Y_2) + P(|Y_2| \leq y | Y_1 > Y_2) P(Y_1 > Y_2). \end{aligned}$$

By symmetry, $P(|Y_2| \leq y | Y_1 > Y_2) = P(|-Y_2| \leq y | -Y_1 > -Y_2) = P(|Y_2| \leq y | Y_1 < Y_2)$. Since Y_1 and Y_2 are i.i.d., this in turn equals $P(|Y_1| \leq y | Y_2 < Y_1)$. So

$$\begin{aligned} P(|M| \leq y) &= P(|Y_1| \leq y | Y_1 \leq Y_2) P(Y_1 \leq Y_2) + P(|Y_2| \leq y | Y_1 > Y_2) P(Y_1 > Y_2) \\ &= P(|Y_1| \leq y | Y_1 \leq Y_2) P(Y_1 \leq Y_2) + P(|Y_1| \leq y | Y_2 < Y_1) P(Y_2 < Y_1) \\ &= P(|Y_1| \leq y). \end{aligned}$$

Similarly, we also have $|\max(Y_1, Y_2)| \sim |Y_1|$. □

3.11 Order Statistics and the Rényi Representation

Definition 3.11.1 The *order statistics* of r.v.s $Y_1, \dots, Y_n$ are the sorted list of the Y_j, denoted by $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. For example, $Y_{(1)} = \min(Y_1, \dots, Y_n), Y_{(n)} = \max(Y_1, \dots, Y_n)$, and for n odd $Y_{((n+1)/2)}$ is the median of $Y_1, \dots, Y_n$.

3.11.1 Exponentials and Order Statistics

Order statistics are particularly pleasant with i.i.d. Exponentials, so we begin by considering that case.

✎ 3.11.2 Suppose that $Y_1, \dots, Y_n \sim$ Expo are i.i.d. Show that the minimum is also Exponentially distributed, with n times the rate: $Y_{(1)} \sim \frac{1}{n}\text{Expo}$ (the rate parameter is defined to be the reciprocal of the scale parameter).

The memoryless property is particularly important in studying survival times, which often arises in biostatistics and reliability analysis. Consider two appliances, say a stove and a refrigerator, whose survival times are independent (scaled) Expos with different rates. At time T, one of the two appliances fails, with the other still working. What information does T convey about *which* appliance fails? It turns out that the answer is *none.*

Theorem 3.11.3 (Competing Risks Theorem) *Let $Y_1 = X_1/\lambda_1$ and $Y_2 = X_2/\lambda_2$ be independent (scaled) Exponentials, with $X_1, X_2 \sim$ Expo and $\lambda_1, \lambda_2 > 0$ constants. Define*

$$W \equiv \min(Y_1, Y_2) \text{ and } B_0 \equiv I_{Y_1 < Y_2},$$

where I_A is the indicator random variable for an event A. Then $W \perp\!\!\!\perp B_0$.

The following Lemma is useful for proving the Competing Risks Theorem.

Lemma 3.11.4 *Let $0 < p < 1$ be a constant and $U \sim$ Unif. Define*

$$B \equiv I_{U \leq p} \text{ and } M \equiv \min\left(\frac{U}{p}, \frac{1-U}{1-p}\right).$$

Then the indicator r.v. $B \sim$ Bern(p) is independent of $M \sim$ Unif.

The proofs of the above Theorem and Lemma are left as exercises at the end of the chapter, with hints.

✎ 3.11.5 Show (by representation) that if $Y_j \sim \lambda_j^{-1}\text{Expo}$, $1 \leq j \leq n$, are independent with (possibly) different rates, then we have

$$Y_{(1)} \sim \lambda^{-1}\text{Expo and } P(Y_{(1)} = Y_1) = \lambda_1/\lambda,$$

where $\lambda \equiv \lambda_1 + \cdots + \lambda_n$ is the total rate.

Hint: for the second part when $n = 3$, note that the event $Y_1 = \min(Y_1, Y_2, Y_3)$ is the same as the event $Y_1 \leq \min(Y_2, Y_3)$.

The *Rényi Representation* is a representation for *all* the order statistics of i.i.d. $Y_1, \dots, Y_n \sim$ Expo (jointly). Consider the difference $Y_{(2)} - Y_{(1)}$, the gap between the first and second order statistics. The memoryless property implies that

$$Y_{(2)} - Y_{(1)} \sim \frac{1}{n-1}\text{Expo and } Y_{(2)} - Y_{(1)} \perp\!\!\!\perp Y_{(1)};$$

in terms of the appliance story, note that at the first time at which an appliance fails, the other $n-1$ are "good as new", and the additional time needed for one of them to fail is the minimum of $n-1$ i.i.d. Expos (a more formal proof of this, conditioning on $Y_{(1)}$, is given in a little while). It then follows that Y_2 can be represented as

$$Y_{(2)} = Y_{(2)} - Y_{(1)} + Y_{(1)} \sim \frac{1}{n-1}X_1 + \frac{1}{n}X_2,$$

with X_1, X_2 i.i.d. Expo. This is the sum of independent Exponentials *with different scales.* Note that a Gamma would be obtained if they had the same scale, but the scales are not exactly the same. However, if n is large, then it follows from the above that $Y_{(2)}$ is approximately $\frac{1}{n}$Gamma(2) in distribution.

To get some intuition into the distribution of $Y_{(2)} - Y_{(1)}$, consider the following story. There are n companies, each of which independently receives phone calls all day long at a rate of one call per minute. Assume that the phone calls for each company follow a Poisson Process. That is, the number of calls in any time interval of length t is Pois(t), with the number of calls in disjoint intervals independent. It follows that the waiting times between consecutive phone calls are independent Exponentials.

Now consider the time of the *first* call to company i to be our Y_i. Then $Y_{(1)}$ is the time of the first call received by *any* company. This is distributed Exponentially with rate n. Now consider $Y_{(2)} - Y_{(1)}$, which is the additional waiting time until the "next first call". The company that receieved the first call of the day can no longer receive its "first call" so there are $n-1$ companies that have yet to receive a call. Now remember that the Exponential distribution is memoryless! For any company j other than the one that had its first call already, $Y_j - Y_{(1)}$ is itself Exponential. So recursively we are in the same situation as before with $n-1$ companies instead of n, and by the memoryless property $Y_{(2)} - Y_{(1)}$ is Exponential with rate $n-1$, independent of $Y_{(1)}$.

To prove this by conditional probability, the idea is to *condition both on $Y_{(1)}$ and on which Y_i is $Y_{(1)}$* (the problem is simplified by knowing which specific Y_i is the minimum; by symmetry, $P(Y_{(1)} = Y_i) = 1/n$ for all $i \in \{1, \dots, n\}$). So

$$P(Y_{(2)} - Y_{(1)} > t | Y_{(1)}) = \frac{1}{n}\sum_{i=1}^{n} P(Y_{(2)} - Y_{(1)} > t | Y_{(1)}, Y_{(1)} = Y_i) = P(Y_{(2)} - Y_{(1)} > t | Y_{(1)}, Y_{(1)} = Y_1).$$

By the memoryless property and independence of the Y_j, for any a we have

$$\begin{aligned} P(Y_{(2)} - Y_{(1)} > t | Y_{(1)} = a, Y_{(1)} = Y_1) &= P(Y_2 - a > t, \dots, Y_n - a > t | Y_{(1)} = a, Y_{(1)} = Y_1) \\ &= P(Y_2 - a > t, \dots, Y_n - a > t | Y_1 = a, Y_2 > a, \dots, Y_n > a) \end{aligned}$$

$$= P(Y_2 - a > t | Y_2 > a) \cdots P(Y_n - Y_{(1)} > t | Y_n > a) = e^{-(n-1)t}.$$

Then $Y_{(2)} - Y_{(1)}$ is independent of $Y_{(1)}$ since its conditional distribution given $Y_{(1)}$ does not involve $Y_{(1)}$. We also have $Y_{(2)} - Y_{(1)} \sim (n-1)^{-1}\text{Expo}$, as claimed.

Continuing in this way, we have the *Rényi Representation.*

Theorem 3.11.6 (Rényi Representation) *For $Y_1, \dots, Y_n$ i.i.d.* Expo*, the order statistics $(Y_{(1)}, \dots, Y_{(n)})$ can be* jointly *represented as*

$$Y_{(k)} \sim \sum_{j=1}^{k} \frac{1}{n-j+1} X_j,$$

where the X_j's are also i.i.d. Exponentials.

Note that in this representation the same X_j's can be used for all the $Y_{(k)}$'s, so the Rényi Representation can be used to study joint distributions for the order statistics, not just marginal distributions. A useful general approach to order statistics is to first reduce the problem to the order statistics of Exponentials (using the PIT), and then to apply the Rényi Representation.

3.11.2 Uniforms and Order Statistics

Now let $U_1, \dots, U_n$ be i.i.d. Uniform. Then $U_{(j)} \sim \text{Beta}(j, n-j+1)$.

In fact, we can *jointly* represent the order statistics of the U_j in terms of ratios of sums of Exponentials.

Theorem 3.11.7 *Let $U_1, \dots, U_n$ be i.i.d. Uniform and*

$$W_j = \frac{X_1 + \cdots + X_j}{X_1 + \cdots + X_{n+1}},$$

with $X_1, \dots, X_{n+1}$ i.i.d. Expo. Then we have the following joint *representation for the order statistics of the U_j:*

$$(U_{(1)}, \dots, U_{(n)}) \sim (W_1, \dots, W_n).$$

We will prove the above result (a joint representation for the order statistics of Uniform r.v.s) using the Rényi Representation (a joint representation for the order statistics of Exponential r.v.s).

Proof Let

$$V_j = -\log U_j \sim \text{Expo}.$$

By the Rényi Representation, we can *jointly* represent

$$\begin{aligned} V_{(1)} &\sim \frac{1}{n}Y_1, \\ V_{(2)} &\sim \frac{1}{n}Y_1 + \frac{1}{n-1}Y_2, \\ &\dots \\ V_{(n)} &\sim \frac{1}{n}Y_1 + \frac{1}{n-1}Y_2 + \cdots + Y_n, \end{aligned}$$

with $Y_1, \dots, Y_n$ i.i.d. Expo. Since the transformation from U_j to V_j is strictly decreasing, the order is *reversed*:

$$V_{(n-j+1)} = -\log U_{(j)}.$$

So

$$U_{(j)} = \exp\left(-V_{(n-j+1)}\right) \sim B_1 B_2 \dots B_{n-j+1},$$

with

$$B_k = \exp\left(-\frac{Y_k}{n-k+1}\right) \sim \text{Beta}(n-k+1, 1),$$

using the facts that $\exp(-Y_k) \sim U_1$ and $U_1^{1/a} \sim \text{Beta}(a, 1)$. The B_k are independent since the Y_k are independent. Using the same method as in Proposition 3.7, we can represent the B_k as

$$B_k \sim \frac{X_1 + \cdots + X_{n-k+1}}{X_1 + \cdots + X_{n-k+2}},$$

with the right-hand sides independent across k (so we have a joint representation for the B_k. The desired result then appears through a telescoping product:

$$U_{(j)} \sim \frac{X_1 + \cdots + X_n}{X_1 + \cdots + X_{n+1}} \cdot \frac{X_1 + \cdots + X_{n-1}}{X_1 + \cdots + X_n} \cdots \cdot \frac{X_1 + \cdots + X_j}{X_1 + \cdots + X_{j+1}} = \frac{X_1 + \cdots + X_j}{X_1 + \cdots + X_{n+1}}.$$

□

Exercises

3.1 (**World series**) Two baseball teams (A and B) are playing a best-of-7 series of games. Let p be the probability of A winning an individual game and the results of games played be independent. In terms of p and a Binomial CDF, what is the probability that A wins the match? Explain why (for this calculation) one can assume that any unplayed games are in fact played, even though once a team reaches 4 wins no further games are played (because the outcome has been determined).

3.2 (**Symmetry with fair coins and one-apart sample sizes**) Let $X \sim \text{Bin}(n, 1/2)$ and $Y \sim \text{Bin}(n+1, 1/2)$ be independent. Show that $P(X < Y) = 1/2$.

Hint: why does $P(X < Y) = P(n - X < n + 1 - Y)$?

3.3 (**Odd Binomial**) Let $X \sim \text{Bin}(n, 1/2)$. Find the probability that X is odd.

Hint: one approach is to condition on the first $n-1$ trials.

3.4 (**Parity check code**) A binary message is sent over a noisy channel. The message is a sequence $x_1, x_2, \dots, x_n$ of n bits ($x_i \in \{0, 1\}$). Since the channel is noisy, there is a chance that any bit might be corrupted, resulting in an error (a 0 becomes a 1 or vice versa). Assume that the error events are independent. Let p be the probability that an individual bit has an error ($0 < p < 1/2$). Let $y_1, y_2, \dots, y_n$ be the received message.

To help detect errors, the nth bit is reserved for a parity check: x_n is defined to be 0 if $x_1 + x_2 + \cdots + x_{n-1}$ is even, and 1 if $x_1 + x_2 + \cdots + x_{n-1}$ is odd. When the message is received, the recipient checks whether y_n has the same parity as $y_1 + y_2 + \cdots + y_{n-1}$. If the parity is wrong, the recipient knows that at least one error occurred; otherwise, the recipient assumes that there were no errors. Find the probability of undetected errors (simplify).
Hint: use the binomial theorem, along with the following trick. If a is a sum over even k and b is a sum over odd k, then consider $a + b$ and $a - b$ in solving for a and b.

3.5 (t **and Beta**) Show that if $T \sim t_m$, then $\frac{1}{m}T^2 \sim \frac{B}{1-B}$, with $B \sim \text{Beta}(1/2, m/2)$.

3.6 (**Fisher's method for combining p-values**) Consider a hypothesis test, where a "test statistic" T is computed, and the null hypothesis is rejected if the observed value is deemed extreme under the null hypothesis. Assume that T is continuous, and the test under consideration rejects the null hypothesis if T exceeds a certain threshold value. The *p-value* is defined to be $P_0(T \geq t_{obs})$, where t_{obs} is the observed value of T and the subscripted 0 indicates that the probability is with respect to the distribution implied by the null hypothesis.

(a) Show that if the null hypothesis is true, then the p-value (viewed as a r.v.) is Uniform.

(b) Suppose we have several p-values, obtained from independent experiments testing the same hypothesis. It seems that several independent results that moderately suggest rejecting the null can be at least as convincing as one result strongly suggesting rejecting the null, but it is not obvious how to combine the p-values into one overall value (this is a problem of *meta-analysis*).
R.A. Fisher proposed the following method for combining the p-values, say $p_1, \dots, p_n$. Take $R \equiv -2\log(p_1 p_2 \cdots p_n)$, and let the "combined p-value" be $P(\chi^2_{2n} \geq r_{obs})$, where r_{obs} is the observed value of R. Explain why the χ^2_{2n} distribution is used here.

3.7 (**Laplace and Logistic Representations**) Let X_1, X_2 be i.i.d. Expo. Prove the following representations (without calculus):
(a) The difference $X_1 - X_2 \sim$ Laplace.

(b) The difference of logarithms $\log X_1 - \log X_2 \sim$ Logistic.

3.8 (**Three Gammas**) Let G_1, G_2, G_3 be independent, $G_i \sim \text{Gamma}(a_i)$, with a_1, a_2, a_3 fixed positive constants. Consider

$$B_1 \equiv \frac{G_1}{G_1 + G_2}, B_2 \equiv \frac{G_1 + G_2}{G_1 + G_2 + G_3}, \text{ and } S \equiv G_1 + G_2 + G_3.$$

Prove that B_1, B_2, and S are fully independent, and identify their three marginal distributions (they will be Betas and Gammas), by the following two methods:

(a) using the Jacobian method and densities.

(b) by representation (using the fact that B_1 and $G_1 + G_2$ are independent with known distributions).

3.9 ($\Gamma(1/2)$ **via** Beta$(1/2, 1/2)$) Use the expression for the normalizing constant of the Beta$(1/2, 1/2)$ distribution in terms of the gamma function to obtain another proof that $\Gamma(1/2) = \sqrt{\pi}$ (and thus another derivation of the Normal normalizing constant).
Hint: the integral $\int \frac{1}{\sqrt{p(1-p)}} dp = \text{asin}(p - q) + C$ is useful, where $0 < p < 1$ and $q \equiv 1 - p$.

3.10 (**Product of Uniforms vs. Squared Uniform**) Let U_1, U_2, U_3 be i.i.d. $\sim$ Unif. Show that

$$P(U_1 U_2 < U_3^2) = \frac{5}{9}.$$

Hint: take logs.

3.11 (**Uniform Power of Uniform Product**) Let U_1, U_2, U_3 be i.i.d. $\sim$ Unif. Show that

$$(U_1 U_2)^{U_3} \sim U_1.$$

3.12 (**Uniform length-respecting property**) Let U be a r.v. supported on $[0, 1]$ which *respects lengths*, in the sense that if I_1, I_2 are subintervals of $[0, 1]$ with length$(I_1) \leq$ length(I_2), then $P(U \in I_1) \leq P(U \in I_2)$. Show that $U \sim$ Unif.

3.13 (**Cauchy shift of Cauchy**) Let Y have a Cauchy distribution centered at θ, so the density is

$$f(y|\theta) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2}.$$

Suppose that θ has a Cauchy distribution (centered at 0). Find the marginal distribution of Y.

3.14 (**Cauchy squared**) Let $C \sim$ Cauchy. Show that $C^2 + \frac{1}{C^2} \sim 4C^2 + 2$.

3.15 (**Weibull hazard**) Let $W \sim$ Wei(β) with $\beta > 1$. Show that, in contrast to the memoryless property of Exponentials, for any positive a, b we have $P(W \geq a + b | W \geq a) > P(W \geq b)$.

3.16 (**Beta CDF**) Prove Proposition 3.7.12.

Hint: consider order statistics.

3.17 **(Competing Risks Theorem)** In this problem, you will prove the Competing Risks Theorem 3.11.3 in several parts. Let the setup and notation be as in the statement of the theorem, and let $U_0 \equiv X_1/(X_1 + X_2)$ and $T \equiv X_1 + X_2$.

(a) Show that $W \sim (\lambda_1 + \lambda_2)^{-1}$Expo and $B_0 \sim$ Bern$(\frac{\lambda_1}{\lambda_1 + \lambda_2})$. Hint for the latter: use what you know about U_0 rather than an integral!

(b) Prove the Competing Risks Theorem, assuming Lemma 3.11.4.

(c) Prove Lemma 3.11.4. Hint: compute $P(U \leq p | M \geq m)$.

3.18 (**Point in the unit disk**) Let (X, Y) be a uniformly distributed point in the unit disk in $\mathbb{R}^2$.

(a) Prove the following representations:

$$X \sim \frac{t_3}{\sqrt{3+t_3^2}}, \text{ and } X^2 \sim \text{Beta}(0.5, 1.5).$$

(b) Prove that

$$Y|X \sim SU\sqrt{1-X^2},$$

with S a random sign and $U \sim$ Unif, independently.

(c) Using the result in (b), the fact that $U^2 \sim \text{Beta}(0.5, 1)$ (check this), and Proposition 3.7.10, show that $Y^2 \sim \text{Beta}(0.5, 1.5)$ (so we have $Y^2 \sim X^2$).

3.19 (**Geometric order statistics**) Let $G_1, \dots, G_n \sim \text{Geom}(p)$ be i.i.d. and $p = 1 - e^{-\lambda}$. Show that $G_{(j)} \sim \lfloor X_{(j)} \rfloor$, with $X_1, \dots, X_n$ i.i.d. $\lambda^{-1} \cdot$ Expo.

3.20 (**Two Exponentials**) Let X_1 and X_2 be i.i.d. Expo. Let $X_{(1)} = \min(X_1, X_2)$ and $X_{(2)} = \max(X_1, X_2)$.

(a) Find the conditional distribution of X_1 given $X_1 + X_2$.

(b) Find the distribution of $\frac{X_{(1)}}{X_{(1)}+X_{(2)}}$.

3.21 (**Product of Uniform Order Statistics from Two Samples**) Let $U_1, \dots, U_n, V_1, \dots, V_n$ be i.i.d. Unif. For $1 \leq j \leq m$, write $U_{(j),m}$ for the jth order statistic among $U_1, \dots, U_m$, and define $V_{(j),m}$ similarly. Show that for $1 \leq j \leq m \leq n-1$, $U_{(j),m}V_{(m+1),n} \sim U_{(j),n}$.