

Notes for MATH 231: Probability Theory

Instructor: Han Li
Notes by: Yulia Alexandr

Contents

1	Preliminaries and Basic Concepts	2
1.1	Motivation	2
1.2	Basic Set Theory	5
1.3	Sequences and Series	7
1.4	σ -Algebras and Probability Spaces	12
1.5	The Borel σ -Algebra	17
1.6	One-Dimensional Probability Distributions	21
1.7	Conditional Probability	26
2	Discrete and Continuous Random Variables	32
2.1	Introduction to Random Variables	32
2.2	Continuous Probability Distributions	41
2.3	Convolution of Probability Density Functions	48
2.4	Product Measure	52
3	Random Vectors	53
3.1	Joint Distribution of Random Variables	53
3.2	Discrete Random Vectors	56
3.3	Independence of Random Variables	60
3.4	Jointly Continuous Random Variables	65
3.5	Quotients of Random Variables	70
3.6	F and T Distributions	72
4	Limit Theorems	76
4.1	Poisson Distribution	76
4.2	Poisson Limit Theorem	81
4.3	Local Limit Theorem	85
4.4	Central Limit Theorem	92
5	Markov Chains (Optional)	95

1 Preliminaries and Basic Concepts

1.1 Motivation

We will start by introducing several examples meant to help you develop some intuition about what probability theory studies and get a feeling of how complicated things become quite fast. Some examples will be as simple and familiar to you as a coin flip, and some will be more abstract, but nevertheless vastly important in mathematics.

Example 1. Suppose we flip a coin twice. It isn't hard to see that the possible outcomes (respecting the order) are: 2 heads, 2 tails, 1 head and 1 tail, 1 tail and 1 head. Mathematically, we will write it as a set $\{HH, TT, HT, TH\}$ where H stands for heads and T stands for tails. If the coin is fair, we all know that the probability we'll get either one of these combinations is $\frac{1}{4}$. If the coin is biased, the probability distribution may be skewed, for example, $\mathbb{P}(HH) = 0.1, \mathbb{P}(TT) = 0.2, \mathbb{P}(HT) = 0.3, \mathbb{P}(TH) = 0.4$ where \mathbb{P} denotes the probability of a certain configuration showing up. (Also note that all probabilities add up to 1: this is not a coincidence!)

This was the first mathematical model among many other we'll see over the course of this semester. We will usually denote the set of all possible outcomes as Ω . So, in the previous example, $\Omega = \{HH, TT, HT, TH\}$.

Intuitively, we can think of probability as the "weight" we assign to each $\omega \in \Omega$. For example, if $\Omega = \{\omega_1, \omega_2, \dots\}$ then each ω_i will have a certain weight P_i such that $\sum_{i=1}^{\infty} P_i = 1$. If we let $A = \{\omega_{k_1}, \dots, \omega_{k_m}\} \subseteq \Omega$, then $\mathbb{P}(A) = \mathbb{P}_{k_1} + \dots + \mathbb{P}_{k_m}$.

Example 2. Suppose now that we have an infinite sequence of coin flips. Then we ask: what is the probability that 3 consecutive heads occur infinitely many times? Let's start from the beginning. Can we describe all the possible situations as in the previous example? The answer is yes and here they are:

$$\Omega = \{(a_n)_{n=1}^{\infty} : a_n \in \{0, 1\}\}$$

where 1 is equivalent to T and 0 is equivalent to H . That is, our set of possibilities is the set of all possible infinite binary strings. Clearly, it is an infinite set; moreover, it's uncountable (meaning it is NOT in bijection with \mathbb{Z}^+ !) Perhaps, it's time to recall Cantor's Theorem:

Theorem 1. *Let X be any set. Let 2^X be its power set. Then there is no bijection between X and 2^X .*

You may be asking yourself: how is this theorem relevant to our example? Turns out, it is. Note that we can associate Ω to the power set of positive integers, $2^{\mathbb{Z}^+}$. Given any infinite binary string, we can think of a 1 in the i th place as having $i \in \mathbb{Z}^+$ in the subset of \mathbb{Z}^+ we're constructing and similarly we can think of a 0 in the j th place as not including j in the same subset. Thus, each string has a (unique) subset associated to it. The other direction is now

obvious: for any subset A of positive integers, construct an infinite binary string by putting a 1 in the i th spot for each $i \in A$ and putting a 0 in the j th spot for each $j \notin A$. Clearly, the two functions are inverses, and thus we have a bijection.

This example shows that the setup we had in Example 1.1 is insufficient to build probability theory in full generality: we need more tools than we have to deal with uncountable sets. Thus, we will come back to this problem once we introduce all of those tools and answer it later in the semester. Stay tuned!

Example 3. Arbitrarily pick a continuous function $f : \mathbb{R} \rightarrow \mathbb{R}$. What is the probability that $f(x) > 40$ for all $x \in (0, 1)$? (This example may be of use to those of you interested in finance, since this kind of model can be used to describe a behavior of financial markets.)

The main point of the above examples is that a "model" is needed to define probability and probability of events depends on the model defined.

Example 4. Arbitrarily pick a point $x \in [0, 1]$. Let $A \subseteq [0, 1]$. What's the probability that $x \in A$? We don't have any rigorous tools to deal with this problem yet, but let's think of it intuitively. What if $A = [\frac{1}{2}, 1]$? Then, we are inclined to say that the probability we are interested in is simply $\frac{1}{2}$. Similarly, if $A = [a, b]$ where $0 \leq a \leq b \leq 1$, then we may say that the probability is $b - a$. This makes sense to us because, for now, we think of probability as a proportion or length.

So let's naively define the probability in this example as $\ell(A)$, the "length" of A . If A is an interval from a to b (no matter open or closed), the length is naturally defined as $b - a$. But what if A is arbitrary? So far, it's not very clear.

Suppose we define our length function ℓ to satisfy the following five properties:

1. $\ell : 2^{[0,1]} \rightarrow [0, 1]$
2. $\ell(E) \leq \ell(F)$ whenever $E \subseteq F \subseteq [0, 1]$, i.e. ℓ is a non-decreasing function.
3. $\ell(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \ell(A_i)$, i.e. the function is σ -additive.
4. $\ell([a, b]) = b - a$ for all $[a, b] \subseteq [0, 1]$.
5. $\ell(x + E) = \ell(E)$ for all $E, x + E \subseteq [0, 1]$ where $x + E := \{x + y : y \in E\}$, i.e. ℓ is shift-invariant.

This looks quite promising, doesn't it? However, the next theorem may disappoint you a bit:

Theorem 2. *There does NOT exist a function ℓ satisfying the properties 1-5 above.*

Proof. We begin by defining a binary relation on $[\frac{1}{3}, \frac{2}{3}]$ as follows: $x \sim y \iff x - y \in \mathbb{Q}$. We will first show that \sim is an equivalence relation, i.e. it is reflexive, symmetric, and transitive. Indeed, for all $x, y, z \in [\frac{1}{3}, \frac{2}{3}]$ we have:

- $x \sim x$, since $x - x = 0 \in \mathbb{Q}$.
- $x \sim y \implies y \sim x$, since if $x - y = r \in \mathbb{Q}$, then $y - x = -r \in \mathbb{Q}$.
- $x - y \in \mathbb{Q}$ and $y - z \in \mathbb{Q} \implies (x - y) + (y - z) = x - z \in \mathbb{Q} \implies x \sim z$.

So indeed $x \sim z$ is an equivalence relation. Thus, by the Axiom of Choice (look it up!), we can construct set E by letting it contain exactly one element from each equivalence class of \sim .

Claim 1: $r_1 + E \cap r_2 + E = \emptyset$ for all $r_1 \neq r_2 \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}$.

Proof. Suppose, toward a contradiction, that $r_1 + x = r_2 + y$ for some $x, y \in E$. (WLOG, assume $r_2 > r_1$.) Then, equivalently, $x - y = r_2 - r_1 = r_3 \in \mathbb{Q}$. This implies that x and y are in the same equivalence class, but since, by our initial assumption, they're both in E , this is a contradiction. From the first claim, we deduce that $\biguplus_{r \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}} r + E$ is a disjoint union. \square

Claim 2: $[\frac{1}{3}, \frac{2}{3}] \subseteq \biguplus_{r \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}} r + E$.

Proof. Let $y \in [\frac{1}{3}, \frac{2}{3}]$. Then there exists $x \in E$ s.t. $y - x = r \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}$ by the definition of E . This implies that $y = r + x \in r + E \subseteq \biguplus_{r \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}} r + E$, as desired. \square

Claim 3: $\biguplus_{r \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}} r + E \subseteq [0, 1]$.

Proof. Let $r \in [-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}$ and $x \in E \subseteq [\frac{1}{3}, \frac{2}{3}]$ be arbitrary. Then $r + x \in [0, 1]$. Hence, $r + E \subseteq [0, 1] \implies \biguplus_r r + E \subseteq [0, 1]$. \square

Now we are ready to finish up the proof. Combining all properties and claims we obtain the following:

$$\begin{aligned}
\frac{1}{3} = \ell\left(\left[\frac{1}{3}, \frac{2}{3}\right]\right) &\leq \ell\left(\biguplus_r r + E\right) \\
&= \sum_r \ell(r + E) \\
&= \sum_r \ell(E) \\
&\leq \ell([0, 1]) = 1
\end{aligned} \tag{1}$$

There is a problem with the above expression. Since $[-\frac{1}{3}, \frac{1}{3}] \cap \mathbb{Q}$ is infinite, the summation above is infinite as well. Thus, $\ell(E)$ is added infinitely many times. However, this summation

must be at least $\frac{1}{3}$, so $\ell(E) \neq 0$. Therefore, $\ell(E) > 0$. But also, the sum has to be at most 1, which is impossible, since there exists no positive real number you can add infinitely many times to get a finite number. Thus, we obtain the desired contradiction, and this completes the proof. \square

This is rather a fascinating result! So what went wrong? Were some of the properties imposing conditions that were too strong? Looking at the five properties again, 2,3,4, and 5 seem so natural, and, intuitively, we want all of them to be satisfied for the math to make sense. So, it must be the property 1 that causes trouble. It turns out, the first property is indeed too strong. That is, not all subsets of Ω can have probability or "length" assigned to them. (Using some jargon that you'll see in the next sections, not all subsets of $[0,1]$ are measurable – clearly, E isn't!) So, if we want to study probability properly, we need to define it for a class of "interesting" subsets. This is what makes probability theory hard (but also interesting.)

Example 5. Suppose we're given a unit circle and an equilateral triangle inscribed in it. We define a chord to be a line segment with both endpoints lying on the circle. Note that doing simple calculations shows that each side of the triangle has length $\sqrt{3}$. Then we ask the following question: if we randomly pick a chord on the circle, what is the probability that that chord has length greater than $\sqrt{3}$?

Solution 1: WLOG, pick any point on the circle and let it be the first endpoint of our chord. Then we can draw the equilateral triangle inside the circle one of whose endpoints is a node of that triangle. Then, the second endpoint must land on the arc defined by the opposite side of the triangle for the chord to be greater than $\sqrt{3}$. That arc is exactly $\frac{1}{3}$ of the circumference of the circle, so the probability is $\frac{1}{3}$.

Solution 2: Note that a chord is completely determined by its midpoint. This midpoint must lie on some radius (some line segment connecting the center to the boundary). Fix a radius and draw all chords whose midpoint lies on that radius. Exactly half of them are longer than $\sqrt{3}$ – those lying above the chord passing through the middle of the radius. Thus, the probability must be $\frac{1}{2}$.

What did just happen? Did we get two different numbers as solutions to the same exact problem?! What we saw just now is called Bertrand Paradox (look it up!) The way mathematicians resolve it is by specifying "randomness" in each case.

1.2 Basic Set Theory

In this section, we'll revisit some of the basic definitions and notation you are probably already familiar with but which will be used in this course all the time.

Let I be an indexing set (possibly uncountable). Let A_i be a set for all $i \in I$. Then we define a union and an intersection as follows.

Definition 1.

$$\bigcup_{i \in I} A_i := \{x : \exists i \in I \text{ s.t. } x \in A_i\}$$

$$\bigcap_{i \in I} A_i := \{x : x \in A_i \forall i \in I\}$$

Example 6. For example, $\mathbb{R} = \bigcup_{i \in \mathbb{Z}} \left[i - \frac{1}{2}, i + \frac{1}{2}\right]$, which, according to the definition above, means that $x \in \mathbb{R} \iff \exists i \in \mathbb{Z} \text{ s.t. } x \in \left[i - \frac{1}{2}, i + \frac{1}{2}\right]$. In other words, every real number is at most $\frac{1}{2}$ away from some integer.

Example 7. Another example would be $\mathbb{R} = \bigcup_{x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)} \{\tan x\}$.

Example 8. To see how the intersection works, observe that:

$$\{0\} = \bigcap_{i \in \mathbb{Z}^+} \left(-\frac{1}{i}, \frac{1}{i}\right)$$

meaning that $x = 0 \iff \forall i \in \mathbb{Z}, -\frac{1}{i} < x < \frac{1}{i}$, which is obviously true.

Convention 1. Let $I = \{m, m+1, m+2, \dots\}, m \in \mathbb{Z}$. Then we will write $\bigcup_{i \in I} A_i$ as $\bigcup_{i=m}^{\infty} A_i$. We will do the same for intersection.

Definition 1. Let $\{A_i\}$ be a family of sets. We say $\{A_i\}_{i \in I}$ is pairwise disjoint if for all $i, j \in I, i \neq j$, we have $A_i \cap A_j = \emptyset$.

Convention 2. 1.11 If $\{A_i\}_{i \in I}$ is pairwise disjoint, then $\bigcup_{i \in I} A_i$ will be written as $\biguplus_{i \in I} A_i$ to emphasize disjointness. Similarly for $\biguplus_{i=m}^{\infty} A_i$.

Example 9. 1.12 Let us give a classic example of a disjoint union. Let X be a set and R be an equivalence on X , meaning that $R \subseteq X \times X$ s.t.:

- $(x, x) \in R$ for all $x \in X$.
- $(x, y) \in R \implies (y, x) \in R$.
- $(x, y) \in R$ and $(y, z) \in R \implies (x, z) \in R$.

Notationally, we will write $x \sim y \iff (x, y) \in R$. Now, recall that X/R , the set of all equivalence classes, is $\{[x] : x \in X\}$ where $[x] = \{y \in X : (x, y) \in R\}$. We also remember from Discrete Math that $X = \biguplus_{[x] \in X/R} [x]$, i.e. that equivalence classes form a partition on X – every $x \in X$ belongs to exactly one and only one equivalence class. So, the disjoint union of equivalence classes forms the entire set X .

On to a concrete example: Let $X = \mathbb{Z}$, $R = \{(x, y) : 2|(x - y)\}$. (You should check R is indeed an equivalence relation.) The set of equivalence classes of R is $R/X = \{[0], [1]\}$, where $[0]$ is the set of all even integers and $[1]$ is the set of all odd integers. It is now obvious that $\mathbb{Z} = [0] \uplus [1]$.

Let us now recall some more set theory and talk about some of its fundamental laws.

Distributive Laws:

$$1. B \cap \left(\bigcup_{i \in I} A_i \right) = \bigcup_{i \in I} (B \cap A_i)$$

$$2. B \cup \left(\bigcap_{i \in I} A_i \right) = \bigcap_{i \in I} (B \cup A_i)$$

De Morgan's Laws:

$$1. \left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c$$

$$2. \left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c$$

We note that for the complement operation to make sense in De Morgan's Laws, we make an assumption that all sets are contained in one large set.

The proofs of the laws above are simple set theory exercises which you can try on your own or look up. We will do the proof of the first distributive law to demonstrate how it works:

Proof. Observe that:

$$\begin{aligned} x \in B \cap \left(\bigcup_{i \in I} A_i \right) &\iff x \in B \text{ and } x \in \left(\bigcup_{i \in I} A_i \right) \\ &\iff x \in B \text{ and } \exists i \in I \text{ s.t. } x \in A_i \\ &\iff \exists i \in I \text{ s.t. } x \in B \text{ and } x \in A_i \\ &\iff x \in \bigcup_{i \in I} (B \cap A_i) \end{aligned}$$

This concludes our proof. □

1.3 Sequences and Series

You must have seen some sequences and series in your Calculus II class already where you learned about different convergence tests and so on. Those will be definitely show up again and again in Analysis and Probability Theory. And while in your Calculus classes you have developed some tools to help you work with infinite series, in Analysis, it is essential to understand what infinite series really *are*. So let us define some rigorous notation that we'll keep referring to very often in this course.

Let $\{a_n\}, \{s_n\} \subseteq \mathbb{R}$ be two sequences of real numbers. We let $\{a_n\}$ be arbitrary, and let each $s_m \in \{s_n\}_{n \in \mathbb{N}}$ be defined as follows:

$$s_m = a_1 + a_2 + \dots + a_m = \sum_{k=1}^m a_k$$

We call $\{s_n\}$ the sequence of *partial sums* of the sequence $\{a_n\}$. Note also that given this sequence of partial sums, we can easily recover $\{a_n\}$, as follows:

$$a_n = \begin{cases} s_n - s_{n-1} & \text{if } n \geq 2 \\ s_1 & \text{if } n = 1 \end{cases}$$

Definition 2. We will break the definition of convergence into two parts:

1. We say that $\{s_n\}$ converges to the limit s if for all $\varepsilon > 0$ there exists $N > 0$ s.t. for all $n > N$, $|s_n - s| < \varepsilon$.
2. We say that $\sum_{n=1}^{\infty} a_n$ converges to s if $\lim_{n \rightarrow \infty} s_n = s$.

The first part of the definition says that for any positive real number ε , no matter how small, all but finitely many terms of $\{s_n\}$ are within ε distance from s , i.e. lie in $(s - \varepsilon, s + \varepsilon)$. The second part says that we don't *really* sum infinitely many numbers when working with infinite series, but rather take the limit of the sequence of the corresponding partial sum.

Example 10. Prove that $\lim_{n \rightarrow \infty} \frac{1}{n} = 0$.

Proof. Let $\varepsilon > 0$ be given. By the definition of the limit above, we would like to show that $|\frac{1}{n} - 0| = \frac{1}{n} < \varepsilon$. Note that if this were true, it would be equivalent to the expression $n > \frac{1}{\varepsilon}$. This gives us an idea of what N should be: we let $N \in \mathbb{Z}$ s.t. $N > \frac{1}{\varepsilon}$. Then for all $n > N$, we have that $n > N > \frac{1}{\varepsilon} \implies \frac{1}{n} < \varepsilon$, as desired. \square

So, if we define $\{s_n\} = \frac{1}{n}$ for all $n \in \mathbb{N}$, we can recover the sequence $\{a_n\}$ associated to it:

$$a_n = \begin{cases} \frac{1}{n} - \frac{1}{n-1} & \text{if } n \geq 2 \\ 1 & \text{if } n = 1 \end{cases}$$

Using the results above, we conclude that $\sum_{n=1}^{\infty} a_n = 0$. It also becomes intuitively clear once you write a few terms of the sum out:

$$\sum_{n=1}^{\infty} a_n = 1 + \left(\frac{1}{2} - 1\right) + \left(\frac{1}{3} - \frac{1}{2}\right) + \left(\frac{1}{4} - \frac{1}{3}\right) + \dots$$

Unsurprisingly, all the terms will *eventually* cancel out, and so we'll get the desired zero!

Example 11. Prove that if $|q| < 1$, then $\lim_{n \rightarrow \infty} \frac{1 - q^n}{1 - q} = \frac{1}{1 - q}$.

Proof. We will prove this result starting with the assumption that $\lim_{n \rightarrow \infty} q^n = 0$ for $|q| < 1$. (You can prove it for yourself or look up the proof!) With this assumption and using the result you will prove in HW1(3) about limit arithmetic, we get that:

$$\lim_{n \rightarrow \infty} \frac{1 - q^n}{1 - q} = \frac{1}{1 - q} \left(\lim_{n \rightarrow \infty} 1 - \lim_{n \rightarrow \infty} q^n \right) = \frac{1}{1 - q}$$

and this concludes the proof. □

Again, we can recover our sequence $\{a_n\}$ as follows:

$$a_n = \begin{cases} \frac{(1 - q^n) - (1 - q^{n-1})}{1 - q} & \text{if } n \geq 2 \\ 1 & \text{if } n = 1 \end{cases} = \begin{cases} q^{n-1} & \text{if } n \geq 2 \\ 1 & \text{if } n = 1 \end{cases}$$

So in this case $\{a_n\}$ is nicely and uniformly defined for all $n \in \mathbb{N}$, namely $a_n = q^{n-1}$ for all $n \geq 1$. Then:

$$\frac{1 - q^n}{1 - q} = 1 + q + q^2 + \dots + q^{n-1} = \sum_{k=1}^n a_k$$

Using the result we proved above, we deduce:

$$\sum_{n=0}^{\infty} q^n = \lim_{n \rightarrow \infty} \sum_{k=0}^n q^k = \lim_{n \rightarrow \infty} \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q}$$

which is the result you should remember from Calculus – this is the Geometric Series! In general, a good way to memorize the formula is (for $|q| < 1$):

$$a + aq + aq^2 + \dots = \frac{a}{1 - q} = \frac{(\text{first term})}{1 - (\text{ratio})}$$

Axiom 1 (The Axiom of Completeness). Let $s_1 \leq s_2 \leq s_3 \leq \dots$ be a monotone increasing sequence of real numbers and let $s_n \leq M \in \mathbb{R}$ for all $n \in \mathbb{N}$. Then $\{s_n\}$ converges to some $s \in \mathbb{R}$.

What this axiom says in English is that every bounded increasing sequence of real numbers converges to a real number. We will also remark that the limit s is in fact also bounded by M , i.e. $s \leq M$. This is somewhat intuitive, but you will also rigorously prove this (the Comparison Theorem) on HW1(2).

The assumption that we work with real numbers is very important. Real numbers are so great exactly because they are complete! The Axiom of Completeness will not hold if we replace all reals with rationals, because a bounded increasing sequence of rational numbers may NOT converge to a rational number, so in this sense, \mathbb{Q} is incomplete, i.e. there are "gaps." An example of this fact would be to consider the sequence $s_n = \left(1 + \frac{1}{n}\right)^n$, which is a sequence of rational numbers, but, as we all know, converges to the irrational number e . You will learn more about it in your Analysis class, if you haven't yet.

Theorem 3. If $a_n \geq 0$ for all $n \in \mathbb{N}$ and $\sum_{k=1}^n a_k \leq M$ for all $n \in \mathbb{N}$, then $\sum_{n=1}^{\infty} a_n$ converges and $\sum_{n=1}^{\infty} a_n \leq M$.

Proof. Since $a_n \geq 0$ for all $n \in \mathbb{N}$, then $s_n \geq s_{n-1}$ for each n . So we get an increasing sequence that is bounded by assumption. Thus, the result follows directly from the Axiom of Completeness. \square

We will also remark that it is a convention to write $\sum_{n=1}^{\infty} a_n = \infty$ if $\{s_n\}$ is unbounded when $a_n \geq 0$.

Corollary 1. If $\sum_{n=1}^{\infty} a_n$ converges, then $|\sum_{n=1}^{\infty} a_n| \leq \sum_{n=1}^{\infty} |a_n|$.

We all know and love this fact in the finite case. However, what happens in this infinite case is that we again *take limit* instead of actually summing infinitely many numbers. That is, what the corollary above says is that $|\lim_{n \rightarrow \infty} s_n| \leq \lim_{n \rightarrow \infty} t_n$ where $t_n = \sum_{k=1}^n |a_k|$. Now, let us actually prove that this is true.

Proof. Note that the result is obvious if $\sum_{n=1}^{\infty} |a_n| = \infty$, so assume this is not the case.

Let us start by identifying what we actually want to show. (Often times it is extremely helpful and immediately gives you an idea of how to proceed with the proof!) What we want to show is the following:

$$-\sum_{n=1}^{\infty} |a_n| \leq \sum_{n=1}^{\infty} a_n \leq \sum_{n=1}^{\infty} |a_n|$$

Equivalently,

$$-\lim_{n \rightarrow \infty} \sum_{n=1}^n |a_n| \leq \lim_{n \rightarrow \infty} \sum_{n=1}^n a_n \leq \lim_{n \rightarrow \infty} \sum_{n=1}^n |a_n|$$

This inequality follows directly from the Comparison Theorem (HW1(2)), since:

$$-\sum_{n=1}^n |a_n| \leq \sum_{n=1}^n a_n \leq \sum_{n=1}^n |a_n|$$

which is easily proven in the finite case by simply taking squares on both sides. (Do it!) \square

Example 12. For any $|x| < \frac{1}{2}$, we have $|\ln(1+x) - x| \leq x^2$.

Proof. Recall that since $|x| < 1$, the Taylor expansion for $\ln(1+x)$ is:

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

Therefore, using Corollary 1, we get:

$$\begin{aligned}
|\ln(1+x) - x| &= \left| \sum_{n=2}^{\infty} (-1)^{n+1} \frac{x^n}{n} \right| \\
&\leq \sum_{n=2}^{\infty} \left| \frac{x^n}{n} \right| \\
&\leq \sum_{n=2}^{\infty} \frac{|x|^n}{2} \\
&= \frac{\frac{x^2}{2}}{1 - |x|} \\
&= \frac{x^2}{2(1 - |x|)}
\end{aligned}$$

where the last equalities follow from the observation that we obtain the geometric series with ratio $|x|$. Now we also recall that $|x| < \frac{1}{2}$, and hence $2(1 - |x|) > 1$. Therefore,

$$|\ln(1+x) - x| \leq \frac{x^2}{2(1 - |x|)} \leq x^2$$

□

This is a very useful fact, and it may be a good idea to restate it in a slightly different manner. It is clear that $0 \leq a \leq b \iff \exists |\theta_b| \leq 1$ s.t. $a = \theta_b b$. Using this description of inequality in Example 12, we get that for any $|x| < \frac{1}{2}$ there exists θ_x such that $|\theta_x| \leq 1$ and $\ln(1+x) = x + \theta_x x^2$. This will turn out to be useful quite soon.

Let us now go over the Taylor expansion of $\ln(1+x)$ using definitions we introduced in this section. That is, what does

$$\ln(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}$$

really mean? Well, according to the definition, it means that for every $|x| < 1$, $\ln(1+x) = \lim_{n \rightarrow \infty} \sum_{k=1}^n (-1)^{k+1} \frac{x^k}{k}$. Equivalently, for every $|x| < 1$ and every $\varepsilon > 0$, there always exists $N = N(\varepsilon, x)$ (this notation means that N is allowed to depend on both ε and x), s.t. for every $n > N$, we have:

$$\left| \ln(1+x) - \sum_{k=1}^n (-1)^{k+1} \frac{x^k}{k} \right| < \varepsilon$$

Example 13. Show that for every $|x| < 1$, we have $|e^x - 1| \leq 2|x|$.

Proof. Recall that the Taylor expansion for e^x , ($x \in \mathbb{R}$) is:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

So, doing some algebra, we obtain:

$$\begin{aligned}
|e^x - 1| &= \left| \sum_{n=1}^{\infty} \frac{x^n}{n!} \right| \\
&\stackrel{(2.6)}{\leq} \sum_{n=1}^{\infty} \frac{|x|^n}{n!} \\
&\stackrel{(*)}{\leq} \sum_{n=1}^{\infty} \frac{|x|^n}{2^{n-1}} \\
&\stackrel{(**)}{\leq} \frac{|x|}{1 - \frac{|x|}{2}}
\end{aligned}$$

where $(*)$ follows from the observation that $n! > 2^{n-1}$ for all $n \in \mathbb{N}$ (you can prove this by induction, if you like). $(**)$ follows from noting that we obtain geometric series with the ratio $\frac{|x|}{2}$.

Now, since $|x| < 1$, it follows that $1 - \frac{|x|}{2} > \frac{1}{2}$. Thus,

$$|e^x - 1| \leq \frac{|x|}{1 - \frac{|x|}{2}} \leq \frac{|x|}{\frac{1}{2}} = 2|x|$$

□

1.4 σ -Algebras and Probability Spaces

Remember that in the section 1.1 we discovered that not all the sets in the power set of $[0, 1]$ can be assigned "length," i.e. not all the sets in $2^{[0, 1]}$ can be measured. It is rather bad news, since we still want to compute (or measure) probability for any event that occurs. The example in that section tells us that if we actually want to define a proper probability function that, in a way, "measures" the sets we input, we must be less ambitious and restrict the domain of that function to only contain the sets that can be measured. But what are those sets? What is meant by "measured"? This section will introduce the concepts that are used to answer some of the questions you may have.

Definition 3.3.1 Let Ω be a set. We say that $\mathcal{F} \subseteq 2^\Omega$ is a **σ -algebra** on Ω if:

1. $\emptyset \in \mathcal{F}$,
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$,
3. For any sequence of sets $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$, we have that $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

So, from the definition above, we can see that a σ -algebras are certain collections (families) of subsets of Ω .

Definition 4. We define a **probability space (or model)** to be a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where \mathcal{F} is a σ -algebra on Ω and $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ is a function such that:

1. $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\Omega) = 1,$
2. $\mathbb{P}\left(\bigsqcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ for any family of disjoint sets $\{A_n\} \subseteq \mathcal{F}.$

We call \mathbb{P} a **probability measure** on (Ω, \mathcal{F}) .

To make the language as precise as possible, we call each $\omega \in \Omega$ a **sample**, each $A \in \mathcal{F}$ an **event**. We refer to $\mathbb{P}(A)$ as the probability of the event A .

Remark:

- In view of the exercise 4, not all subsets of Ω need to have probability.
- The formula in 4-(2) is independent of the ordering of the elements in $\{A_n\}$. This is a consequence of HW1(4a).

Proposition 1. Let \mathcal{F} be a σ -algebra on Ω . Then:

1. $\Omega \in \mathcal{F},$
2. For any $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$, we have:

- a. $\bigcup_{k=1}^n A_k \in \mathcal{F},$

- b. $\bigcap_{k=1}^n A_k \in \mathcal{F},$

- c. $\bigcap_{n=1}^{\infty} A_n \in \mathcal{F},$

3. If $A, B \in \mathcal{F}$, then $A - B \in \mathcal{F}$, where $A - B = \{x \in \Omega : x \in A \text{ and } x \notin B\}.$

Proof. Since $\emptyset \in \mathcal{F}$ by the first axiom of being a σ -algebra and since every σ -algebra is closed under complement, we have that $\emptyset \in \mathcal{F} \implies \emptyset^c \in \mathcal{F} \implies \Omega \in \mathcal{F}$. Thus (1) holds. To prove (2a), simply note that since \mathcal{F} is a σ -algebra, it is closed under countable union. Taking all but finitely many sets to be empty gives us closure under finite union as well, thus proving (2a). Next, we will prove (2c) and see that (2b) will follow from it. Using the second and third axioms of being a σ -algebra, we get that:

$$\forall n \in \mathbb{N}, A_n \in \mathcal{F} \implies A_n^c \in \mathcal{F} \implies \bigcup_{n=1}^{\infty} A_n^c \in \mathcal{F}$$

Using De Morgan's Law, we know that:

$$\bigcup_{n=1}^{\infty} A_n^c = \left(\bigcap_{n=1}^{\infty} A_n \right)^c$$

And using closure under complement again, we obtain:

$$\left(\bigcap_{n=1}^{\infty} A_n\right)^c \in \mathcal{F} \implies \bigcap_{n=1}^{\infty} A_n \in \mathcal{F}$$

and this concludes the proof. Note that closure under finite intersection now follows from simply taking all but finitely many sets in the countable intersection to be $\Omega \in \mathcal{F}$. Finally, to prove (3), first note that $A - B = A \cap B^c$. Since \mathcal{F} is closed under complement, we have $B \in \mathcal{F} \implies B^c \in \mathcal{F}$. Then, by (2b), we conclude that $A - B = A \cap B^c \in \mathcal{F}$. \square

Example 14. The following are examples of σ -algebras on Ω :

1. 2^Ω ,
2. $\{\emptyset, \Omega\}$,
3. $\{\emptyset, A, A^c, \Omega\}$ for any $A \subseteq \Omega$.

It would be a good practice for you to check that each of these examples satisfies all of the three axioms.

Example 15. Let $\Omega = \mathbb{N}$. Let $\mathcal{F} = \{A \subseteq \mathbb{N} : \text{either } A \text{ or } A^c \text{ is finite}\}$. Then \mathcal{F} is NOT a σ -algebra on \mathbb{N} .

It is easy to see that the first two axioms are satisfied, since $\emptyset \in \mathcal{F}$ and $A \in \mathcal{F} \implies A^c \in \mathcal{F}$, trivially. However, if we let $A = \{2k : k \in \mathbb{N}\} = \bigcup_{k=0}^{\infty} \{2k\}$, then we see that for every $k \in \mathbb{N}$, $\{2k\} \in \mathcal{F}$, as it is a finite set. However, the infinite union is not in \mathcal{F} , since the resulting set A is the set of all even natural numbers, which is neither finite, nor co-finite. So, \mathcal{F} is not closed under countable union and thus is not a σ -algebra. However, it is closed under *finite* unions (check it!) and hence is an **algebra**.

Definition 5. Let \mathcal{F} be any σ -algebra on Ω . Let $x \in \Omega$. Define: $\delta_x : \mathcal{F} \rightarrow [0, 1]$ as follows:

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

We refer to δ_x as **Dirac's delta measure**.

Our next claim is that $(\Omega, \mathcal{F}, \delta_x)$ is a probability space. Let us verify this:

- Clearly, $\delta_x(\Omega) = 1$ and $\delta_x(\emptyset) = 0$,
- Let $\{A_n\}$ be a collection of disjoint sets in \mathcal{F} . To see why $\delta_x\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \delta_x(A_n)$, note that, since the sets are disjoint, then x belongs to at most one set in the collection. That is, if $x \in \biguplus_{n=1}^{\infty} A_n$, then there is only one set A_k containing x , and hence the only non-zero term in the sum on the right side will be 1 in the k th spot. Similarly, if x is not in the union, it cannot be in any set of the collection, and the conclusion follows as well.

Example 16. Let $\Omega = \{f : f : [0, 1] \rightarrow [0, \infty) \text{ and } f \text{ is continuous}\}$. Let $\mathcal{F} = 2^\Omega$. We think of each function in Ω as a function of time. Let $x \in \Omega$ such that $x(t) = 38$ for all $t \in [0, 1]$ (that is, x is a constant function on the specified interval. We note that $(\Omega, \mathcal{F}, \delta_x)$ is a probability model for a market where no transaction occurs in the next business day. That is, if we ask: what is the probability that tomorrow's stock price is at least 35 (for all $t \in [0, 1]$)? Rephrasing it in math terms, we defined a set $A = \{f \in \Omega : f(t) \geq 35, t \in [0, 1]\}$ and asked what is $\delta_x(A)$. Since the constant function $x \in A$, by the definition of Dirac's delta measure, $\delta_x(A) = 1$. That is, we are absolutely positive that tomorrow's stock price will be above 35, because it is always 38. One of the possible explanations for that is that no transactions occur tomorrow, and thus stock prices will not be affected. Clearly, this is not the most sophisticated model and probably is of little use in the modern day stock market. But this baby example demonstrates how probability measure can model real-world situations. To build more complicated models, your probability measure will have to be much more involved – it may come from the stochastic differential equations (look up the Black-Scholes equation, for instance) you're working with. Those of you who decide to pursue a career in financial math, may see it some day.

Theorem 4. Let $\mathbb{P}_1, \mathbb{P}_2, \dots, \mathbb{P}_n$ be probability measures on (Ω, \mathcal{F}) and let $\{\alpha_n\}$ be a sequence such that $\alpha_n \geq 0$ for all $n \in \mathbb{N}$ and $\sum_{n=1}^{\infty} \alpha_n = 1$. Define

$$\mathbb{P}(A) = \left(\sum_{n=1}^{\infty} \alpha_n \mathbb{P}_n \right)(A) = \sum_{n=1}^{\infty} \alpha_n \mathbb{P}_n(A)$$

Then \mathbb{P} is a probability measure on (Ω, \mathcal{F}) .

Before we prove the above theorem, we would like to remark that it completely characterizes probability measure on countable Ω . Recall that in the first section, we said that if $\Omega = \{\omega_1, \omega_2, \dots\}$, then we think of probability as some weight assignment to the members of Ω . So, in the light of the above theorem, we can think of it as assigning the weight α_i to $\omega_i \in \Omega$ for each $i \in \mathbb{N}$. Then, if we take $A = \{\omega_{n_1}, \omega_{n_2}, \dots\} \subseteq \Omega$, then

$$\mathbb{P}(A) = \left(\sum_{n=1}^{\infty} \alpha_n \delta_{\omega_n} \right)(A) = \sum_{n=1}^{\infty} \alpha_n \delta_{\omega_n}(A) = \sum_{\omega_{n_i} \in A} \alpha_{n_i}$$

That is, we only sum the weights of the members of A .

Proof. In order to show that \mathbb{P} is a probability measure, we need to check that the two conditions are satisfied:

1. Clearly, $\mathbb{P}(\emptyset) = \sum_{n=1}^{\infty} \alpha_n \cdot 0 = 0$ and $\mathbb{P}(\Omega) = \sum_{n=1}^{\infty} \alpha_n \cdot \mathbb{P}_n(\Omega) = \sum_{n=1}^{\infty} \alpha_n = 1$.

2. Let $\{A_k\}$ be a sequence of pairwise disjoint sets in \mathcal{F} . Then:

$$\begin{aligned}
\mathbb{P}\left(\biguplus_{k=1}^{\infty} A_k\right) &= \sum_{n=1}^{\infty} \alpha_n \mathbb{P}_n\left(\biguplus_{k=1}^{\infty} A_k\right) \\
&\stackrel{(*)}{=} \sum_{n=1}^{\infty} \alpha_n \sum_{k=1}^{\infty} \mathbb{P}_n(A_k) \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \alpha_n \mathbb{P}_n(A_k) \\
&\stackrel{(**)}{=} \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \alpha_n \mathbb{P}_n(A_k) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(A_k)
\end{aligned} \tag{2}$$

where $(*)$ follows from the fact that each \mathbb{P}_n is known to be a probability measure, and thus is countably additive. $(**)$ follows from HW1(4b), since the terms we are summing are always non-negative.

□

Definition 6. Let $\mathcal{E} \subseteq 2^{\Omega}$, and define

$$\Lambda = \{\mathcal{A} \subseteq 2^{\Omega} : \mathcal{A} \text{ is a } \sigma\text{-algebra on } \Omega \text{ and } \mathcal{E} \subseteq \mathcal{A}\}$$

Then $\sigma(\mathcal{E}) = \bigcap_{\mathcal{A} \in \Lambda} \mathcal{A} = \{A \subseteq \Omega : A \in \mathcal{A} \forall \mathcal{A} \in \Lambda\}$ is the smallest σ -algebra containing \mathcal{E} , called the σ -algebra **generated by** \mathcal{E} .

What the above definition is saying is that if we take some set Ω , consider all σ -algebras on it and then take their intersection, then the intersection itself is a σ -algebra, and moreover, is the smallest σ -algebra defined on Ω . You may be a bit critical of this definition right now. First of all, how do we know that Λ is not empty? Or how do we know that the intersection of σ -algebras is a σ -algebra itself? Or even if so, how do we know it is the smallest one possible? In fact, we don't know any of these yet, and we must prove it. So let us prove the following claim.

Proposition 2. The following are true:

- $\Lambda \neq \emptyset$,
- $\sigma(\mathcal{E}) \in \Lambda$,
- $\sigma(\mathcal{E}) \subseteq \mathcal{A}$ for all $\mathcal{A} \in \Lambda$.

Proof. Observe that the power set of Ω , 2^{Ω} is a σ -algebra containing all other σ -algebras. Hence, $2^{\Omega} \in \Lambda$, and thus Λ is not empty. To see why the second one is true, we need to prove that an intersection of σ -algebras is itself a σ -algebra. Let $\{\mathcal{A}_i\}_{i \in I}$ be a collection of σ -algebras on Ω and denote $\bigcap_{i \in I} \mathcal{A}_i = \mathcal{A}$. Then:

1. $\emptyset \in \mathcal{A}_i$ for all $i \in I \implies \emptyset \in \mathcal{A}$,
2. $A \in \mathcal{A} \implies A \in \mathcal{A}_i$ for all $i \in I \implies A^c \in \mathcal{A}_i$ for all $i \in I \implies A^c \in \mathcal{A}$,
3. Suppose $\{A_n\} \subseteq \mathcal{A}$. Then, $\{A_n\} \subseteq \mathcal{A}_i$ for all $i \in I$. Since each A_i is a σ -algebra, we have that $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_i$ for all $i \in I$. But then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

To prove the last claim, let $A \in \sigma(\mathcal{E})$. Then, since $\sigma(\mathcal{E})$ is defined as the intersection of all σ -algebras containing \mathcal{E} , A is in every σ -algebra containing \mathcal{E} . That is, $A \in \mathcal{A}$ for all $\mathcal{A} \in \Lambda$. Since A was arbitrary, this concludes the proof. \square

Example 17. Let $\mathcal{E} = \{A\} \subseteq 2^\Omega$. Prove that $\sigma(\mathcal{E}) = \{\emptyset, A, A^c, \Omega\} \subseteq 2^\Omega$.

Proof. Exercise to the reader. (Do it!) \square

1.5 The Borel σ -Algebra

Proposition 3. Recall that in the previous section, we claimed the following: Let Ω be a set, $\mathcal{E} \subseteq 2^\Omega$ a collection of subsets of Ω , and

$$\sigma(\mathcal{E}) = \{A \subseteq \Omega : A \in \mathcal{A} \text{ for any } \sigma\text{-algebra } \mathcal{A} \text{ on } \Omega \text{ containing } \mathcal{E}\}.$$

Then,

1. the family $\sigma(\mathcal{E})$ is a σ -algebra on Ω , and moreover $\mathcal{E} \subseteq \sigma(\mathcal{E})$;
2. for any σ -algebra \mathcal{A} on Ω which contains \mathcal{E} , we have $\sigma(\mathcal{E}) \subseteq \mathcal{A}$.

This justifies the terminology of $\sigma(\mathcal{E})$ for being the smallest σ -algebra containing \mathcal{E} .

Proof. To begin, let us prove that an intersection of σ -algebras is itself a σ -algebra. Let $\{\mathcal{A}_i\}_{i \in I}$ be a collection of σ -algebras on Ω and denote $\bigcap_{i \in I} \mathcal{A}_i = \mathcal{A}$. Then:

1. $\emptyset \in \mathcal{A}_i$ for all $i \in I \implies \emptyset \in \mathcal{A}$,
2. $A \in \mathcal{A} \implies A \in \mathcal{A}_i$ for all $i \in I \implies A^c \in \mathcal{A}_i$ for all $i \in I \implies A^c \in \mathcal{A}$,
3. Suppose $\{A_n\} \subseteq \mathcal{A}$. Then, $\{A_n\} \subseteq \mathcal{A}_i$ for all $i \in I$. Since each A_i is a σ -algebra, we have that $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}_i$ for all $i \in I$. But then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Since $\sigma(\mathcal{E})$ is the intersection of all σ -algebras containing \mathcal{E} , then the intersection itself must contain \mathcal{E} as well. Hence $\mathcal{E} \subseteq \sigma(\mathcal{E})$, and this concludes the proof of (1).

To prove the second claim, let $A \in \sigma(\mathcal{E})$. Then, since $\sigma(\mathcal{E})$ is defined as the intersection of all σ -algebras containing \mathcal{E} , A is in every σ -algebra containing \mathcal{E} , i.e. $A \in \mathcal{A}$ for any σ -algebra \mathcal{A} containing \mathcal{E} . Since A was arbitrary, we obtain the desired conclusion. \square

Definition 7. The Borel σ -algebra \mathcal{B} is the σ -algebra on \mathbb{R} generated by

$$\mathcal{A} = \{(-\infty, a] : a \in \mathbb{R}\}.$$

We shall call any $B \in \mathcal{B}$ a Borel subset of \mathbb{R} .

Example 18. Let $a, b \in \mathbb{R}$ with $a \leq b$. Show that $(-\infty, a)$, $(a, +\infty)$, $[a, +\infty)$, (a, b) , $[a, b)$, $(a, b]$, $[a, b]$, and $\{a\}$ are all Borel subsets of \mathbb{R} .

Proof. By the previous exercise, we know that $\mathcal{A} \subseteq \sigma(\mathcal{A}) = \mathcal{B}$. We also know that \mathcal{B} is a σ -algebra, so it must contain all possible complements and (countable) unions of elements in \mathcal{A} . Let us consider each of the Borel sets separately:

1. $(-\infty, a) = \bigcup_{n=1}^{\infty} (-\infty, a - \frac{1}{n}]$. Since each $(-\infty, a - \frac{1}{n}] \in \mathcal{A}$, and \mathcal{B} is a σ -algebra, then the countable union must be in \mathcal{B} as well. Hence, $(-\infty, a) \in \mathcal{B}$.
2. $(a, \infty) = (-\infty, a]^c$. Since $(-\infty, a] \in \mathcal{A}$ and \mathcal{B} is closed under complement, then $(-\infty, a]^c \in \mathcal{B}$.
3. $[a, \infty) = (-\infty, a)^c$. From (1), we know that $(-\infty, a) \in \mathcal{B}$, and thus its complement is in \mathcal{B} as well.
4. $(a, b) = (a, \infty) \cap (-\infty, b)$. Both of these sets are in \mathcal{B} by (1) and (2). We also proved in Section 1.4 that σ -algebras are closed under finite intersections, so the result follows.
5. $[a, b) = [a, \infty) \cap (-\infty, b)$. Both of these sets are in \mathcal{B} by (1) and (3), and thus $[a, b) \in \mathcal{B}$.
6. $(a, b] = (a, \infty) \cap (-\infty, b]$. The first set is in \mathcal{B} by (2) and the second set is in $\mathcal{A} \subseteq \mathcal{B}$. Thus, their intersection is in \mathcal{B} .
7. $[a, b] = [a, \infty) \cap (-\infty, b]$. The first set is in \mathcal{B} by (3) and the second set is in $\mathcal{A} \subseteq \mathcal{B}$. Thus, their intersection is in \mathcal{B} .
8. $\{a\} = [a, \infty) \cap (-\infty, a]$. Both sets are in \mathcal{B} , and thus their intersection is as well.

□

Example 19. Let $A \in \mathcal{B}$ be a Borel subset of $\subseteq \mathbb{R}$. Show that $\mathcal{B}(A) = \{A \cap B : B \in \mathcal{B}\}$ is a σ -algebra on A . We shall call $\mathcal{B}(A)$ the Borel σ -algebra on A .

Proof. We need to show that $\mathcal{B}(A)$ satisfies the three axioms of being a σ -algebra:

- Clearly, $\emptyset \in \mathcal{B}$, and hence $A \cap \emptyset = \emptyset \in \mathcal{B}(A)$. Also, $A \in \mathcal{B}$ by assumption, and hence $A \cap A = A \in \mathcal{B}(A)$.
- Let $C \in \mathcal{B}(A)$. Then $C = A \cap B$ for some Borel set B . Then $(\mathbb{R} \setminus B) \in \mathcal{B}$, since \mathcal{B} is a σ -algebra. Thus $A \cap (\mathbb{R} \setminus B) \in \mathcal{B}(A)$. However, $A \cap (\mathbb{R} \setminus B) = A \setminus (A \cap B)$. Thus, $A \setminus C = A \setminus (A \cap B) = A \cap (\mathbb{R} \setminus B) \in \mathcal{B}(A)$ is desired.

- Let $\{C_i\}_{i \in I} \subseteq \mathcal{B}(A)$. Then $C_i = A \cap B_i$ for each $i \in I$. Then:

$$\begin{aligned}\bigcup_{i=1}^{\infty} C_i &= \bigcup_{i=1}^{\infty} (A \cap B_i) \\ &= A \cap \left(\bigcup_{i=1}^{\infty} B_i \right) \\ &= A \cap B\end{aligned}$$

where $B = \bigcup_{i=1}^{\infty} B_i$. Since each B_i is Borel, then the union of Borel sets is also Borel, and thus B is Borel. Hence, $\bigcup_{i=1}^{\infty} C_i = A \cap B \in \mathcal{B}(A)$, as desired.

□

Recall that it is not a good idea to define length for all subsets of $[0, 1]$. However, the following theorem says that we can do this for all Borel sets. Before stating the theorem, let us define a finite interval I to be either (a, b) , $[a, b)$, $(a, b]$ or $[a, b]$, where a, b are real numbers with $a < b$, and call $|I| = b - a$ the length of the interval.

Theorem 5. (c.f. Example 4) For any finite interval I there exists a unique function $m : \mathcal{B}(I) \rightarrow [0, +\infty)$ such that

1. $m(\emptyset) = 0$;
2. $m(I_1) = |I_1|$ for any subinterval $I_1 \subseteq I$;
3. $m(\biguplus_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} m(A_n)$ for any pairwise disjoint sequence $\{A_n\} \subseteq \mathcal{B}(I)$;
4. $m(x + E) = m(E)$ for any $E, x + E \in \mathcal{B}(I)$.

At Wesleyan the proof of Theorem 5 is usually covered in Math255 (Fundamentals of Real Analysis II) or in a graduate real analysis course. The set function $m : \mathcal{B}(I) \rightarrow [0, +\infty)$ is called the Lebesgue measure on $(I, \mathcal{B}(I))$. It is not a probability measure whenever $m(I) = |I|$ is not equal to 1. However, if we define $\mathbb{P} : \mathcal{B}(I) \rightarrow [0, 1]$ to be

$$\mathbb{P}(A) = \frac{m(A)}{|I|}, \quad A \in \mathcal{B}(I).$$

Then $(I, \mathcal{B}(I), \mathbb{P})$ is a probability space, and moreover \mathbb{P} satisfies Property 4 of Theorem 5. With respect to this model, for any Borel subset A of I it makes sense to talk about the probability for a random point of I lying in A , which, by definition, is equal to $\mathbb{P}(A)$.

Let us finish this section by proving a lemma that will be used later.

Lemma 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\{A_n\} \subseteq \mathcal{F}$ a sequence of events. Let $B_1 = A_1$, and $B_n = A_n - (\bigcup_{i=1}^{n-1} A_i)$ for any $n > 1$. Show that

1. $\{B_n\}$ is a sequence of pairwise disjoint events;

Proof. Let B_n and B_m be given and assume that $n < m$. Let $x \in B_n$ be arbitrary. Then, by the way B_n is defined, $x \in A_n$. However, then since $n < m$, we have that $x \notin B_m = A_m \setminus \left(\bigcup_{k=1}^{m-1} A_k \right)$. Since x was an arbitrary element of B_n , none of the elements in B_n are in B_m , and hence $B_n \cap B_m = \emptyset$. \square

2. $\bigcup_{k=1}^n A_k = \uplus_{k=1}^n B_k$ for any $n \geq 1$;

Proof. First note that the disjointness of B_k sets follows from part (a). So all we need to prove is the equality of the two unions. We will prove the claim by induction on n . Note that we are given that $A_1 = B_1$, and thus the base case holds. Let now $n \in \mathbb{N}$ be given. To prove the inductive case, assume that $\bigcup_{k=1}^{n-1} A_k = \bigcup_{k=1}^{n-1} B_k$. Then:

$$\begin{aligned}
\bigcup_{k=1}^n A_k &= \left(\bigcup_{k=1}^{n-1} A_k \right) \cup A_n \\
&\stackrel{(*)}{=} \left(\bigcup_{k=1}^{n-1} B_k \right) \cup A_n \\
&= \left(\bigcup_{k=1}^{n-1} B_k \right) \cup \left(\left(B_n \cup \bigcup_{k=1}^{n-1} A_k \right) \cap A_n \right) \\
&= \left(\bigcup_{k=1}^{n-1} B_k \right) \cup \left(\left(B_n \cup \bigcup_{k=1}^{n-1} B_k \right) \cap A_n \right) \\
&= \left(\bigcup_{k=1}^{n-1} B_k \right) \cup (B_n \cap A_n) \\
&= \left(\bigcup_{k=1}^{n-1} B_k \right) \cup B_n \\
&= \bigcup_{k=1}^n B_k
\end{aligned}$$

where $(*)$ follows from the inductive hypothesis. \square

3. $\bigcup_{n=1}^{\infty} A_n = \uplus_{n=1}^{\infty} B_n$.

Proof. Let $x \in \bigcup_{n=1}^{\infty} A_n$. Then $x \in A_m$ for some $m \in \mathbb{N}$. So $x \in \bigcup_{n=1}^m A_n = \bigcup_{n=1}^m B_n$, where the last equality follows from part (b). But then, $x \in \bigcup_{n=1}^{\infty} B_n$, since adding more elements to the union can only enlarge it. Hence, $\bigcup_{n=1}^{\infty} A_n \subseteq \bigcup_{n=1}^{\infty} B_n$. The reverse containment is similar (do it!) We then conclude that $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$. \square

1.6 One-Dimensional Probability Distributions

Proposition 4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Show the following holds:

1. $\mathbb{P}\left(\biguplus_{n=1}^k A_n\right) = \sum_{n=1}^k \mathbb{P}(A_n)$ for any pairwise disjoint $\{A_n\} \subseteq \mathcal{F}$;
2. $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$ for all $A \in \mathcal{F}$;
3. $\mathbb{P}(A) \leq \mathbb{P}(B)$ for all $A, B \in \mathcal{F}$ with $A \subseteq B$.

Proof. To prove (1), simply note that \mathbb{P} is countably additive. To show that it is also finitely additive, simply let all but finitely many sets be empty. In our case, let $A_i = \emptyset$ for all $i > k$. Then, $\mathbb{P}(A_i) = \mathbb{P}(\emptyset) = 0$ for all $i > k$. Thus,

$$\mathbb{P}\left(\biguplus_{n=1}^k A_n\right) = \mathbb{P}\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^k \mathbb{P}(A_n) + \sum_{n=k+1}^{\infty} 0 = \sum_{n=1}^k \mathbb{P}(A_n)$$

To prove (2), we will use part (1). Note that:

$$\begin{aligned} \Omega &= A \uplus A^c \implies \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c) \\ &\implies 1 = \mathbb{P}(A) + \mathbb{P}(A^c) && \text{by (1)} \\ &\implies \mathbb{P}(A) = 1 - \mathbb{P}(A^c) \end{aligned}$$

Finally, to prove (3), note that $B = A \uplus (B - A)$. Hence, by (1), we get:

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B - A) \implies \mathbb{P}(B) \geq \mathbb{P}(A)$$

since $\mathbb{P}(B - A)$ is always non-negative. □

Note that the above proposition is a simple corollary from the definition of probability measure, but is just as important to keep in mind as the definition itself.

Theorem 6. [Continuity Theorem of Probability Measure] Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then:

1. If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$, then $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$;
2. If $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$, then $\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$;

Proof. To prove (1), recall from Lemma 1-2 that for all $n \in \mathbb{N}$, we have $\bigcup_{k=1}^n A_k = \biguplus_{k=1}^n B_k$

where $B_1 = A_1, B_n = A_n - \bigcup_{k=1}^{n-1} A_k$. Then:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) &= \mathbb{P}\left(\biguplus_{n=1}^{\infty} B_n\right) && \text{Lemma 1-3} \\
&= \sum_{n=1}^{\infty} \mathbb{P}(B_n) && \sigma\text{-additivity of } \mathbb{P} \\
&= \lim_{n \rightarrow \infty} \sum_{k=1}^n \mathbb{P}(B_k) && \text{definition of series} \\
&= \lim_{n \rightarrow \infty} \mathbb{P}\left(\biguplus_{k=1}^n B_k\right) && \text{Prop. 4-1} \\
&= \lim_{n \rightarrow \infty} \mathbb{P}\left(\bigcup_{k=1}^n A_k\right) \\
&= \lim_{n \rightarrow \infty} \mathbb{P}(A_n) && \text{nested sets assumption}
\end{aligned}$$

To prove (2), we'll use De Morgan's Law. Note that $\left(\bigcap_{n=1}^{\infty} A_n\right)^c = \bigcup_{n=1}^{\infty} A_n^c$. Since, by assumption, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$, then $A_1^c \subseteq A_2^c \subseteq A_3^c \subseteq \dots$. By (1), we know that $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^c\right) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n^c)$. Also, by 4-2, we know that $\mathbb{P}(A_n) = 1 - \mathbb{P}(A_n^c)$, so:

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right) &= 1 - \mathbb{P}\left(\left(\bigcap_{n=1}^{\infty} A_n\right)^c\right) && \text{Prop. 4-2} \\
&= 1 - \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n^c\right) && \text{De Morgan's Law} \\
&= 1 - \lim_{n \rightarrow \infty} \mathbb{P}(A_n^c) && \text{part (1)} \\
&= \lim_{n \rightarrow \infty} (1 - \mathbb{P}(A_n^c)) && \text{limit arithmetic} \\
&= \lim_{n \rightarrow \infty} (\mathbb{P}(A_n)) && \text{4-2}
\end{aligned}$$

□

In the future, whenever we talk about sets that we measure probability of, we'll always assume they are in the σ -algebra defined on the probability space we're in.

Proposition 5. For any $A \subseteq \bigcup_{n=1}^{\infty} A_n$, we have $\mathbb{P}(A) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

Proof. Let $\{A_n\}$ be a sequence of sets, and $A \subseteq \bigcup_{n=1}^{\infty} A_n$. Define the disjoint sequence of sets $\{B_n\}$ as in Theorem 6. Clearly, $B_n \subseteq A_n$ for all $n \in \mathbb{N}$. Thus, by 4-3, $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$ for all

$n \in \mathbb{N}$. Hence:

$$\begin{aligned}
\mathbb{P}(A) &\leq \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) && 5.1-3 \\
&= \mathbb{P}\left(\bigoplus_{n=1}^{\infty} B_n\right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(B_n) && \sigma\text{-additivity of } \mathbb{P} \\
&\leq \sum_{n=1}^{\infty} \mathbb{P}(A_n) && 4-3
\end{aligned}$$

This property is called **σ -subadditivity**. □

Corollary 2. Let \mathbb{P} be a probability measure on $(\mathbb{R}, \mathcal{B})$. Then:

1. $\lim_{n \rightarrow \infty} \mathbb{P}((-\infty, n]) = 1$;
2. $\lim_{n \rightarrow \infty} \mathbb{P}((-\infty, x + \frac{1}{n}]) = \mathbb{P}((-\infty, x])$.

Proof. To prove (1), note that $(-\infty, 1] \subseteq (-\infty, 2] \subseteq (-\infty, 3] \subseteq \dots$. So we have an increasing sequence of nested intervals. Thus we can apply 6-1 and get:

$$\lim_{n \rightarrow \infty} \mathbb{P}((-\infty, n]) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} (-\infty, n]\right) = \mathbb{P}(\mathbb{R}) = 1$$

To prove (2), note that $(-\infty, x + 1] \supseteq (-\infty, x + \frac{1}{2}] \supseteq (-\infty, x + \frac{1}{3}] \supseteq \dots$. That is, we have a decreasing sequence of nested intervals, and can apply 6-2. We get:

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left(-\infty, x + \frac{1}{n}\right]\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} \left(-\infty, x + \frac{1}{n}\right]\right) = \mathbb{P}((-\infty, x])$$

□

Definition 8. We will introduce some terminology:

1. We will denote the set of all probability measures defined on $(\mathbb{R}, \mathcal{B})$ by $\text{Prob}(\mathbb{R}, \mathcal{B})$. That is, $\text{Prob}(\mathbb{R}, \mathcal{B}) = \{\mathbb{P} : \mathbb{P} \text{ is a probability measure on } (\mathbb{R}, \mathcal{B})\}$.
2. Any $\mathbb{P} \in \text{Prob}(\mathbb{R}, \mathcal{B})$ is called a **one-dimensional probability distribution**.
3. For any $\mathbb{P} \in \text{Prob}(\mathbb{R}, \mathcal{B})$, we define a function $F(x) := \mathbb{P}((-\infty, x])$ called the **cumulative distribution function (cdf)** of \mathbb{P} .

Informally speaking, cdf measures probability (or weight) up to a certain point. Also note that $\lim_{x \rightarrow \infty} F(x) = 1$.

We are now ready to ask the following very important question: How should we properly

describe a probability distribution? Clearly, we can't just assign weight to each set in our σ -algebra – or we might find ourselves doing so for the rest of eternity, since there may be uncountably many such sets! There must be a smarter way. And there is! Turns out, cdf's completely uniquely characterize probability distributions. This result is stated formally in the next theorem:

Theorem 7. Let $\mathbb{P}_1, \mathbb{P}_2 \in \text{Prob}(\mathbb{R}, \mathcal{B})$. If $\mathbb{P}_1((-\infty, x]) = \mathbb{P}_2((-\infty, x])$ for all $x \in \mathbb{R}$, then $\mathbb{P}_1 = \mathbb{P}_2$.

Proof. You are proving it in HW2-5. □

For probability measures that can be written as combinations of Dirac Delta measures, it is convenient to write them down explicitly.

Example 20. We will now define some of the very important probability distributions:

- Bernoulli: Let $p, q \in [0, 1]$ with $p + q = 1$. We define $\text{Ber}(p) = p\delta_1 + q\delta_0$. This is called Bernoulli distribution with parameter p . It describes the probability distribution of a random variable (to be formally defined in the next section!) that takes the value 1 with probability p and takes the value 0 with probability q .
- Binomial: We describe binomial distribution with parameters p and n by $\text{Bin}(n, p) = \sum_{k=1}^n \binom{n}{k} p^k q^{n-k} \delta_k$. We will see that this is the probability distribution of the number of successes in a sequence of n independent experiments, where each success occurs with probability p .
- Geometric: We define geometric distribution with parameter p by $\text{Geo}(p) = \sum_{k=1}^{\infty} p q^{k-1} \delta_k$ where $p, q \in [0, 1], p + q = 1$. The geometric distribution gives the probability that the first occurrence of success requires a certain number of independent trials, each with success probability p .

Recall that by Theorem 4, $\sum_{n=1}^{\infty} \alpha_n \mathbb{P}_n$ is a probability measure if each \mathbb{P}_n is so and $\sum_{n=1}^{\infty} \alpha_n = 1, \alpha_n \geq 0$ for all n . Similarly, we can think of the geometric distribution as assigning weights p, pq, pq^2, \dots to the elements in our sample space. Then, $\sum_{k=1}^{\infty} p q^{k-1} = \frac{p}{1-q} = \frac{p}{p} = 1$.

Now suppose $\mathbb{P} = \text{Geo}(p)$, then we can measure $\mathbb{P}((2.5, 4.4)) = pq^2 + pq^3$, since

$$\sum_{n=1}^{\infty} \alpha_n \delta_{x_n}(A) = \sum_{x_n \in A} \alpha_n$$

That is, we only sum the weights of points in our set.

It is important to remark that there are probability measures that cannot be written as combinations of Dirac Delta measures.

Example 21. Let \mathbb{P} be the Lebesgue measure on $((0, 1), \mathcal{B}(0, 1))$, i.e the unique probability measure such that $\mathbb{P}(c, d) = d - c$ for all $0 < c \leq d < 1$. Note it is not a combination of Dirac Delta measures, because the probability of each particular point is simply 0. It is also important to note that Lebesgue measure is not a one-dimensional distribution, because it is only defined on an interval. But we can make it such by letting:

$$\tilde{\mathbb{P}} = \mathbb{P}(B \cap (0, 1))$$

Claim: $\tilde{\mathbb{P}} \in \text{Prob}(\mathbb{R}, \mathcal{B})$.

Proof. We need to prove that the two axioms hold:

1. Clearly, $\tilde{\mathbb{P}}(\mathbb{R}) = \mathbb{P}(\mathbb{R} \cap (0, 1)) = \mathbb{P}(0, 1) = 1$ and $\tilde{\mathbb{P}}(\emptyset) = \mathbb{P}(\emptyset \cap (0, 1)) = \mathbb{P}(\emptyset) = 0$.
2. Let $\{A_n\} \subseteq \mathcal{B}$ be pairwise disjoint. Notice that $(0, 1) \cap \left(\biguplus_{n=1}^{\infty} A_n\right) = \biguplus_{n=1}^{\infty} (A_n \cap (0, 1))$, by the distributive law. So:

$$\begin{aligned} \tilde{\mathbb{P}}\left(\biguplus_{n=1}^{\infty} A_n\right) &= \mathbb{P}\left(\left(\biguplus_{n=1}^{\infty} A_n\right) \cap (0, 1)\right) \\ &= \mathbb{P}\left(\biguplus_{n=1}^{\infty} (A_n \cap (0, 1))\right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(A_n \cap (0, 1)) \\ &= \sum_{n=1}^{\infty} \tilde{\mathbb{P}}(A_n) \end{aligned}$$

and this concludes the proof. □

Now, we are ready to compute cdf of \tilde{P} . By definition,

$$F(x) = \tilde{\mathbb{P}}((-\infty, x]) = \mathbb{P}((-\infty, x] \cap (0, 1)) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

If we were to graph it, we would get a continuous function differentiable everywhere except at 0 and 1. This distribution has a name (and thus is important!)

Definition 9. The measure $\tilde{\mathbb{P}} \in \text{Prob}(\mathbb{R}, \mathcal{B})$ defined in 21 is called the **uniform distribution** on $(0, 1)$. It is characterized by the cdf:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

We would like to remark two things:

1. The uniform distribution on $[0, 1]$, $[0, 1)$, $(0, 1]$, $(0, 1)$ are the same. This is because the Lebesgue measure of a particular point is 0 (each point has length 0).
2. We can generalize the notion of uniform distribution to an arbitrary interval (a, b) on the real line as follows:

$$F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

Note that for $a < x < b$, $F(x)$ represents the ratio of the length of the interval (a, x) to the length of (a, b) .

1.7 Conditional Probability

You might have heard about different paradoxes such as the Birthday Problem or the Monty Hall Problem, which people have been debating about for decades. Many people have their own theories about how those paradoxes are resolved. If you ask a probability theorist, however, he or she will tell you that those problems are not paradoxes at all, but simple consequences on the notion of conditional probability. Conditional probability turns out to be extremely important when it comes to modeling real-life situations, as often events we're interested in are dependent on one another; and this, naturally, affects the probabilities of their occurrences.

Example 22 (Birth Paradox). Consider a family with two children. Suppose we don't know their genders, and each child is equally likely to be a boy or a girl. That is, there are four possibilities: boy and boy, girl and girl, boy and girl, girl and boy (naturally, we assume the children are distinguishable...) Rephrasing this in the language of mathematics, our sample space is $\Omega = \{bb, gg, bg, gb\}$. Since our set is finite, we can take $\mathcal{F} = 2^\Omega$ and let $\mathbb{P} = \sum_{\omega \in \Omega} \frac{1}{4} \delta_\omega$. So, our probability space is $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose we want to calculate the probability of the event that there is exactly one boy and one girl. So $A = \{bg, gb\}$. Then, clearly, $\mathbb{P}(A) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Now, let's additionally assume that we found out that the family has at least one girl. What is the probability of the event A now, that we have this information available? Intuitively, we are inclined to say that the probability has now changed to $\frac{2}{3}$, since the event that both kids are boys is now eliminated. But clearly, $\frac{2}{3} \neq \frac{1}{2}$, so how is it possible for the same event to have different probabilities?!

The answer to the above question is: conditional probability. It turns out, in the two above situations, event A occurs in two different spaces – and those spaces have different probability measures defined on them.

Lemma 2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Define $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$ as:

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

Then $\mathbb{P}(\bullet|B)$ is a probability measure on (Ω, \mathcal{F})

Proof. We need to verify two things:

1. Observe that by definition:

$$\mathbb{P}(\emptyset|B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = \frac{0}{\mathbb{P}(B)} = 0$$

Also:

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$$

2. Let $\{A_n\}_{n \in \mathbb{N}}$ be a collection of disjoint sets in \mathcal{F} . Then:

$$\begin{aligned} \mathbb{P}\left(\biguplus_{n=1}^{\infty} A_n | B\right) &= \frac{\mathbb{P}\left(\left(\biguplus_{n=1}^{\infty} A_n\right) \cap B\right)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}\left(\biguplus_{n=1}^{\infty} (A_n \cap B)\right)}{\mathbb{P}(B)} \\ &= \frac{\sum_{n=1}^{\infty} \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} && \mathbb{P} \text{ is } \sigma\text{-additive} \\ &= \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} \\ &= \sum_{n=1}^{\infty} \mathbb{P}(A_n | B) \end{aligned}$$

□

Definition 10. The probability space $(\Omega, \mathcal{F}, \mathbb{P}(\bullet|B))$ is called the **model of conditioning with respect to B** . We call $\mathbb{P}(A|B)$ the **conditional probability of A given B** . The initial model $(\Omega, \mathcal{F}, \mathbb{P})$ is called the **absolute model**. (And we refer to \mathbb{P} as absolute probability measure).

As mentioned above, conditional probability is the solution to the birth paradox. Let us see how. Define B to be the event that the family has at least one girl. Then $\mathbb{P}(B) = \frac{3}{4}$. So, in the second case, when we are given the information recorded by the event B , we need to calculate probability of A **given** B . Now we know how to do it:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/2}{3/4} = \frac{2}{3}$$

So now we see how this is not a paradox at all: in one case we calculate the absolute probability of A and in the second we calculate the conditional probability of A . As per the definition above, those are different probability measures (and subsequently we work in different probability spaces!)

Example 23. Ten dice are rolled. What is the probability that 2 or more of them are aces (1's), given that there is at least one ace?

Before we present the solution, note that whenever the sentence is structured like " P given Q ," you immediately know that you need to work in the conditional probability space. Then P will be your event A and Q will be your event B .

Proof. So let A be the event that there are 2 or more aces and let B be the event that there is at least one ace. Let us first construct the appropriate absolute probability space. We define:

$$\Omega = \{(a_1, \dots, a_{10}) : a_i \in \{1, \dots, 6\}, i \in \mathbb{N}\}$$

Again, since Ω is finite, we have the "luxury" to use $\mathcal{F} = 2^\Omega$. Since not stated otherwise, we assume that all outcomes are equally likely – and thus we'll equip our space with the discrete uniform measure $\mathbb{P} = \sum_{\omega \in \Omega} \frac{1}{|\Omega|} \delta_\omega$. In our case, $|\Omega| = 6^{10}$, so:

$$\mathbb{P} = \sum_{\omega \in \Omega} \frac{1}{|\Omega|} \delta_\omega = \sum_{\omega \in \Omega} \frac{1}{6^{10}} \delta_\omega$$

So, we can calculate probabilities we need separately:

$$\mathbb{P}(B) = \sum_{\omega \in \Omega} \frac{1}{6^{10}} \delta_\omega(B) = \frac{|B|}{|\Omega|} = \frac{|\Omega| - |\Omega \setminus B|}{|\Omega|} = \frac{6^{10} - 5^{10}}{6^{10}}$$

Also:

$$\mathbb{P}(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{6^{10} - 5^{10} - 10 \cdot 5^9}{6^{10}}$$

Finally, we can conclude:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{6 - 5^{10} - 10 \cdot 5^9}{6^{10} - 5^{10}} \approx 0.61$$

□

Example 24. [Sampling without replacement] Take two *ordered* samples from $\{1, \dots, n\}$ *without replacement*. Suppose $i \neq j$ are two numbers in $\{1, \dots, n\}$. What is the probability that the second sample is j , given that the first sample is i ?

Proof. Again, we know how to turn this sentence in English into the language of probability theory. Let A denote the event that the second sample is j and B the event that the first sample is i . The number we would like to calculate is $\mathbb{P}(A|B)$. The absolute probability space we are working in is $(\Omega, \mathcal{F}, \mathbb{P})$ where each component is defined as follows:

$$\Omega = \{(a_1, a_2) : i, j \in \{1, \dots, n\}, a_1 \neq a_2\}$$

Note that Ω is exactly $X_{n,2}$ from your Homework 1. Again, we let $\mathcal{F} = 2^\Omega$ and the probability measure is the discrete uniform measure. Notice that

$$A = \{(x, j) : x \in \{1, \dots, n\}, x \neq j\}, \quad B = \{(i, x) : x \in \{1, \dots, n\}, x \neq i\}.$$

So $|A| = |B| = n - 1$, and $|A \cap B| = 1$. Again, we calculate probabilities separately:

$$\mathbb{P}(B) = \frac{|B|}{|\Omega|} = \frac{n-1}{n(n-1)} = \frac{1}{n}$$

Also:

$$\mathbb{P}(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{1}{n(n-1)}$$

Therefore:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{n-1}$$

□

Let us quote a nice remark of Feller (An Introduction to Probability Theory and Its Applications, vol1, 3ed, pp117):

*This expresses the fact that the second choice refers to a population of $n - 1$ elements, each of which has the same probability of being chosen. In fact, the most natural **definition** of random sampling is: “Whatever the first r choices, at the $(r + 1)$ st step each of the remaining $n - r$ elements has probability $1/(n - r)$ to be chosen.”*

Definition 11. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that $A, B \in \mathcal{F}$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

We need to remark two things here:

1. If $\mathbb{P}(B) > 0$, then A and B are independent if and only if $P(A) = P(A|B)$. Intuitively we can think of independence as the situation when the occurrence of one event does not affect the probability of the occurrence of the other.
2. A and B being independent **does not imply** $A \cap B = \emptyset$. To see why, note that whenever $\mathbb{P}(A), \mathbb{P}(B) > 0$ and A and B are independent, then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) > 0$, and so $A \cap B \neq \emptyset$. So, in many cases, independence implies non-empty intersection!

Example 25. 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the infinite coin toss model with $p \in (0, 1)$. Remember that we proposed the following:

$$\mathbb{P}\{X_{n_1} = a_{n_1}, \dots, X_{n_k} = a_{n_k}\} = p^{\sum_{i=1}^{\infty} a_{n_i}} q^{k - \sum_{i=1}^{\infty} a_{n_i}}$$

Now we can see that this was exactly because we assumed that all random variables X_i for all $i \in \mathbb{N}$ were independent of each other. Indeed, what happens in the i th toss has no affect on what happens in the j th toss for all $i \neq j$. In other words, the events $\{X_i = 1\}$ and $\{X_j = 0\}$ are independent and thus $\mathbb{P}\{X_i = 1, X_j = 0\} = \mathbb{P}\{X_i = 1\}\mathbb{P}\{X_j = 0\} = pq$. Same reasoning clearly works for bigger events.

2. If we look at the Example 24 again, we can see that A and B are **not** independent, since:

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{1}{n} \cdot \frac{1}{n} \neq \frac{1}{n} \cdot \frac{1}{n-1} = \mathbb{P}(A \cap B)$$

Intuitively, it also should make sense to you: when sampling without replacement, each draw affects the possible outcomes of subsequent draws.

Theorem 8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A, B_1, B_2, \dots \in \mathcal{F}$. Suppose that $A \subseteq \bigcup_{n=1}^{\infty} B_n$. Show that:

1. $\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A \cap B_n)$;
2. If moreover, $\mathbb{P}(B_n) > 0$ for all n , then $\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) \mathbb{P}(A|B_n)$.

Proof. To prove (1), note that since $A \subseteq \bigcup_{n=1}^{\infty} B_n$, we have that $A = A \cap \left(\bigcup_{n=1}^{\infty} B_n \right)$. Using the distributive law yields $A = A \cap \left(\bigcup_{n=1}^{\infty} B_n \right) = \bigcup_{n=1}^{\infty} (A \cap B_n)$. Therefore, by the second axiom of being a probability measure, we get:

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A \cap B_n)$$

To prove (2), recall that, by definition, $\mathbb{P}(A|B_n) = \frac{\mathbb{P}(A \cap B_n)}{\mathbb{P}(B_n)}$. Therefore, $\mathbb{P}(A \cap B_n) = \mathbb{P}(A|B_n) \mathbb{P}(B_n)$. Hence, by (a), we get the desired claim. \square

The formula in (2) above turns out to be very important in probability theory. It is called the *Law of Total Probability*. As an example, suppose A represents some kind of symptom, and B_1, B_2, \dots represent various diseases this symptom could be associated with. So if we want to know the probability that someone has the symptom, you would have to take into an account all possible diseases that cause it. Moreover, you might want to ask a question like: What is the probability that someone has disease B_k , given they have the symptom A ? Turns out, the Law of Total Probability can be used to answer this as well (look up Bayes' rule, if curious!)

Example 26. We are given two urns, I and II. Suppose urn I has a white balls and b red balls, and suppose urn II has α white balls and β red balls, where $a, b, \alpha, \beta \geq 1$. Randomly draw one ball from urn I, put it into urn II. Then, randomly pick a ball from urn II and examine its color. What is the probability it is white?

Proof. First, we need to define our events. Let A denote the event that the last (examined) ball is white. Let B_1 be the event that the first drawn ball is white. Let B_2 be the event that the first drawn ball is red. Clearly, $A \subseteq B_1 \cup B_2$. By the Law of Total Probability, we have:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A|B_1) \mathbb{P}(B_1) + \mathbb{P}(A|B_2) \mathbb{P}(B_2) \\ &= \frac{\alpha + 1}{\alpha + \beta + 1} \cdot \frac{a}{a + b} + \frac{\alpha}{\alpha + \beta + 1} \cdot \frac{b}{a + b} \end{aligned}$$

\square

Now, let us consider a concrete example when $\alpha = a = 2$ and $\beta = b = 1$. By the formula we've just computed, we get that $\mathbb{P}(A) = \frac{2}{3}$. Let us try to compute the same probability combinatorially, by defining the appropriate sample space and explicitly looking at all possible events.

Let us label all the balls in two urns, to be able to distinguish them in the sample space. Then suppose the distribution of balls is as follows:

Urn I: ①, ② are white and ③ is red.

Urn II: ④, ⑤ are white and ⑥ is red.

Then, the sample space Ω is given as follows:

$$\Omega = \left\{ \begin{array}{ccc} (1, 4), & (2, 4), & (3, 4), \\ (1, 5), & (2, 5), & (3, 5), \\ (1, 6), & (2, 6), & (3, 6), \\ (1, 1), & (2, 2), & (3, 3) \end{array} \right\}$$

We will let $\mathcal{F} = 2^\Omega$ and $\mathbb{P} = \sum_{\omega \in \Omega} \frac{1}{12} \delta_\omega$, as per usual. The desirable outcomes for us are all the tuples that end in 1, 2, 4, or 5 (colored in blue). There are exactly 8 of them, and the total size of Ω is 12. Therefore, we can see that the probability is indeed $\frac{8}{12} = \frac{2}{3}$.

Let us now twist the problem a little bit: we are still given two urns, I and II. Suppose urn I has a white balls and b red balls, and suppose urn II has α white balls and β red balls, where $a, b, \alpha, \beta \geq 1$. Randomly draw **two** balls from urn I, put them into urn II. Then, randomly pick **two** balls from urn II and examine their colors. What is the probability both of them are white?

Proof. Again, we'll define the events first. Let A denote the event that both examined balls are white. Let B_1 be the event that in the first draw, there are two white balls. Let B_2 be the event that in the first draw, there is one white and one red ball. Let B_3 be the event that in the first draw, there are two red balls. Then, again, by the Law of Total Probability, we get:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2) + \mathbb{P}(A|B_3)\mathbb{P}(B_3) \\ &= \frac{\binom{a}{2}}{\binom{a+b}{2}} \cdot \frac{\binom{\alpha+2}{2}}{\binom{\alpha+\beta+2}{2}} + \frac{\binom{a}{1}\binom{b}{1}}{\binom{a+b}{2}} \cdot \frac{\binom{\alpha+1}{2}}{\binom{\alpha+\beta+2}{2}} + \frac{\binom{b}{2}}{\binom{a+b}{2}} \cdot \frac{\binom{\alpha}{2}}{\binom{\alpha+\beta+2}{2}} \end{aligned}$$

□

2 Discrete and Continuous Random Variables

2.1 Introduction to Random Variables

Before we introduce the notion of random variable, let us review the cumulative distribution function (cdf) we talked about last time. Recall that we defined the Bernoulli distribution as follows: $p\delta_1 + q\delta_0$ where $p, q \in [0, 1]$ and $p + q = 1$. Let us see what the cdf of this distribution looks like:

$$F(x) = \mathbb{P}((-\infty, x]) = \begin{cases} 0 & \text{if } x < 0 \\ q & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

If you plot the above function, you will see that it is a step function that is right-continuous, but not left-continuous. Now, let us start talking about a very important notion in probability theory – random variables.

Definition 12. A **random variable** on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $\xi : \Omega \rightarrow \mathbb{R}$ that is $(\mathcal{F}, \mathcal{B})$ -measurable, i.e. $\xi^{\leftarrow}(B) = \{\omega \in \Omega : \xi(\omega) \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}$.

Convention 3. Since the notation gets cumbersome quite fast with all the set and pre-image symbols, we will establish the following conventions:

- $\{\omega \in \Omega : \xi(\omega) \in B\} = \{\xi \in B\} \subseteq \Omega$;
- $\xi^{\leftarrow}((-\infty, x]) = \{\omega \in \Omega : \xi(\omega) \in (-\infty, x]\} = \{\xi \in (-\infty, x]\} = \{\xi \leq x\} \subseteq \Omega$

Proposition 6. 6.3 The following are true of random variables:

1. $\xi : \Omega \rightarrow \mathbb{R}$ is a random variable if and only if $\{\xi \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$.
2. If $\mathcal{F} = 2^\Omega$, then any $\xi : \Omega \rightarrow \mathbb{R}$ is a random variable.

Proof. The proof of (1) is on your homework, and the proof of (2) is trivial: if the σ -algebra is the whole power set, then every set in it is measurable. \square

Notice that by (1) random variables are exactly those functions for which the question “what is the probability that the value of the function is $\leq x$?” makes sense, for any $x \in \mathbb{R}$.

Example 27. Let $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}(0, 1)$ and \mathbb{P} be the Lebesgue measure, i.e. $\mathbb{P}([x, y]) = y - x$ for all $0 < x \leq y < 1$. So, $(\Omega, \mathcal{F}, \mathbb{P})$ is our probability space. Then the following are examples of random variables:

1. $\xi : \Omega \rightarrow \mathbb{R}$ where $\xi(\omega) = a + (b - a)\omega$ is a random variable ($a < b$).
2. $\eta : \Omega \rightarrow \mathbb{R}$ where $\eta(\omega) = \omega^2$ is a random variable.

Example 28. Let $\Omega = \{(a_n)_{n=1}^\infty : a_n \in \{0, 1\}\}$ where $p \in [0, 1], p + q = 1$. We would like to model an unlimited sequence of coin tosses for which each toss has probability p of being a head. In this course, we will use head (vs. tail) interchangeably with success (vs. failure) and 1 (vs. 0). Then let $X_n(\omega) = a_n$ for all $\omega \in \Omega, n \in \mathbb{N}$. Note that at this point, X_n is not a random variable *yet*, as we have not specified the probability space – the σ -algebra and the measure.

Consider the following table that can give you some intuition of what kind of events we want to have in our probability space and what their "expected" or "proposed" probability would be:

"Events"	"Expected" or "Proposed" Probability
$\{X_1 = 1\}$	p
$\{X_1 = 1, X_2 = 1\}$	pq
$\{X_1 = 0, X_3 = 0\}$	pq
$\{X_{n_1} = a_{n_1}, \dots, X_{n_k} = a_{n_k}\}$	$p^{\sum_{i=1}^k a_{n_i}} q^{k - \sum_{i=1}^k a_{n_i}}$

Note that the third row can be justified as follows:

$$\{X_1 = 1, X_3 = 1\} = \{X_1 = 0, X_2 = 1, X_3 = 0\} \uplus \{X_1 = 0, X_2 = 0, X_3 = 0\}$$

must have probability

$$ppq + pq = pq(p + q) = pq$$

Also note that $\sum_{i=1}^k a_{n_i}$ in the fourth row is the number of successes (or 1's) that appear. Now it is time to build the probability model on Ω . To do so, we define:

$$\mathcal{E} = \left\{ \{X_{n_1} = a_{n_1}, \dots, X_{n_k} = a_{n_k}\} : k \in \mathbb{Z}^+, n_1 < n_2 < \dots < n_k, a_{n_i} \in \{0, 1\} \right\} \subseteq 2^\Omega$$

We will take the σ -algebra generated by \mathcal{E} to be \mathcal{F} . We will note that $\mathcal{F} = \sigma(\mathcal{E}) \neq 2^\Omega$. This fact is non-trivial, but we won't prove it here.

Theorem 9. For each $p \in (0, 1)$, there is a unique probability measure \mathbb{P} on (Ω, \mathcal{F}) such that for all $k \in \mathbb{Z}^+, n_1 < \dots < n_k, a_{n_i} \in \{0, 1\}$, we have:

$$\mathbb{P}(\{X_{n_1} = a_{n_1}, \dots, X_{n_k} = a_{n_k}\}) = p^{\sum_{i=1}^k a_{n_i}} q^{k - \sum_{i=1}^k a_{n_i}}$$

Definition 13. We will call $(\Omega, \mathcal{F}, \mathbb{P})$ above the **infinite coin toss model** with parameter $0 < p < 1$. (This is a non-standard terminology).

Example 29. The followings are examples of random variables in $(\Omega, \mathcal{F}, \mathbb{P})$.

- Let $X_n : \Omega \rightarrow \mathbb{R}, X_n(\omega) = a_n$ where $\omega = (a_n)_{n=1}^\infty$. Then the random variable X_n represents the result of the n -th trial for each sample ω
- Define $S_n : \Omega \rightarrow \mathbb{R}$ as $S_n(\omega) = X_1(\omega) + \dots + X_n(\omega) = a_1 + \dots + a_n$. Note that the random variable S_n represents the cumulative number of successes in the first n trials (or the number of heads in the first n coin tosses) of each sample ω .

- Define $\eta : \Omega \rightarrow \mathbb{R}$ as follows:

$$\eta(\omega) = \begin{cases} 0 & \text{if } \omega = (0, 0, \dots, 0, \dots) \\ n & \text{if } \omega = (a_n)_{n=1}^{\infty} \text{ s.t. } a_1 = \dots = a_{n-1} = 0, a_n = 1 \end{cases}$$

The random variable η represents the epoch of the first success of each sample ω . (Notice that a success exists in every sample other than $\omega = (0, 0, \dots, 0, \dots)$.)

- Define $W_k : \Omega \rightarrow \mathbb{R}$ as:

$$W_k(\omega) = \begin{cases} 0 & \text{if } \sum_{i=1}^{\infty} a_i < k \\ n & \text{if } \sum_{i=1}^{n-1} a_i = k-1, a_n = 1 \end{cases}$$

The random variable W_k represents the epoch of the k -th success for each sample ω (whenever it exists). Also note that $\eta = W_1$.

Here is an intuitive way to think about random variables. In a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, each sample $\omega \in \Omega$ carries a lot of information, and a random variable ξ reveals one specific feature (like the cumulative number of Heads in the first n coin tosses or the epoch of the k -th success) of each sample ω by recording it as a number $\xi(\omega)$.

Definition 14. Let $\xi : \Omega \rightarrow \mathbb{R}$ be a random variable.

1. The push-forward measure

$$\mathbb{P}_{\xi}(B) = \mathbb{P}(\{\xi \in B\})$$

for all $B \in \mathcal{B}$ is called the **law or probability distribution** of ξ .

2. The cdf of \mathbb{P}_{ξ} , that is, the function $F_{\xi}(x) = \mathbb{P}_{\xi}((-\infty, x]) = \mathbb{P}(\{\xi \leq x\})$, is called the cdf of ξ .

It is important to keep in mind that the law of a random variable is a one dimensional probability distribution, or equivalently a probability measure defined on $(\mathbb{R}, \mathcal{B})$ instead of the original space (Ω, \mathcal{F}) . When we talked about one-dimensional distributions, we mentioned that a one dimensional probability distribution is uniquely determined (or characterized) by its cdf. Because F_{ξ} is, by definition, the cdf of \mathbb{P}_{ξ} , once we determine F_{ξ} then the law \mathbb{P}_{ξ} is also uniquely determined. Now let us go back to the example [27](#) and recall the random

variable $\xi(\omega) = a + (b - a)\omega$. Let us compute its cdf:

$$\begin{aligned}
 F_\xi(x) &= \mathbb{P}\{\xi \leq x\} \\
 &= \mathbb{P}\{\omega \in (0, 1) : a + (b - a)\omega \leq x\} \\
 &= \mathbb{P}\left\{\omega \in (0, 1) : \omega \leq \frac{x - a}{b - a}\right\} \\
 &= \mathbb{P}\left(\left(-\infty, \frac{x - a}{b - a}\right] \cap (0, 1)\right) \\
 &= \begin{cases} 0 & \text{if } x < a \\ \frac{x - a}{b - a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}
 \end{aligned}$$

Surprise! ξ is a uniformly distributed random variable (because F_ξ matches the cdf characterization of uniform distribution).

Now let us calculate the cdf of $\eta(\omega) = \omega^2$:

$$\begin{aligned}
 F_\eta(x) &= \mathbb{P}\{\eta \leq x\} \\
 &= \mathbb{P}\{\omega \in (0, 1) : \omega^2 \leq x\} \\
 &= \mathbb{P}\{\omega \in (0, 1) : |\omega| \leq \sqrt{x}\} \\
 &= \mathbb{P}([-\sqrt{x}, \sqrt{x}] \cap (0, 1)) \\
 &= \begin{cases} 0 & \text{if } x < 0 \\ \sqrt{x} & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}
 \end{aligned}$$

For discrete random variables, i.e. those taking only countably (or finitely) many values, we can describe the law explicitly as follows:

Extra-Credit 1: Prove the following:

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\xi : \Omega \rightarrow \mathbb{R}$ be a random variable. Suppose that ξ only takes values in the countable (denumerable) set x_1, x_2, x_3, \dots where $x_i \neq x_j$ for all $i \neq j$. Then the law of ξ is:

$$\sum_{n=1}^{\infty} \mathbb{P}\{\xi = x_n\} \delta_{x_n}$$

The extra credit problem is worth 3 points of your semester total (1.5% of your grade). You can discuss general concepts (e.g. what the Dirac Delta measure is, how combinations of such measures work, etc.) with anyone, but not any technical aspects related to this problem; you're not allowed to collaborate on it with anyone; no internet. Put as many details in your solutions as possible, and justify every step and every equality you derive. The assignment

will be graded by Prof. Li for correctness. Due Tue 9/25 at the beginning of class. (No late submissions!)

Now, let's go back to math. If we assume the above result, we can compute the law of $X_n(\omega) = a_n$ from Example 29:

$$\mathbb{P}\{X_n = 1\}\delta_1 + \mathbb{P}\{X_n = 0\}\delta_0 = p\delta_1 + q\delta_0$$

Note that the law of the random variable X is the Bernoulli distribution of parameter p .

Let us revise the infinite coin toss model again. Let $p \in (0, 1)$ and let $(\Omega, \mathcal{F}, \mathbb{P})$ be our probability space where $\Omega = \{(a_n)_{n=1}^\infty, a_n \in \{0, 1\}\}$ and $\mathcal{F} = \sigma(\mathcal{E})$ where \mathcal{E} is defined as:

$$\mathcal{E} = \left\{ X_{n_1} = a_1, \dots, X_{n_k} = a_k : k \in \mathbb{Z}^+, n_1 < n_2 < \dots < n_k, a_{n_i} \in \{0, 1\} \right\} \subseteq 2^\Omega$$

Remember that we said that $\mathbb{P}\{X_{n_1} = a_{n_1}, \dots, X_{n_k} = a_{n_k}\} = p^{\sum_{i=1}^k a_{n_i}} q^{1 - \sum_{i=1}^k a_{n_i}} \leq (\max\{p, q\})^k$. An example of a concrete event in the probability space would be $\{X_1 = 1, X_3 = 0\} = \{(1, a_2, 0, a_4, \dots) : a_i \in \{0, 1\}\} \subseteq \Omega$. Using the "formula" above, it is clear that the probability of this event is pq . Let us explore this space a little bit more and see what other events can take place and what their probabilities are.

Proposition 7. The following are true in $(\Omega, \mathcal{F}, \mathbb{P})$ defined above:

1. For all $\omega \in \Omega$, $\{\omega\} \in \mathcal{F}$ and $\mathbb{P}(\{\omega\}) = 0$.
2. For any finite subset $A \subseteq \Omega$ such that $A \in \mathcal{F}$, $\mathbb{P}(A) = 0$.

Proof. To prove (1), let $\omega = (a_1, a_2, \dots)$. Note that $\omega \notin \mathcal{E}$, since the tuple is infinite. For all $n \in \mathbb{N}$, define $A_n = \{X_1 = a_1, \dots, X_n = a_n\} \in \mathcal{E} \subseteq \mathcal{F}$. Notice that $\{\omega\} = \bigcap_{n=1}^\infty A_n$, so $\{\omega\} \in \mathcal{F}$, since \mathcal{F} is a σ -algebra and thus is closed under countable intersection. Notice now that A_n is a decreasing sequence of sets, i.e. $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$, so by Theorem 6, we obtain:

$$0 \leq \mathbb{P}(\{\omega\}) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n) \leq (\max\{p, q\})^n \rightarrow 0$$

Since $\lim_{n \rightarrow \infty} (\max\{p, q\})^n = 0$ (because $p, q < 1$), by Squeeze theorem, we conclude that $\mathbb{P}(\{\omega\}) = 0$.

To prove (2), note that any finite set can be represented as a union of singletons. By (1), each singleton is an event in \mathcal{F} and, since \mathcal{F} is a σ -algebra, it is closed under finite unions. Hence:

$$A = \{\omega_1, \dots, \omega_n\} = \biguplus_{i \in \{1, \dots, n\}} \{\omega_i\} \subseteq \mathcal{F}$$

Since the union is clearly disjoint, by finite additivity of probability measure, we obtain $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(\{\omega_i\}) = 0$. \square

Proposition 8. Let $A = \{\omega = (a_n)_{n=1}^\infty : \sum_{n=1}^\infty a_n < \infty\}$. Then $A \in \mathcal{F}$ and $\mathbb{P}(A) = 0$.

Before we prove the above proposition, it is important to understand that A represents the event that in the sequence of infinitely many coin tosses, we get only finitely many heads. In other words, the proposition is saying that with probability 1, heads will show up infinitely many times in such an experiment.

Proof. For each $n \in \mathbb{N}$, define $A_n = \{(a_n)_{n=1}^\infty : a_{n+1} = a_{n+2} = \dots = 0\} \in \mathcal{F}$ (since it is a finite set). For each n , A_n represents the event that after the n th toss, we get tails only. Then $|A_n| = 2^n$ for all n , since there are only n positions of freedom in each case. So each A_n is finite, and thus has probability 0, as per Proposition 7. Note that $A = \bigcup_{n=1}^\infty A_n \subseteq \mathcal{F}$. So, by σ -subadditivity of probability measure, we obtain that

$$0 \leq \mathbb{P}(A) \leq \sum_{n=1}^\infty \mathbb{P}(A_n) = 0$$

So indeed $\mathbb{P}(A) = 0$, as desired. \square

Definition 15. Let $S \subseteq \mathbb{R}$. We say that a random variable ξ on a general probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is S -valued if $\xi(\omega) \in S$ for all $\omega \in \Omega$. If moreover S is countable, then ξ is called discrete.

Theorem 10. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $S \subseteq \mathbb{R}$ be a countable set. Let $\xi : \Omega \rightarrow S$ be a function such that $\{\xi = x\} \in \mathcal{F}$ for all $x \in S$. Then:

1. ξ is a random variable;
2. The law of ξ satisfies

$$\mathbb{P}_\xi = \sum_{x \in S} \mathbb{P}\{\xi = x\} \delta_x$$

Before we prove the above proposition, we will emphasize it once again that the law completely answers questions like "what is the probability that $a < \xi < b$ " or "what is the probability that $\xi \leq x$ ", because $\mathbb{P}\{a < \xi < b\} = \mathbb{P}_\xi(a, b)$ and $\mathbb{P}\{\xi \leq x\} = \mathbb{P}_\xi((-\infty, x])$. More generally, for any Borel set B , the probability that ξ is in B is expressed as $\mathbb{P}(\{\xi \in B\}) = \mathbb{P}_\xi(B)$.

Proof. To prove (1), let $B \in \mathcal{B}$. Then:

$$\begin{aligned} \xi^\leftarrow(B) &= \{\xi \in B\} \\ &= \{\omega \in \Omega : \xi(\omega) \in B\} \\ &= \{\xi \in B \cap S\} && \text{since } \xi \text{ is } S\text{-valued} \\ &= \biguplus_{x \in B \cap S} \{\xi = x\} \end{aligned}$$

Note that for all $x \in S$, $\{\xi = x\} \in \mathcal{F}$ by assumption and the union above is countable, as S is so. Since \mathcal{F} is closed under countable unions, we conclude that $\xi^\leftarrow(B) \in \mathcal{F}$ and hence indeed ξ is a random variable.

To prove (2), let $B \in \mathcal{B}$. Note that the right-hand side can be written as:

$$\begin{aligned} \left(\sum_{x \in S} \mathbb{P}\{\xi = x\} \delta_x \right)(B) &= \sum_{x \in S} \mathbb{P}\{\xi = x\} \delta_x(B) \\ &= \sum_{x \in S \cap B} \mathbb{P}\{\xi = x\} \end{aligned}$$

The left-hand side can in turn be written as:

$$\begin{aligned} \mathbb{P}_\xi(B) &= \mathbb{P}(\{\xi \in B\}) && \text{by definition} \\ &= \mathbb{P}\left(\biguplus_{x \in B \cap S} \{\xi = x\} \right) && \text{by part (1)} \\ &= \sum_{x \in B \cap S} \mathbb{P}\{\xi = x\} && S \text{ is countable} \end{aligned}$$

Notice that LHS=RHS and this concludes the proof. \square

Example 30. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the infinite coin toss model with $p \in (0, 1)$ and $\omega = (a_n)_{n=1}^\infty$.

1. Consider the random variable $X_n(\omega) = a_n$. Note that it reveals whether the n th coin toss is a head or a tail. Clearly, $S = \{0, 1\}$, i.e. X_n is $\{0, 1\}$ -valued. Then, since we know that $\mathbb{P}\{X_n = 0\} = q$ and $\mathbb{P}\{X_n = 1\} = p$, Theorem 10 tells us that $\mathbb{P}_{X_n} = q\delta_0 + p\delta_1$.
2. Consider the random variable $S_n(\omega) = a_1 + \dots + a_n$, which represents the accumulated number of heads up to epoch n . Let $k \in S = \{0, 1, \dots, n\}$. Then, intuitively, we know that $\mathbb{P}\{S_n = k\} = \binom{n}{k} p^k q^{n-k}$ (i.e. S_n in fact follows the Binomial Distribution). You can convince yourself this is true by considering the case when $k = 1$. However, we will also make this rigorous. First, for some set $A \subseteq \mathbb{N}$, we will define the indicator function $\mathbb{1}_A : \mathbb{N} \rightarrow \{0, 1\}$ as follows:

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

Now, let $Y_{n,k} = \{A \subseteq \{1, \dots, n\} : |A| = k\}$, like in the homework. Spend some time to convince yourself that:

$$\{S_n = k\} = \biguplus_{A \in Y_{n,k}} \{X_i = \mathbb{1}_A(i) : i \in \{1, \dots, n\}\}$$

The intuitive way to think about it is to note that the union over all elements of $Y_{n,k}$ gives you all possible arrangements of the results of n coin tosses such that exactly k of them are heads (or 1's). That is because each element of $Y_{n,k}$ tells you the indices of the k positions that are occupied with heads – and the rest have to be tails. Using the indicator function just gives us a way to express this in terms of the values of random variables we are interested in and ensure that each of the terms of the big union is

easily seen to be an element of \mathcal{E} , the generating set of \mathcal{F} . After expressing our event in these terms, we can easily compute the probability:

$$\begin{aligned}\mathbb{P}\{S_n = k\} &= \sum_{A \in Y_{n,k}} \mathbb{P}\{X_i = \mathbb{1}_A(i); 1 \leq i \leq n\} \\ &= \sum_{A \in Y_{n,k}} p^k q^{n-k} \\ &= p^k q^{n-k} |Y_{n,k}| \\ &= \binom{n}{k} p^k q^{n-k}\end{aligned}$$

Now we can also calculate the law of S_n :

$$\begin{aligned}\mathbb{P}_{S_n} &= \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \delta_k \\ &= \sum_{k=0}^n \mathbb{P}\{S_n = k\} \delta_k\end{aligned}$$

Extra Credit 2: Consider the random variable

$$W_r(\omega) = \begin{cases} 0 & \text{if } \sum_{n=1}^{\infty} a_n < r \\ n & \text{the epoch of the } r\text{th success} \end{cases}$$

Note that W_r can take countably many values.

1. Show that W_r is a random variable.
2. Compute, with a proof, the law of W_r .

Proof. The crucial thing to notice is that W_r can be re-written as:

$$W_r = \begin{cases} 0 & \text{if } \sum_{i=1}^{\infty} a_i < k \\ n & \text{if } S_{n-1} = r-1, a_n = 1 \end{cases}$$

To give a concrete example, consider a sample $\omega = (1, 0, 0, 1, 0, 0, 0, 0, 1, \dots)$. Then $W_3(\omega) = 9$, since the third success happens on the 9th toss. \square

First, we need to find the range of W_r . It is clear that for any sample $\omega \in \Omega$, we have $W_r(\omega) \in \{0\} \cup \{r, r+1, r+2, \dots\}$, and hence it is our range. So, what we would like to show is that $\{W_r = k\} \in \mathcal{F}$ for all k in the range, to show that it is the random variable (this is equivalent to the definition of measurability, since W_r takes only countable many values).

Case $k = 0$: Observe that:

$$\begin{aligned}\{W_r = 0\} &= \{\omega \in \Omega : W_r(\omega) = 0\} \\ &= \{\omega \in \Omega : \sum_{n=1}^{\infty} a_n < r\} \\ &\subseteq \{\omega \in \Omega : \sum_{n=1}^{\infty} a_n < +\infty\}\end{aligned}$$

The latter is countable, since it can be represented as a union of countable sets (A_n where the sum equals n for all $n \in \mathbb{N}$) and thus the desired set is countable as well. Hence, we can enumerate it: $\{W_r = 0\} = \{\omega_1, \omega_2, \dots\} = \biguplus_{n=1}^{\infty} \{\omega_n\} \in \mathcal{F}$, so $\mathbb{P}\{W_r = 0\} = 0$, since we have proven in before that the probability of each singleton in the infinite coin toss model is zero.

Case $k \neq 0$: Recall that in the last section we defined $Y_{n,k}$ as $Y_{n,k} = \{A \subseteq \{1, \dots, n\} : |A| = k\}$ and you've proven in the homework that $|Y_{n,k}| = \binom{n}{k}$

$$\begin{aligned}\{W_r = k\} &= \{S_{k-1} = r-1, X_k = 1\} \\ &= \{\omega \in \Omega : S_{k-1}(\omega) = r-1, X_k(\omega) = 1\} \\ &= \{S_{k-1} = r-1\} \cap \{X_k = 1\} \\ &= \left(\biguplus_{A \in Y_{k-1, r-1}} \{X_i = \mathbb{1}_A(i) : 1 \leq i \leq k-1\} \right) \cap \{X_k = 1\} \\ &= \biguplus_{A \in Y_{k-1, r-1}} \{X_i = \mathbb{1}_A(i) \text{ and } X_k = 1 : 1 \leq i \leq k-1\}\end{aligned}$$

So each of these sets is easily seen to be a member of the generating set and so their countable union is in the σ -algebra. Indeed, then, W_r is a random variable. Now note that

$$\begin{aligned}\mathbb{P}\{W_r = k\} &= \sum_{A \in Y_{k-1, r-1}} \mathbb{P}\{X_i = \mathbb{1}_A(i) \text{ and } X_k = 1 : 1 \leq i \leq k-1\} \\ &= \sum_{A \in Y_{k-1, r-1}} p^r q^{k-r} \\ &= \binom{k-1}{r-1} p^r q^{k-r} \\ &= \mathbb{P}\{S_{k-1} = r-1\} \cdot \mathbb{P}\{X_k = 1\}.\end{aligned}$$

In other words, we have just proved, as a byproduct, that the two events $\{S_{k-1} = r-1\}$ and $\{X_k = 1\}$ are independent. So, finally we conclude that

$$\mathbb{P}_{W_r} = \sum_{k=r}^{\infty} \binom{k-1}{r-1} p^r q^{k-r} \delta_k$$

This distribution has a name – it is called the **Negative Binomial Distribution** with parameters r and p , denoted by $\text{NB}(r, p)$. It is good to observe that $\text{NB}(1, p) = \sum_{k=1}^{\infty} pq^{k-1}\delta_k = \text{Geo}(p)$. That is, the geometric distribution is the law of the waiting time (which is the third of Example 29) for the first success in an unlimited sequence of coin tosses.

2.2 Continuous Probability Distributions

Recall that the set of all distribution functions on \mathbb{R} is defined as follows:

$$\text{Dist}(\mathbb{R}) = \left\{ F(x) : \begin{array}{l} \textcircled{1} F(x) \leq F(y) \quad \forall x < y; \\ \textcircled{2} F \text{ is right-continuous;} \\ \textcircled{3} \lim_{n \rightarrow \infty} F(x) = 1; \\ \textcircled{4} \lim_{n \rightarrow \infty} F(x) = 0 \end{array} \right\}$$

Recall that in the homework you proved that there is a bijective map between $\text{Prob}(\mathbb{R}, \mathcal{B})$, all probability measures on $(\mathbb{R}, \mathcal{B})$, and all distribution functions, $\text{Dist}(\mathbb{R})$. The map you defined was $\mathbb{P} \mapsto F(x) = \mathbb{P}((-\infty, x])$ sending each one dimensional probability distribution to its cumulative distribution function.

Before now, we've only interacted with probability measures that are combinations of Dirac Delta measures (except uniform distribution). There are more out there. Because of the one-to-one correspondence above, it suffices to produce an “interesting” distribution function and we'll automatically get a new probability measure. Let us begin by stating some definitions.

Definition 16. A *probability density function* (pdf) is a function f , defined on \mathbb{R} , satisfying the following:

1. $f(t) \geq 0$ for all $t \in \mathbb{R}$;
2. $f(t)$ is continuous at all but finitely many points in \mathbb{R} ;
3. $\int_{-\infty}^{\infty} f(t)dt = 1$.

Lemma 3. For any pdf $f(t)$ there exists a unique $\mathbb{P} \in \text{Prob}(\mathbb{R}, \mathcal{B})$ such that $\mathbb{P}((-\infty, x]) = \int_{-\infty}^x f(t)dt$.

Proof. Trivial! (Not really, but you will see it in Math 225). □

It is worth noting that part (2), i.e. piece-wise continuity is OK, because in reality what we're doing is looking at the area under the curve and finitely many discontinuities won't affect anything, since area below those points is ε -small.

Definition 17.

1. We say that $\mathbb{P} \in \text{Prob}(\mathbb{R}, \mathcal{B})$ is *continuous* if there exists a pdf $f(t)$ such that

$$\mathbb{P}((-\infty, x]) = \int_{-\infty}^x f(t) dt$$

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

2. A random variable ξ is called *continuous* on some probability space $(\Omega, \mathcal{F}, \mu)$ if the law is continuous.

We will call $f(t)$ the *density* of \mathbb{P} and ξ , respectively.

Remark:

1. If f is a density of $\mathbb{P} \in \text{Prob}(\mathbb{R}, \mathcal{B})$ and $g(t) = f(t)$ for all but finitely many $t \in \mathbb{R}$, then $g(t)$ is also a density of \mathbb{P} . (So, density is NOT unique!) This also comes from the fact that we're dealing with integrals and modifications at finitely many points won't affect the area under the curve.
2. If \mathbb{P} has a density $f(t)$, then $\mathbb{P}(B) = \int_B f(t) dt$ for all $B \in \mathcal{B}$. In particular, $\mathbb{P}(\{x\}) = \int_x^x f(t) dt = 0$. This means that the Dirac Delta measure doesn't have a density.

Then, the natural question one may ask is: Given \mathbb{P} , or $F(x)$, how do we decide if \mathbb{P} has a density or not? The following theorem provides the answer.

Theorem 11. Let $\mathbb{P} \in \text{Prob}(\mathbb{R}, \mathcal{B})$ and $F(x)$ be the cdf of \mathbb{P} . Let

$$f(t) = \begin{cases} F'(t) & \text{if } F \text{ is differentiable at } t \\ 0 & \text{otherwise} \end{cases}$$

If $f(t)$ above is a pdf, then \mathbb{P} is continuous with density $f(t)$.

Proof. Omitted \odot . The issue is that, if F is everywhere differentiable and also F' is continuous, then the statement follows from the Fundamental Theorem of Calculus. But in probability theory it is often the case that F fails to be differentiable at finitely many points. So the proof will show that $f(t)$ being a pdf guarantees that the Fundamental Theorem of Calculus still holds true for F . (Take Math 225 and Math 255!). \square

Example 31.

1. Let $U(a, b)$ be the uniform distribution on the interval (a, b) . Recall:

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$$

Using the formula above, we can calculate piece-wise derivatives as follows:

$$f(t) = \begin{cases} 0 & \text{if } t \leq a \\ \frac{1}{b-a} & \text{if } a < t < b \\ 0 & \text{if } t \leq b \end{cases}$$

Note that $F(x)$ is not differentiable at a and b , so at those points $f(t) = 0$ by definition in Theorem 11. Let us check whether $f(t)$ is a density. Observe:

$$\int_a^b \frac{1}{b-a} dt = 1$$

and the other conditions are clearly satisfied. So we conclude from Theorem 11 that the uniform distribution is a continuous probability distribution with a density $f(t)$.

2. Consider the Dirac Delta measure δ_0 . Recall its cdf is given by:

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

When we differentiate this function piece-wise, we get that $f(t) \equiv 0$. Hence, it can't be a pdf, as it doesn't satisfy the third axiom. This is consistent with the fact that δ_0 is not a continuous probability distribution.

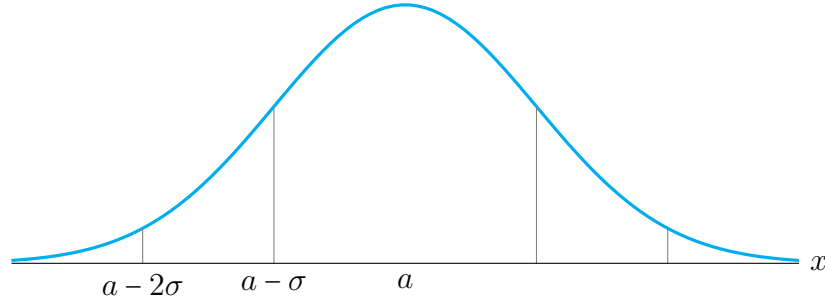
We've established that there is a bijection between distribution functions on \mathbb{R} and probability measures on $(\mathbb{R}, \mathcal{B})$, as long as we can supply an "interesting" distribution function, we'll get a unique "interesting" measure. Today we will see several very important and interesting distributions in probability theory. Also recall that for continuous distributions defined on $(\mathbb{R}, \mathcal{B}, \mathbb{P})$ with the pdf $f(t)$, we have $F(x) = \mathbb{P}((-\infty, x]) = \int_{-\infty}^x f(t)dt$, so each continuous distribution is completely characterized by the pdf, as we can recover its cdf using the formula above.

Definition 18. The following continuous distributions are defined by their pdfs:

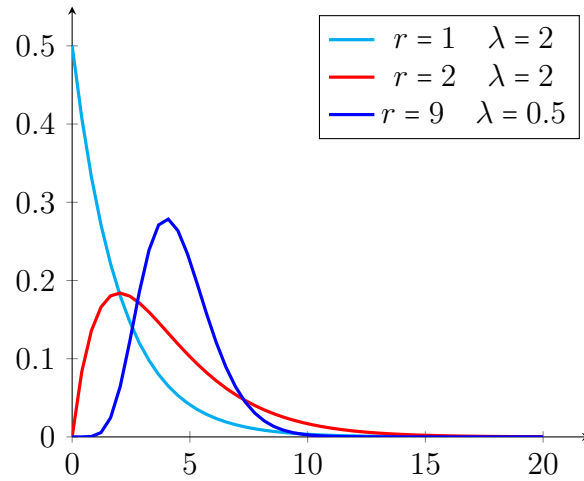
Distribution	Notation	Parameters	pdf
Normal	$\mathcal{N}(a, \sigma^2)$	$a \in \mathbb{R}, \sigma > 0$	$\mathcal{N}_{a, \sigma^2}(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-a)^2}{2\sigma^2}}$
Γ (Gamma)	$G(\lambda, r)$	$\lambda, r > 0$	$g_{\lambda, r}(t) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases}$

where $\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$.

Let us plot these distributions' density functions:. The normal distribution is the well-familiar bell curve:



The Gamma distribution plot is a bit more involved. We will only plot it for $t > 0$, since it is clear what its plot looks like for $t \leq 0$. The plot depends on the value of the parameters. If $r \in (0, 1]$, we get a strictly decreasing function, while if $r > 1$, we get a curve that obtains a local maxima once. The rate of increase/decrease, in turn, depends on the value of the parameter λ . This phenomena is demonstrated below, with the Gamma pdf plotted for different values of r and λ .



Note that both $\mathcal{N}(a, \sigma^2)$ and $G(\lambda, r)$ are probability measures on $(\mathbb{R}, \mathcal{B})$.

Proposition 9. Let ξ be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\xi \sim \mathcal{N}(0, 1)$ [this notation means that $\mathbb{P}_\xi = \mathcal{N}(0, 1)$, so ξ follows the standard normal distribution]. Then:

$$\xi^2 \sim G\left(\frac{1}{2}, \frac{1}{2}\right)$$

.

Proof. Let us first clarify what ξ^2 means. Recall that $\xi : \Omega \rightarrow \mathbb{R}$, so we define $\xi^2 : \Omega \rightarrow \mathbb{R}$ as $\xi^2(\omega) = \xi(\omega) \cdot \xi(\omega)$. It is easy to check that ξ^2 is a random variable (do it!).

Now let us proceed with the proof of the claim. Consider $F_{\xi^2}(x)$, the cdf of ξ^2 . We have:

$$\begin{aligned} F_{\xi^2}(x) &= \mathbb{P}\{\xi^2 \leq x\} \\ &= \mathbb{P}\{\omega \in \Omega : \xi^2(\omega) \leq x\} \end{aligned}$$

We will now consider three cases to compute the cdf explicitly:

Case $x < 0$: Suppose $x < 0$. Then $\{\xi \leq x\} = \emptyset$, since a squared real number is always non-negative. Therefore, $F_{\xi^2}(x) = \mathbb{P}(\emptyset) = 0$.

Case $x = 0$: Suppose $x = 0$. Then $\{\xi \leq x\} = \{\xi = 0\}$. Since we're in the continuous world, the probability at each single point is 0, so we get $\mathbb{P}\{\xi \leq x\} = \mathbb{P}\{\xi = 0\} = 0$.

Case $x > 0$: Suppose $x > 0$. Then $\{\xi^2 \leq x\} = \{-\sqrt{x} \leq \xi \leq \sqrt{x}\}$, so we obtain:

$$\begin{aligned} F_{\xi^2} &= \mathbb{P}\{-\sqrt{x} \leq \xi \leq \sqrt{x}\} \\ &= \mathcal{N}_{0,1}([-\sqrt{x}, \sqrt{x}]) && \text{since } \xi \sim \mathcal{N}(0, 1) \\ &= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt && \text{continuous} \end{aligned}$$

For the last equality, recall that for any continuous probability distribution \mathbb{P} with density $f(t)$, we have

$$\mathbb{P}(B) = \int_B f(t) dt$$

for any Borel set $B \in \mathcal{B}$. Although the definition of the integral over an arbitrary Borel set is quite involved, we are able to give a rigorous proof of this fact for finite intervals. Note that we know that, by definition, $\mathbb{P}((-\infty, x]) = \int_{-\infty}^x f(t) dt$, so:

$$\begin{aligned} \mathbb{P}([a, b]) &= \mathbb{P}((a, b]) && \text{continuous} \\ &= \mathbb{P}((-\infty, b]) - \mathbb{P}((-\infty, a]) \\ &= \int_{-\infty}^b f(t) dt - \int_{-\infty}^a f(t) dt \\ &= \int_a^b f(t) dt \end{aligned}$$

So, the cdf is expressed as follows:

$$F_{\xi^2} = \begin{cases} 0 & \text{if } x \leq 0 \\ \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt & \text{if } x > 0 \end{cases}$$

Therefore, by 11, we'll differentiate piece-wise to obtain the pdf. Clearly, whenever $t < 0$, we have $f(t) = 0$. Suppose $t > 0$. Recall that the Fundamental Theorem of Calculus and Chain Rule tells us the following:

$$\frac{d}{dx} \left(\int_{v(x)}^{u(x)} f(t) dt \right) = f(u(x))u'(x) - f(v(x))v'(x)$$

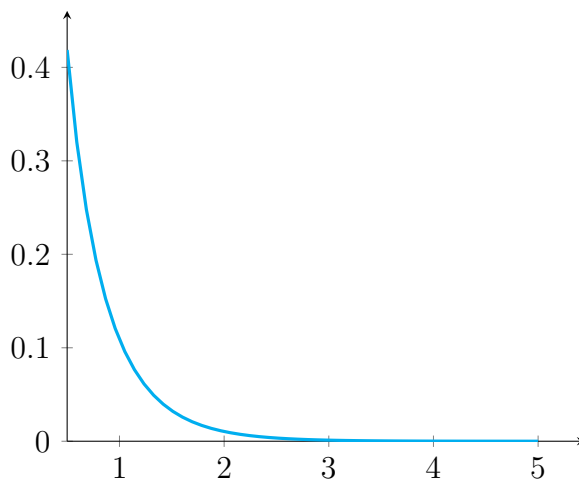
Applying it to our example and taking derivative with respect to x and evaluating at t , we get that for $t > 0$:

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} e^{-t/2} \left(\frac{1}{2} \cdot \frac{1}{\sqrt{t}} \right) + \frac{1}{\sqrt{2\pi}} e^{-t/2} \left(\frac{1}{2} \cdot \frac{1}{\sqrt{t}} \right) &= \frac{1}{\sqrt{2\pi}} e^{-t/2} \frac{1}{\sqrt{t}} \\ &= \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{\pi}} \cdot e^{-t/2} \cdot t^{-1/2} \end{aligned}$$

Now, notice that $G\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{\Gamma(1/2)} \cdot \left(\frac{1}{2}\right)^{1/2} \cdot t^{-1/2} \cdot e^{-t/2}$ for $t > 0$ and 0 otherwise. The only thing we need to verify to show that the equality holds is that $\Gamma(1/2) = \sqrt{\pi}$. Fortunately (!), however, we have done Homework 3, and know that it is the case! Note now that we can stop and not consider the case when $t = 0$, because we see that the functions agree at all but (possibly) one point, so they induce the same probability measure. So indeed the proposition holds. \square

Definition 19. $G\left(\frac{1}{2}, \frac{1}{2}\right)$ is called the χ^2 distribution (read as: Chi Square).

Here's the plot for the χ^2 distribution:



The reason you've probably already heard about the normal distribution before taking this class is that it plays a big role in applied probability theory. However – you may be wondering: how can a continuous distribution be useful when working with finite data sets? The answer is, people found out that the normal distribution happens to nicely approximate discrete

data sets, and, naturally, the larger the size of the set is, the better the approximation is. Let us define two very important concepts: expectation and variance, which will take us a step closer to proving this phenomena.

Definition 20. We will define *expectation* and *variance* for discrete and continuous distributions as follows:

Type of r.v.	Expectation (μ)	Variance (σ^2)
Discrete r.v. ξ with $\mathbb{P}_\xi = \sum_{k=1}^{\infty} \mathbb{P}_k \delta_{x_k}$	$\sum_{k=1}^{\infty} \mathbb{P}_k x_k$	$\sum_{k=1}^{\infty} \mathbb{P}_k (x_k - \mu)^2$
Continuous r.v. with density $f(t)$	$\int_{-\infty}^{+\infty} t f(t) dt$	$\int_{-\infty}^{+\infty} (t - \mu)^2 f(t) dt$

Moreover, we require the above series to converge **absolutely**. (This requirement is necessary to ensure that no matter how we permute the terms of the series, we get the same answer, so that the expectation/variance are well-defined).

It is worth noting that discrete and continuous distributions do not exhaust $\text{Prob}(\mathbb{R}, \mathcal{B})$. There are also singular distributions and distributions of mixed type, but we will not talk about them in this course.

OK, let's now use some real-life intuition to see why expectation is defined in this way.

Example 32. Suppose in a town we have n_k families with exactly k members, $n, k \in \mathbb{N}^+$. Then what is the *average* number of people per family? Back in middle school, we learned that to compute this average you need to divide the number of people by the total number of families. This would give us the expression:

$$\frac{n_1 + 2n_2 + \cdots + \ell n_\ell}{n_1 + n_2 + \cdots + n_\ell} = \sum_k \mathbb{P}_k x_k$$

where $\mathbb{P}_k = \frac{n_k}{n_1 + \cdots + n_\ell}$. This is exactly the expectation in the discrete case! (Clearly, the underlying probability model is the set of families in the town together with the power set σ -algebra and discrete uniform measure. The random variable maps each family to the number of people in it.)

For some random variable ξ , we'll denote the expectation of ξ by $\mathbb{E}\xi$. Similarly, we'll denote its variance by $\text{Var}\xi$.

Example 33. We'll consider the two examples:

1. Suppose $\xi \sim \text{Ber}(p)$. Then:

$$\mathbb{E}\xi = p \cdot 1 + q \cdot 0 = p$$

Similarly,

$$\text{Var}\xi = p(1-p)^2 + q(0-p)^2 = pq^2 + qp^2 = pq(q+p) = pq$$

The square root of variance is called the *standard deviation* of ξ .

2. Suppose $\xi \sim \mathcal{N}(a, \sigma^2)$. Then, using the formula for continuous random variables:

$$\begin{aligned}\mathbb{E}\xi &= \int_{-\infty}^{+\infty} t \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-a)^2}{2\sigma^2}} dt \\ &= \int_{-\infty}^{+\infty} (\sigma y + a) \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/2} \sigma dy & y = (t-a)/\sigma \\ &= \int_{-\infty}^{+\infty} \sigma y \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy + \int_{-\infty}^{+\infty} a \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy\end{aligned}$$

Now observe that $\sigma y \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$ is an odd (integrable!) function integrated over a symmetric interval, and hence the integral is 0. So we obtain:

$$\begin{aligned}\mathbb{E}\xi &= \int_{-\infty}^{+\infty} a \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= \sqrt{2\pi} \cdot \frac{1}{\sqrt{2\pi}} a & \text{by HW 3--use Fubini!} \\ &= a\end{aligned}$$

So the first parameter of the normal distribution is its expectation. Let us now compute the variance:

$$\begin{aligned}\text{Var}\xi &= \int_{-\infty}^{+\infty} (t-a)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-a)^2}{2\sigma^2}} dt \\ &= \int_{-\infty}^{+\infty} \sigma^2 y^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-y^2/2} \sigma dy & y = \frac{t-a}{\sigma} \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} y^2 e^{-y^2/2} dy \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \left(-ye^{-y^2/2} \Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} e^{-y^2/2} dy \right) & \text{integration by parts} \\ &= \sigma^2 \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy & \text{L'Hopital Rule} \\ &= \sigma^2 & \text{by HW 3}\end{aligned}$$

So, the variance is σ^2 (the second parameter), and hence the standard deviation is σ .

2.3 Convolution of Probability Density Functions

Before we start talking about convolutions and other fun things, let us take a step back and recall the fundamental concepts we've learned so far.

Recall that if we have an arbitrary probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a **random variable** on this space is defined as a function $\xi : \Omega \rightarrow \mathbb{R}$ that is $(\mathcal{F}, \mathcal{B})$ -measurable. It is important to

remember that random variables record one specific property of each sample as a number. The **key problem** we're concerned with is the following: What is the probability that ξ takes a value between a and b , i.e. what is $\mathbb{P}\{\omega \in \Omega : a < \xi(\omega) < b\}$? Here are four things we need to keep in mind in order to answer this:

1. The question only makes sense whenever $\{\xi \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$, since we saw in the very beginning how important it is to make sure all the sets whose probability we'd like to measure are, in fact, measurable. Since \mathcal{B} is generated by the set of all closed left-rays on the real line, what we want is that $\{\xi \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}$.

To further convince you that the condition above is indeed necessary, consider the following counterexample that demonstrates how things go wrong when it isn't satisfied. Suppose that we have some set Ω and also that $\mathcal{F} \neq 2^\Omega$. Let $A \in \Omega$ such that $A \notin \mathcal{F}$. Then recall that the indicator function is defined as:

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

Then, we can't evaluate $\mathbb{P}\{\mathbb{1}_A \in (\frac{1}{2}, \frac{3}{2})\}$, because $\mathbb{1}_A^{-1}(\frac{1}{2}, \frac{3}{2}) = A \notin \mathcal{F}$! So we see that, since $\mathbb{P}(A)$ is not defined, this is a problem – our function is not Borel-measurable! Thus, **random variables are precisely the functions for which the key problem makes sense, for any a and b .**

2. The **law** of a random variable solves the key problem completely. Recall that the law of a random variable ξ on $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as $\mathbb{P}_\xi(B) = \mathbb{P}\{\xi \in B\} = \mathbb{P}\{\omega \in \Omega : \xi(\omega) \in B\}$. Once \mathbb{P}_ξ is known, the key problem will be answered completely without any information about Ω , \mathcal{F} , \mathbb{P} , or ξ , the original probability space or the random variable! Now we have a new probability measure $\mathbb{P}_\xi : \mathcal{B} \rightarrow [0, 1]$ on $(\mathbb{R}, \mathcal{B})$, which we know a great deal about. Once the law of ξ is known, the key problem is solved by the formula: $\mathbb{P}\{a < \xi < b\} = \mathbb{P}_\xi((a, b))$.
3. Moreover, the law determines $\mathbb{E}\xi$ and $\text{Var}\xi$ completely. These things reveal to us quantitative information about (like what is the average grade students got on a test) but not any qualitative information recorded in the original sample space (like who got what scores, etc.)
4. All the distributions we've seen so far (normal, binomial, etc.) are, by definition, probability measures on $(\mathbb{R}, \mathcal{B})$. These distributions have names because they are the laws of important random variables but their definitions do not involve random variables!

OK, now let's talk about convolutions. Remember that a distribution \mathbb{P} is continuous when it has a density. Recall, from the last section, that for any $B \in \mathcal{B}$, $\mathbb{P}(B) = \int_B f(t)dt$. So, the normal, Gamma, and uniform distributions are all probability measures on $(\mathbb{R}, \mathcal{B})$ defined

in this way. Now, recall how we used to add discrete random variables and get another random variable. Turns out, this kind of approach doesn't quite fly with continuous random variables – we can't simply add their pdfs. For example, if we have two pdfs, f_1 and f_2 , then $\int f_1 + f_2 = \int f_1 + \int f_2 = 1 + 1 = 2$. So, $f_1 + f_2$ is not a pdf. One option would be, of course, to scale each density by $\frac{1}{2}$, but it turns out that this is not the most useful thing to do. This is where we use the tool called convolution.

Definition 21. Let f_1 and f_2 be pdfs. The convolution of f_1 and f_2 is:

$$(f_1 * f_2)(t) = \int_{-\infty}^{+\infty} f_1(s)f_2(t-s)ds$$

Here are some properties of convolution that you will prove on the homework:

- $f_1 * f_2 = f_2 * f_1$
- $(f_1 * f_2) * f_3 = f_1 * (f_2 * f_3)$

Proposition 10. Let f_1 and f_2 be pdfs. Then:

1. $f_1 * f_2$ is a pdf;
2. If $f_1(t) = f_2(t) = 0$ for all $t \leq 0$, then:

$$f_1 * f_2(t) = \begin{cases} \int_0^t f_1(s)f_2(t-s)ds & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Proof. To prove (1), first note that $f_1 * f_2(t) \geq 0$ for all t , because the function we integrate is non-negative, by assumption. To prove continuity at all but finitely many points, you need to take Math 225. Finally, observe that

$$\begin{aligned} \int_{-\infty}^{+\infty} (f_1 * f_2)(t)dt &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_1(s)f_2(t-s)dsdt \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_1(s)f_2(t-s)dtds && \text{Fubini} \\ &= \int_{-\infty}^{+\infty} f_1(s) \left(\int_{-\infty}^{+\infty} f_2(t-s)dt \right) ds && y = t - s \\ &= \int_{-\infty}^{+\infty} f_1(s) \left(\int_{-\infty}^{+\infty} f_2(y)dy \right) ds \\ &= \int_{-\infty}^{+\infty} f_1(s)ds \\ &= 1 \end{aligned}$$

We'll note that in the continuous case, the pdf of the sum of two random variables is the convolution of their respective pdfs.

Now, to prove (2), first suppose $t \leq 0$. Then:

$$\begin{aligned} f_1 * f_2(t) &= \int_{-\infty}^{+\infty} f_1(s) f_2(t-s) ds \\ &= \int_0^{\infty} f_1(s) f_2(t-s) ds & f_1(s) = 0 \quad \forall t \leq 0 \\ &= 0 \end{aligned}$$

Where the last equality follows from observing that whenever $s > 0$ and $t \leq 0$, then $t-s < 0$, so $f_2(t-s) = 0$.

Next, assume $t > 0$. Then:

$$\begin{aligned} f_1 * f_2(t) &= \int_{-\infty}^{+\infty} f_1(s) f_2(t-s) ds \\ &= \int_0^{\infty} f_1(s) f_2(t-s) ds \\ &= \int_0^t f_1(s) f_2(t-s) ds \end{aligned}$$

Where the last equality follows from the fact that if $s > t$, then $t-s < 0$, so $f_2(t-s) = 0$ for all such values of s . \square

Theorem 12 (Important!).

$$(g_{\lambda, r_1} * g_{\lambda, r_2})(t) = g_{\lambda, r_1+r_2}(t)$$

for all $\lambda, r_1, r_2 > 0$ and $t > 0$.

Proof. First, recall that:

$$g_{\lambda, r}(t) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

where $\Gamma(r) = \int_0^{\infty} t^{r-1} e^{-t} dt$. Note that condition (2) of Proposition 10 is satisfied, so we can use it here. Observe that it is sufficient to show that

$$(g_{\lambda, r_1} * g_{\lambda, r_2}) = C g_{\lambda, r_1+r_2}$$

where C is a constant. This is the so-called "proportion trick." This is a sufficient condition for equality, because, since both sides are density functions, we know that taking integral on both sides we get that $1 = C \cdot 1$, so the constant C is forced to be 1. For simplicity of notation, let $g_{\lambda, r_1} = g_1$ and $g_{\lambda, r_2} = g_2$. By Prop. 10, we have that:

$$g_1 * g_2(t) = \begin{cases} \int_0^t g_1(s) g_2(t-s) ds & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Combining the results above, it then suffices to show that there exists some C such that for all $t > 0$, we have:

$$\int_0^t g_1(s)g_2(t-s)ds = Ct^{r_1+r_2-1}e^{-\lambda t}$$

Note that $\frac{1}{\Gamma(r)}\lambda^r$ is just another constant independent of t , so we don't have to worry about it. Note that WMA that $0 \leq s \leq t$. So we have the following:

$$\begin{aligned} \int_0^t g_1(s)g_2(t-s)ds &= C_1 \int_0^t s^{r_1-1}e^{-\lambda s}(t-s)^{r_2-1}e^{-\lambda(t-s)}ds \\ &= C_1 e^{-\lambda t} \int_0^t s^{r_1-1}(t-s)^{r_2-1}ds \\ &= C_1 e^{-\lambda t} \int_0^1 (t\theta)^{r_1-1}(t-t\theta)^{r_2-1}t d\theta \quad s = \theta t \text{ where } 0 \leq \theta \leq 1 \\ &= C_1 e^{-\lambda t} t^{r_1+r_2-1} \int_0^1 \theta^{r_1-1}(1-\theta)^{r_2-1}d\theta \\ &= C_2 e^{-\lambda t} t^{r_1+r_2-1} \end{aligned}$$

and this concludes the proof. □

2.4 Product Measure

Remember, when we introduced the notion of independence, we saw how convenient it was to compute probabilities when events were independent. Today, we'll talk about the notion called product measure, which is an extremely useful construction that will help us model successive independent experiments. In particular, product probability spaces are those where independence is defined in a natural way – we will take a collection of random variables and put them together to be independent via Cartesian product of their respective sample spaces and the construction of the appropriate σ -algebra. (It would be a good idea to read Feller Chapter V.4 to better familiarize yourself with the topic.)

Let $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ be probability spaces where $1 \leq i \leq n$. We would like to define a probability space on the set $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$. How do we define a σ -algebra on it? The first and natural guess would be to let $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 \times \cdots \times \mathcal{F}_n$. However, turns out, the Cartesian product of σ -algebras is not always a σ -algebra! As a counterexample consider $\Omega_1 = \Omega_2 = \{0, 1\}$ and $\mathcal{F}_1 = \mathcal{F}_2 = 2^{\{0,1\}}$. Let $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$. Then we see that $\{(0, 0)\} = \{0\} \times \{0\}$ and $\{(1, 1)\} = \{1\} \times \{1\}$. Thus $\{(0, 0)\} \in \mathcal{F}$ and $\{(1, 1)\} \in \mathcal{F}$. If \mathcal{F} were a σ -algebra, it must also be the case that $\{(0, 0), (1, 1)\} = \{(0, 0)\} \cup \{(1, 1)\} \in \mathcal{F}$, however $\{(0, 0), (1, 1)\} \notin \mathcal{F}_1 \times \mathcal{F}_2$. (Check this!) Thus, we need to fix this and find an appropriate σ -algebra on Ω . This is not too hard – we simply take the σ -algebra generated by the product of σ -algebras on each set:

$$\mathcal{F} = \sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n) = \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$$

Now, we need a probability measure on (Ω, \mathcal{F}) ...

Theorem 13. There exists a unique probability measure \mathbb{P} on $(\Omega, \mathcal{F}) = (\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n)$ such that $\mathbb{P}(A_1 \times \cdots \times A_n) = \mathbb{P}_1(A_1) \cdots \mathbb{P}_n(A_n)$ for all $(A_1, \dots, A_n) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

It is worth noting that once we fix the value of \mathbb{P} on the generating set of a σ -algebra, it uniquely determines the values that \mathbb{P} assigns to the rest of the sets in \mathcal{F} (this is a consequence of the π - λ theorem). However, it does not give us a tool to compute measures of an arbitrary set in \mathcal{F} – it only guarantees existence and uniqueness of the measure.

Proof. The proof is quite involved and is studied in Math 255. □

3 Random Vectors

3.1 Joint Distribution of Random Variables

In this section, we will generalize the notion of random variables to multidimensional random vectors using the concept of σ -algebras on Cartesian product that we talked about in the last section. Let us state a few definitions:

Definition 22.

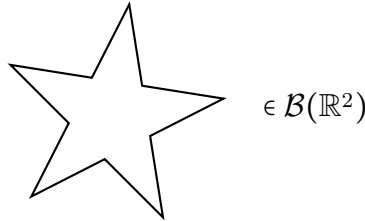
1. We generalize the notion of the Borel σ -algebra to $n \in \mathbb{N}$ dimensions as follows:

$$\mathcal{B}(\mathbb{R}^n) = \mathcal{B} \otimes \cdots \otimes \mathcal{B} = \sigma(\{A_1 \times \cdots \times A_n \subseteq \mathbb{R}^n : A_i \in \mathcal{B} \forall 1 \leq i \leq n\})$$

2. Any $\mathbb{P} \in \text{Prob}(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called an n -dimensional probability distribution.

Example 34. Consider a solid disk $\mathcal{D} = \{x^2 + y^2 < 1\} \in \mathcal{B}(\mathbb{R}^2)$. It is an exercise in analysis to show that \mathcal{D} is a union of countably many rectangles. Therefore \mathcal{D} is in the σ -algebra $\mathcal{B}(\mathbb{R}^2)$.

So keep in mind that, while something as cool as this:



is in the σ -algebra, it is not in the generating set.

Just like in the one-dimensional case, we have a one-to-one correspondence between $\text{Prob}(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ and $\text{Dist}(\mathbb{R}^n)$ where the bijection is defined as $\mathbb{P} \mapsto F(x_1, \dots, x_n) = \mathbb{P}((-\infty, x_1] \times \cdots \times (-\infty, x_n])$.

Example 35. Suppose we're working in \mathbb{R}^2 . Consider the following table:

Types	Discrete	Continuous
Definition	$\mathbb{P} = \sum_{k=1}^{\infty} p_k \delta_{(x_k, y_k)}$ $\sum_{k=1}^{\infty} p_k = 1; p_k \geq 0$	<ul style="list-style-type: none"> • $\mathbb{P} \sim f(t, s)$ • $f(t, s) \geq 0$ • $\iint_{\mathbb{R}^2} f(t, s) dt ds = 1$ • f is continuous at all but finitely many curves
Value of $\mathbb{P}(B)$, $B \in \mathcal{B}$	$= \sum_{k=1}^{\infty} p_k \delta_{(x_k, y_k)}(B)$ $= \sum_{(x_k, y_k) \in B} p_k$	$= \iint_B f(t, s) dt ds$

Recall that we said in one of the lectures that we have several kinds of probability distributions: discrete, continuous, singular and those of the mixed type. This is still true in the multidimensional case, and in fact we'll be able to provide an example of a singular probability this time.

Let us first recall a few things from calculus. Note that a double integral over any curve in \mathbb{R}^2 is simply 0, because, loosely speaking, it doesn't have any thickness. That is, for any curve C , we have $\iint_C f(t, s) dt ds = 0$. (It is important not to confuse this with the line integral of functions, $\int_C f(x, y) ds = \int_a^b f(x(t), y(t)) \sqrt{(x'(t))^2 + (y'(t))^2} dt$ where $\{x = x(t), y = y(t) : a \leq t \leq b\}$ is a parametrization of the curve C). This is why it is OK for $f(t, s)$ to be discontinuous at finitely many curves in the continuous case.

A probability measure on the unit circle is an example of a singular probability measure. Certainly, it cannot be discrete, because the set of points is uncountable. But also, it is not continuous, by our remark above, as the double integral over the unit circle is 0.

We'll also note that any n -dimensional distribution can be written as $\lambda_1 \mathbb{P}_1 + \dots + \lambda_n \mathbb{P}_n$ where $\lambda_1 + \dots + \lambda_n = 1$.

Definition 23. An n -dimensional *random vector* on $(\Omega, \mathcal{F}, \mathbb{P})$ is an ordered collection of random variables $(\xi_1, \dots, \xi_n) = \vec{\xi}$.

Example 36. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an infinite coin toss model, $0 < p < 1$. Recall, we defined a Bernoulli random variable to be $X_i(\omega) = a_i$ where $\omega = (a_i)_{i=1}^{\infty}$. Then $(X_1, X_2), (X_1, X_1), (X_3, S_4)$ are all examples of random vectors on $(\Omega, \mathcal{F}, \mathbb{P})$.

Like in the one-dimensional case, we're concerned with the **key problem**: What is the probability that $(\xi_1, \dots, \xi_n) \in B$ where $B \in \mathcal{B}(\mathbb{R}^n)$? That is, we'd like to be able to compute the value of $\mathbb{P}\{\omega \in \Omega : (\xi_1(\omega), \dots, \xi_n(\omega)) \in B\}$. Let us solve the problem step-by-step:

1. When does the problem make sense?

Proposition 11. $\vec{\xi}$, as a map $\Omega \rightarrow \mathbb{R}^n$, is $(\mathcal{F}, \mathcal{B}(\mathbb{R}^n))$ -measurable.

Proof. We would like to show that for any $B \in \mathcal{B}(\mathbb{R}^n)$, we have $\{\omega \in \Omega : \vec{\xi}(\omega) \in B\} = \{\vec{\xi} \in B\} \in \mathcal{F}$. We've proven that it only suffices to show measurability for the members of the generating set. Let $A_1 \times \cdots \times A_n \in \mathcal{B} \times \cdots \times \mathcal{B}$. Note that:

$$\begin{aligned} \{\vec{\xi} \in A_1 \times \cdots \times A_n\} &= \{\omega \in \Omega : (\xi_1(\omega), \dots, \xi_n(\omega)) \in A_1 \times \cdots \times A_n\} \\ &= \{\omega \in \Omega : \xi_i(\omega) \in A_i, 1 \leq i \leq n\} \\ &= \bigcap_{i=1}^n \{\xi_i \in A_i\} \in \mathcal{F} \end{aligned}$$

where the last equality follows from the fact that each ξ_i is a random variable, so $\{\xi_i \in A_i\} \in \mathcal{F}$, and that \mathcal{F} is closed under countable intersections. \square

So, random vectors are those functions for which the problem makes sense.

Definition 24. The push-forward measure $\mathbb{P}_{\vec{\xi}} = \mathbb{P}_{(\xi_1, \dots, \xi_n)} \in \text{Prob}(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ is called the *joint distribution/law* of the random vector $\vec{\xi}$.

Example 37. Let, again, $(\Omega, \mathcal{F}, \mathbb{P})$ be the infinite coin toss model where $p = \frac{1}{2}$. We'd like to ask: what is the law of (X_1, X_2) ? Well, any combination in $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ is equally likely, so:

$$\mathbb{P}_{(X_1, X_2)} = \frac{1}{4}\delta_{(0,0)} + \frac{1}{4}\delta_{(0,1)} + \frac{1}{4}\delta_{(1,0)} + \frac{1}{4}\delta_{(1,1)}$$

Now, what is the law of (X_1, X_1) ? This time, there are only two possibilities: $(0, 0)$ and $(1, 1)$, so:

$$\mathbb{P}_{(X_1, X_1)} = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)}$$

2. Note that joint distribution completely solves the key problem, as for any $B \in \mathcal{B}(\mathbb{R}^n)$:

$$\mathbb{P}_{(\xi_1, \dots, \xi_n)}(B) = \mathbb{P}\{\omega \in \Omega : (\xi_1(\omega), \dots, \xi_n(\omega)) \in B\}$$

So, when the joint distribution is known, we will be able to answer the main question without having any information about the original probability space or the random variables in the random vector.

In this class, we will be concerned with the following:

- For given random variables ξ_1, \dots, ξ_n , determine their joint distribution.
- Suppose we are given the joint distribution. Determine the properties of random variables and the probability of “interesting” events.

Example 38. Let $B \in \mathcal{B}(\mathbb{R}^2)$ and ξ, η be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$.

1. Suppose $B = A \times \mathbb{R}$. Then:

$$\begin{aligned}\mathbb{P}_{(\xi, \eta)}(B) &= \mathbb{P}\{\omega \in \Omega : (\xi(\omega), \eta(\omega)) \in A \times \mathbb{R}\} \\ &= \mathbb{P}\{\omega \in \Omega : \xi(\omega) \in A \text{ and } \eta(\omega) \in \mathbb{R}\} \\ &= \mathbb{P}\{\xi \in A\}\end{aligned}$$

Note that from this calculation, we can conclude that the distribution of each random variable in the vector is determined by the joint distribution.

2. Suppose $B = \{(x, y) : x + y \leq z\}$. Then:

$$\begin{aligned}\mathbb{P}_{(\xi, \eta)}\{(x, y) : x + y \leq z\} &= \mathbb{P}\{\omega \in \Omega : \xi(\omega) + \eta(\omega) \leq z\} \\ &= \mathbb{P}\{\xi + \eta \leq z\}\end{aligned}$$

This means that knowing the joint distribution, we can compute the cdf of the sum of random variables, and hence uniquely determine the law.

3. Doing the same for $B = \{(x, y) : xy \leq z\}$ will yield that we can also compute the cdf of the product of random variables, and hence uniquely determine the law.

Now, a natural question arises: Is the converse true? That is, do \mathbb{P}_ξ and \mathbb{P}_η completely determine the joint distribution, $\mathbb{P}_{(\xi, \eta)}$? Turns out, no. In fact, you saw the counterexample in 37, where we had $\mathbb{P}_{X_1} = \mathbb{P}_{X_2} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, but $\mathbb{P}_{(X_1, X_2)} \neq \mathbb{P}_{(X_1, X_1)}$.

Problem 1. Let $\xi \sim \text{Bin}(n, p)$ and $\eta \sim \text{Geo}(p)$ where ξ and η are random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. What is the joint distribution of (ξ, η) ?

Proof. This problem does *NOT* make sense – joint distribution cannot be determined from individual laws of random variables! \square

3.2 Discrete Random Vectors

Definition 25. A random vector $\vec{\xi} = (\xi_1, \dots, \xi_n)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called *discrete* if there exists a countable set $S \subseteq \mathbb{R}^n$ such that $(\xi_1(\omega), \dots, \xi_n(\omega)) \in S$ for all $\omega \in \Omega$.

Lemma 4. $\vec{\xi} = (\xi_1, \dots, \xi_n)$ is discrete if and only if ξ_i is discrete for all $1 \leq i \leq n$.

Proof. We need to prove the two directions:

- (\implies) Suppose $\vec{\xi} = (\xi_1, \dots, \xi_n)$ is discrete. Then, by definition, there exists a countable set $S \subseteq \mathbb{R}^n$ such that $\vec{\xi}$ takes values only in S . If ξ_i took uncountably many values for some $1 \leq i \leq n$, then S would be uncountable as well, since a vector can only take as many or more values than individual random variables. Hence each ξ_i takes countably many values and is a discrete random variable.

(\Leftarrow) Suppose each ξ_i is a random variable taking values in a countable set $A_i \subseteq \mathbb{R}$. Then let $S = A_1 \times \cdots \times A_n$. Finite Cartesian product of countable sets is countable, and so S is countable. Now observe that, by the way we defined S , we have $\vec{\xi}(\omega) \in S$ for all $\omega \in \Omega$, so we're done. □

Theorem 14. If (ξ_1, \dots, ξ_n) is a discrete random vector, then:

$$\mathbb{P}_{(\xi_1, \dots, \xi_n)} = \sum_{(x_1, \dots, x_n) \in S} \mathbb{P}\{\xi_1 = x_1, \dots, \xi_n = x_n\} \delta_{(x_1, \dots, x_n)}$$

Proof. Similar to that of Theorem 10. □

Example 39. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the infinite coin toss model with $p = \frac{1}{2}$. Recall that in the last section, we said that:

$$\mathbb{P}_{(X_1, X_2)} = \frac{1}{4} \delta_{(0,0)} + \frac{1}{4} \delta_{(0,1)} + \frac{1}{4} \delta_{(1,0)} + \frac{1}{4} \delta_{(1,1)}$$

Note that this exactly coincides with the theorem above, since $\mathbb{P}\{X_1 = a_1, X_2 = a_2\} = \frac{1}{4}$, by the definition of the model.

Example 40. Suppose we have 3 balls numbered ①, ②, ③ and we put them into three cells numbered $\boxed{1}$, $\boxed{2}$, $\boxed{3}$. Let us describe the probability space we're working in. Clearly, we only have finitely many possibilities, so we can let $\mathcal{F} = 2^\Omega$ and $\mathbb{P} = \sum_{\omega \in \Omega} \frac{1}{|\Omega|} \delta_\omega$. What should Ω be? We will let Ω consist of triples, where each number in the triple will denote the number of the cell each respective ball is placed into. That is, $\Omega = \{(a_1, a_2, a_3) : a_i \in \{1, 2, 3\}\}$ where $\boxed{a_i}$ is the cell of ball ① for $i = 1, 2, 3$. Now, let ξ be the random variable that denoted the number of occupied cells. Let η be the random variable denoting the number of balls in cell $\boxed{1}$. We will consider the random vector (ξ, η) . So, for instance, when $\omega = (2, 2, 1)$, we have that $(\xi(\omega), \eta(\omega)) = (2, 1)$ and when $\omega = (3, 2, 1)$, we have $(\xi(\omega), \eta(\omega)) = (3, 1)$. Whenever we want to describe a joint distribution of a two-dimensional random vector, each component of which takes finitely many values, it is always convenient to make a table that describes each possibility:

$\xi \backslash \eta$	0	1	2	3
1	$\frac{2}{27}$	0	0	$\frac{1}{27}$
2	$\frac{6}{27}$	$\frac{6}{27}$	$\frac{6}{27}$	0
3	0	$\frac{6}{27}$	0	0

While describing each number is perhaps unnecessary, we will explain where the entry (2,2) comes from. This vector represents an event that there are 2 occupied boxes and 2 balls in

the first box. Thus, there are $\binom{3}{2} = 3$ ways to choose the two balls to place in cell $\boxed{1}$ and then there are two choices for where to put the remaining ball. Thus, there are $3 \cdot 2 = 6$ desirable outcomes out of total $3^3 = 27$. So the probability of this event is $\frac{6}{27}$.

Now, suppose we want to figure out the value of $\mathbb{P}\{\xi + \eta \leq 2.5\}$. Recall that this is the same as $\mathbb{P}_{(\xi, \eta)}(B)$ where $B = \{(x, y) : x + y \leq 2.5\}$. In other words, it's the sum of the probabilities of those events when the sum of the values ξ and η take is at most 2.5. Those events are marked blue in the table above. Hence, we conclude that $\mathbb{P}\{\xi + \eta \leq 2.5\} = \frac{8}{27}$.

Now, can we figure out the law of ξ alone? Recall that by Theorem 10, \mathbb{P}_ξ is defined as $\mathbb{P}_\xi = \mathbb{P}\{\xi = 1\}\delta_1 + \mathbb{P}\{\xi = 2\}\delta_2 + \mathbb{P}\{\xi = 3\}\delta_3$. Now, recall that $\mathbb{P}\{\xi = 1\} = \mathbb{P}\{\xi = 1, \eta \in \mathbb{R}\} = \mathbb{P}_{(\xi, \eta)}(B)$ where $B = \{1\} \times \mathbb{R}$. Thus, according to the table:

- $\mathbb{P}\{\xi = 1\} = \frac{3}{27}$
- $\mathbb{P}\{\xi = 2\} = \frac{18}{27}$
- $\mathbb{P}\{\xi = 3\} = \frac{6}{27}$

where the numbers are coming from summing across all possible values of η for each particular ξ . Thus,

$$\mathbb{P}_\xi = \frac{3}{27}\delta_1 + \frac{18}{27}\delta_2 + \frac{6}{27}\delta_3$$

which is referred to as *marginal* distribution.

Example 41. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an infinite coin toss model with $0 < p < 1$. For each $r \in \mathbb{N}^+$, define η_r recursively as $\eta_1 = W_1$ and $\eta_r = W_r - W_{r-1}$ where:

$$W_r(\omega) = \begin{cases} 0 & \text{if } \sum_{i=1}^{\infty} a_i < r \\ n & \text{if } \sum_{i=1}^{n-1} a_i = r - 1, a_n = 1 \end{cases}$$

Let us consider a concrete example to understand what η represents. Let

$$\omega = (0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, \dots)$$

Then:

- $\eta_1(\omega) = 3$
- $\eta_2(\omega) = 4 - 3 = 1$
- $\eta_3(\omega) = 8 - 4 = 4$

and so on. We see that η_r describes the number of experiments between the $r - 1$ st success (not included) and the r th success. Now, let us determine the joint distribution $\mathbb{P}_{(\eta_1, \dots, \eta_n)}$

for some $n \in \mathbb{N}^+$. Note that η_i takes countably many values for all i , so the random vector is discrete. Moreover, recall that

$$\mathbb{P}\{(a_i)_{i=1}^\infty \in \Omega : \sum_{i=1}^\infty a_i < \infty\} = 0$$

So, in fact, the probability of getting finitely many 1's is zero, so we only need to worry about the values η_r takes that are positive integers, since it is always true whenever we have infinitely 1's. By Theorem 14, we need to compute the value of $\mathbb{P}\{\eta_1 = k_1, \dots, \eta_n = k_n\}$ for $k_1, \dots, k_n \in \mathbb{Z}^+$. Let us consider the special case when $n = 2$ and compute $\mathbb{P}\{\eta_1 = 3, \eta_2 = 1\}$. Note, by definition of η_i , we have:

$$\mathbb{P}\{\eta_1 = 3, \eta_2 = 1\} = \mathbb{P}\{X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 1\} = p^2 q^2$$

Now, we can see how to generalize this notion, by translating the problem into computing probabilities of the events we are well-familiar with:

$$\begin{aligned} \mathbb{P}(\eta_1 = k_1, \dots, \eta_n = k_n) &= \begin{cases} (pq^{k_1-1})(pq^{k_2-1}) \dots (pq^{k_n-1}) & \text{if } k_i \in \mathbb{Z}^+ \forall i \\ 0 & \text{if } \exists k_i \notin \mathbb{Z}^+ \end{cases} \\ &= \begin{cases} p^n q^{k_1 + \dots + k_n - n} & \text{if } k_i \in \mathbb{Z}^+ \forall i \\ 0 & \text{if } \exists k_i \notin \mathbb{Z}^+ \end{cases} \end{aligned}$$

Finally, we're ready to compute the law:

$$\mathbb{P}_{(\eta_1, \dots, \eta_n)} = \sum_{k_1, \dots, k_n \in \mathbb{Z}^+} p^n q^{k_1 + \dots + k_n - n} \delta_{(k_1, \dots, k_n)}$$

This distribution is called the *multidimensional geometric distribution*.

Note that the following result is a corollary of the above computation:

$$\sum_{k_1, \dots, k_n \in \mathbb{Z}^+} p^n q^{k_1 + \dots + k_n - n} = 1$$

Now, let $n \in \mathbb{N}^+$ be given. We want to determine the law of η_n . Thankfully, we have the ~~nuclear bomb~~ joint distribution, so we know how to ~~destroy~~ compute it! Since $\mathbb{P}_{\eta_n} = \sum_{k \in \mathbb{Z}^+} \mathbb{P}\{\eta_n = k\} \delta_k$, we need to determine $\mathbb{P}\{\eta_n = k\}$ for each $k \in \mathbb{Z}^+$. First, note that, in multiple dimensions, we define the delta measure in the same way:

$$\delta_{(k_1, \dots, k_n)}(A_1 \times \dots \times A_n) = \begin{cases} 1 & \text{if } (k_1, \dots, k_n) \in A_1 \times \dots \times A_n \\ 0 & \text{otherwise} \end{cases}$$

Then, we proceed as follows:

$$\begin{aligned}
\mathbb{P}\{\eta_n = k\} &= \mathbb{P}_{(\eta_1, \dots, \eta_n)}(\mathbb{R}^{n-1} \times \{k\}) \\
&= \sum_{k_1, \dots, k_n \in \mathbb{Z}^+} p^n q^{k_1 + \dots + k_n - n} \delta_{(k_1, \dots, k_n)}(\mathbb{R}^{n-1} \times \{k\}) \\
&= \sum_{k_1, \dots, k_{n-1} \in \mathbb{Z}^+} p^n q^{k_1 + \dots + k_{n-1} + k - n} \\
&= pq^{k-1} \sum_{k_1, \dots, k_{n-1} \in \mathbb{Z}^+} p^{n-1} q^{k_1 + \dots + k_{n-1} - (n-1)} \\
&= pq^{k-1} \cdot 1 && \text{corollary above} \\
&= pq^{k-1}
\end{aligned}$$

Looks familiar, doesn't it? Finally, we conclude that:

$$\mathbb{P}_{\eta_n} = \sum_{k \in \mathbb{Z}^+} pq^{k-1} \delta_k = \text{Geo}(p)$$

Definition 26. Discrete random variables ξ_1, \dots, ξ_n , taking values in a countable set S , are *independent* if:

$$\mathbb{P}_{(\xi_1, \dots, \xi_n)} = \sum_{x_1, \dots, x_n \in S} \left(\prod_{i=1}^n \mathbb{P}\{\xi_i = x_i\} \right) \delta_{(x_1, \dots, x_n)}$$

Note that, from the definition above, it follows that discrete random variables are independent when

$$\mathbb{P}\left(\bigcap_{i=1}^n \{\xi_i = x_i\}\right) = \mathbb{P}\{\xi_1 = x_1, \dots, \xi_n = x_n\} = \mathbb{P}\{\xi_1 = x_1\} \dots \mathbb{P}\{\xi_n = x_n\}$$

This coincides with the definition we had earlier. So random variables are independent iff when their joint probability equals the product of their individual probabilities.

Moreover, we can now conclude that η_1, \dots, η_n are independent for every $n \in \mathbb{N}^+$.

3.3 Independence of Random Variables

Let us pick up where we left off in the last section. Suppose ξ_1, \dots, ξ_n be discrete random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and assume $\xi_i(\omega) \in S$ for all $i = 1, \dots, n$ and $\omega \in \Omega$ where $S \subseteq \mathbb{R}^n$ is countable. Recall we stated the following result:

Theorem 15. If (ξ_1, \dots, ξ_n) is a discrete random vector, then:

$$\mathbb{P}_{(\xi_1, \dots, \xi_n)} = \sum_{(x_1, \dots, x_n) \in S} \mathbb{P}\{\xi_1 = x_1, \dots, \xi_n = x_n\} \delta_{(x_1, \dots, x_n)}$$

We also defined independence of random variables as follows:

Definition 27. Discrete random variables ξ_1, \dots, ξ_n , taking values in a countable set S , are *independent* if:

$$\mathbb{P}_{(\xi_1, \dots, \xi_n)} = \sum_{x_1, \dots, x_n \in S} \left(\prod_{i=1}^n \mathbb{P}\{\xi_i = x_i\} \right) \delta_{(x_1, \dots, x_n)}$$

That is,

$$\mathbb{P}\{\xi_1 = x_1, \dots, \xi_n = x_n\} = \prod_{i=1}^n \mathbb{P}\{\xi_i = x_i\}$$

for all $x_1, \dots, x_n \in S$. Now, we will finally shed the light on the relationship between the notion of independence of random variables and that of events.

Theorem 16. Let $A, B \in \mathcal{F}$. Recalled we defined the indicator of an arbitrary set $F \in \mathcal{F}$ function $\mathbb{1}_F : \mathbb{N} \rightarrow \{0, 1\}$ as follows:

$$\mathbb{1}_F(i) = \begin{cases} 0 & \text{if } i \notin F \\ 1 & \text{if } i \in F \end{cases}$$

Then A and B are independent if and only if $\mathbb{1}_A$ and $\mathbb{1}_B$ are independent random variables.

Proof. Before proving this important result, note that the indicator functions are indeed random variables, since A and B are in the σ -algebra \mathcal{F} , and hence they are clearly measurable.

\implies Suppose A and B are independent. Then $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Note that we need to show that the four things hold:

- (a) $\mathbb{P}(\mathbb{1}_A = 1, \mathbb{1}_B = 1) = \mathbb{P}(\mathbb{1}_A = 1)\mathbb{P}(\mathbb{1}_B = 1)$
- (b) $\mathbb{P}(\mathbb{1}_A = 1, \mathbb{1}_B = 0) = \mathbb{P}(\mathbb{1}_A = 1)\mathbb{P}(\mathbb{1}_B = 0)$
- (c) $\mathbb{P}(\mathbb{1}_A = 0, \mathbb{1}_B = 1) = \mathbb{P}(\mathbb{1}_A = 0)\mathbb{P}(\mathbb{1}_B = 1)$
- (d) $\mathbb{P}(\mathbb{1}_A = 0, \mathbb{1}_B = 0) = \mathbb{P}(\mathbb{1}_A = 0)\mathbb{P}(\mathbb{1}_B = 0)$

We will prove (a) and (b), since the rest are symmetric. To see why (a) is true, note:

$$\begin{aligned} A \cap B &= \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\} \\ &= \{\omega \in \Omega : \mathbb{1}_A(\omega) = 1 \text{ and } \mathbb{1}_B(\omega) = 1\} \\ &= \{\mathbb{1}_A = 1, \mathbb{1}_B = 1\} \end{aligned}$$

Similarly, we know that $A = \{\omega \in \Omega : \mathbb{1}_A(\omega) = 1\}$ and $B = \{\omega \in \Omega : \mathbb{1}_B(\omega) = 1\}$ so we conclude that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \implies \mathbb{P}(\mathbb{1}_A = 1, \mathbb{1}_B = 1) = \mathbb{P}(\mathbb{1}_A = 1)\mathbb{P}(\mathbb{1}_B = 1)$$

To prove (b), we will also translate sets into the values indicator functions take:

$$\begin{aligned}
\mathbb{P}(\mathbb{1}_A = 1, \mathbb{1}_B = 0) &= \mathbb{P}\{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\} \\
&= \mathbb{P}\{\omega \in \Omega : \omega \in A - (A \cap B)\} \\
&= \mathbb{P}(A) - \mathbb{P}(A \cap B) && A \cap B \subseteq A \\
&= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) && \text{independence} \\
&= \mathbb{P}(A)(1 - \mathbb{P}(B)) \\
&= \mathbb{P}(A)\mathbb{P}(B^c) \\
&= \mathbb{P}(\mathbb{1}_A = 1)\mathbb{P}(\mathbb{1}_B = 0)
\end{aligned}$$

\Leftarrow Now suppose the indicator functions are independent. In particular, $\mathbb{P}(\mathbb{1}_A = 1, \mathbb{1}_B = 1) = \mathbb{P}(\mathbb{1}_A = 1)\mathbb{P}(\mathbb{1}_B = 1)$. Then we immediately get the desired result, by the work we've done in the forward direction.

□

Definition 28. We say that $A_1, \dots, A_n \in \mathcal{F}$ are independent if $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_n}$ are independent random variables.

WARNING:

1. $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) \not\Rightarrow A, B, C \text{ are independent.}$
2. Pairwise independence of $A, B, C \not\Rightarrow A, B, C \text{ are independent.}$

Example 42. Let η_1, \dots, η_n be independent random variables all taking values in \mathbb{Z}^+ . Suppose $\eta_i \sim \text{Geo}(p)$ for $i = 1, \dots, n$ and $p \in (0, 1)$. Then:

$$\eta_1 + \dots + \eta_n \sim \text{NB}(n, p)$$

Proof. It's important to note that previously, we only worked with the concrete coin toss model. Now we are proving something more general, independent of the underlying probability space.

Recall that we've already shown the following three things:

- We computed the joint distribution, $\mathbb{P}_{(\eta_1, \dots, \eta_n)}$ for all $n \in \mathbb{N}^+$ in the previous section;
- $\eta_i \sim \text{Geo}(p)$ for all $i = 1, \dots, n$;
- The above two points imply that η_1, \dots, η_n are independent.

Also, recall that $\text{NB}(n, p) = \sum_{n=k}^{\infty} \binom{k-1}{n-1} p^n q^{k-n} \delta_n$. Now, we begin with the proof. Note:

$$\begin{aligned}
\mathbb{P}\{\eta_1 + \dots + \eta_n = k\} &= \sum_{\substack{k_1 + \dots + k_n = k \\ k_i > 0}} \mathbb{P}\{\eta_1 = k_1, \dots, \eta_n = k_n\} && \text{general} \\
&= \sum_{\substack{k_1 + \dots + k_n = k \\ k_i > 0}} \mathbb{P}\{\eta_1 = k_1\} \dots \mathbb{P}\{\eta_n = k_n\} && \text{independence} \\
&= \sum_{\substack{k_1 + \dots + k_n = k \\ k_i > 0}} p q^{k_1-1} \dots p q^{k_n-1} && \eta_i \sim \text{Geo}(p) \\
&= \sum_{\substack{k_1 + \dots + k_n = k \\ k_i > 0}} p^n q^{k-n} \\
&= p^n q^{k-n} |\{(k_1, \dots, k_n) : k_1 + \dots + k_n = k, k_i > 0\}| \\
&= p^n q^{k-n} \binom{k-1}{k-n} \\
&= \binom{k-1}{n-1} p^n q^{k-n}
\end{aligned}$$

□

and this concludes the proof.

Example 43. Let ξ, η, θ be independent random variables and suppose $\xi, \eta, \theta \sim \text{Geo}(p)$ where $p \in (0, 1)$. Then:

1. $\mathbb{P}\{\xi \leq \eta\} = \frac{1}{1+q}$;
2. $\xi + \eta$ and θ are independent;
3. $\mathbb{P}\{\xi + \eta \leq \theta + 1\} = \frac{1}{(1+q)^2}$ ← Extra Credit #3

Proof. To prove (1), first observe the following:

$$\begin{aligned}
\{\xi \leq \eta\} &= \{\omega \in \Omega : \xi(\omega) \leq \eta(\omega)\} \\
&= \uplus_{k=1}^{\infty} \uplus_{i=1}^k \{\eta = k, \xi = i\}
\end{aligned}$$

Now, we compute:

$$\begin{aligned}
\mathbb{P}\{\xi \leq \eta\} &= \sum_{k=1}^{\infty} \sum_{i=1}^k \mathbb{P}\{\eta = k, \xi = i\} \\
&= \sum_{k=1}^{\infty} \sum_{i=1}^k \mathbb{P}\{\eta = k\} \mathbb{P}\{\xi = i\} && \text{independence} \\
&= \sum_{k=1}^{\infty} \sum_{i=1}^k pq^{k-1} pq^{i-1} \\
&= \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} pq^{k-1} pq^{i-1} && \text{change limits: HW 1} \\
&= \sum_{i=1}^{\infty} p^2 q^{i-2} \sum_{k=i}^{\infty} q^k \\
&= \sum_{i=1}^{\infty} p^2 q^{i-2} \left(\frac{q^i}{1-q} \right) && \text{geometric series} \\
&= \sum_{i=1}^{\infty} pq^{2i-2} \\
&= \frac{p}{1-q^2} && \text{geometric series} \\
&= \frac{p}{(1-q)(1+q)} \\
&= \frac{1}{1-q}
\end{aligned}$$

To prove (2), we do the following:

$$\begin{aligned}
\mathbb{P}\{\xi + \eta = k, \theta = \ell\} &= \mathbb{P}\left(\bigcup_{i=1}^{k-1} \{\xi = i, \eta = k-i, \theta = \ell\}\right) \\
&= \sum_{i=1}^{k-1} \mathbb{P}\{\xi = i, \eta = k-i, \theta = \ell\} \\
&= \sum_{i=1}^{k-1} \mathbb{P}\{\xi = i\} \mathbb{P}\{\eta = k-i\} \mathbb{P}\{\theta = \ell\} && \text{independent} \\
&= \mathbb{P}\{\theta = \ell\} \sum_{i=1}^{k-1} \mathbb{P}\{\xi = i\} \mathbb{P}\{\eta = k-i\} \\
&= \mathbb{P}\{\theta = \ell\} \mathbb{P}\{\xi + \eta = k\}
\end{aligned}$$

Note that the last equality comes from the fact that mutual independence of random variables implies pairwise independence, so ξ and η are independent, since ξ, η, θ are independent by

assumption. To prove this fact, simply observe that:

$$\begin{aligned}\mathbb{P}\{\xi = x_1, \eta = x_2\} &= \mathbb{P}\{\xi = x_1, \eta = x_2, \theta \in \mathbb{R}\} \\ &= \mathbb{P}\{\xi = x_1\}\mathbb{P}\{\eta = x_2\}\mathbb{P}\{\theta \in \mathbb{R}\} \\ &= \mathbb{P}\{\xi = x_1\}\mathbb{P}\{\eta = x_2\}\end{aligned}$$

It is also worth noting that the proof above is a general fact about independent random variables – we never used the fact they are geometrically distributed.

Proof of (3) is left as an extra-credit exercise. Hint: Show $\xi + \eta \sim \text{NB}$ and use the fact in (2). Identify the right Borel set and proceed similarly to (1). \square

3.4 Jointly Continuous Random Variables

Before we begin talking about probability, let us review some multivariable calculus.

Example 44. 15.0 Define a function φ as follows:

$$\varphi(s, t) = \begin{cases} 8st & 0 \leq s \leq t, 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Compute:

1. $\int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} \varphi(s, t) ds) dt$
2. $\int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} \varphi(s, t) dt) ds$

Proof. To evaluate these integrals, we first evaluate the inner part, and then the outer part. We get:

$$\begin{aligned}\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(s, t) ds dt &= 8 \int_0^1 \int_0^t st ds dt \\ &= 8 \int_0^1 \left[\frac{s^2 t}{2} \right]_0^t dt \\ &= 8 \cdot \frac{t^4}{8} \Big|_0^1 \\ &= 1\end{aligned}$$

Changing the order of integration, we obtain:

$$\begin{aligned}
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(s, t) ds dt &= 8 \int_0^1 \int_s^1 st \, dt ds \\
&= 8 \int_0^1 \left[\frac{st^2}{2} \right]_s^1 ds \\
&= 8 \int_0^1 \left[\frac{s}{2} - \frac{s^3}{2} \right] ds \\
&= \int_0^1 4s - 4s^3 ds \\
&= 2 - 1 \\
&= 1
\end{aligned}$$

□

Definition 29. We say random variables ξ_1, \dots, ξ_n on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are jointly continuous if $\mathbb{P}_{(\xi_1, \dots, \xi_n)}$ is continuous, i.e. there exists an n -dimensional pdf $\varphi(t_1, \dots, t_n)$ such that

$$\mathbb{P}_{(\xi_1, \dots, \xi_n)}(B) = \int \int \dots \int_B \varphi(t_1, \dots, t_n) dt_1 \dots dt_n$$

for all $B \in \mathcal{B}(\mathbb{R}^n)$.

Example 45. Consider uniform distribution over the solid disk $x^2 + y^2 \leq 1$. Its pdf is defined by:

$$\varphi(s, t) = \begin{cases} \frac{1}{\pi} & s^2 + t^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

It is a good idea to check it is indeed a pdf!

Example 46. Consider particles randomly distributed in \mathbb{R}^3 , i.e. x, y, z coordinates are random variables. There is a function $\psi : \mathbb{R}^3 \rightarrow \mathbb{C}$ called the wave function such that the following property holds:

$$\int_{\mathbb{R}^3} |\psi|^2 = 1$$

The probability that a particle is in a set $B \in \mathcal{B}(\mathbb{R}^3)$ is:

$$\int_B |\psi|^2$$

So we see that $|\psi|^2$ is a joint distribution of the three random variables.

Proposition 12. The following two things hold:

1. If ξ_1, \dots, ξ_n are jointly continuous, then ξ_i is continuous for all $1 \leq i \leq n$.
2. The converse of (1) is *NOT* true.

Proof. To prove (1), let $n = 2$ for simplicity (the result can simply be extended to any n by induction). Suppose $(\xi, \eta) \sim \varphi(s, t)$. Then:

$$\begin{aligned}\mathbb{P}\{\xi \leq x\} &= \mathbb{P}\{\omega \in \Omega : \xi(\omega) \leq x, \eta(\omega) \in \mathbb{R}\} \\ &= \mathbb{P}_{(\xi, \eta)}((-\infty, x] \times \mathbb{R}) \\ &= \int_{-\infty}^x \left(\int_{-\infty}^{\infty} \varphi(s, t) dt \right) ds\end{aligned}$$

Therefore, $\xi \sim \int_{-\infty}^{\infty} \varphi(s, t) dt$, so it has a pdf and thus is continuous.

To give a counterexample for (2), let $\xi \sim U(0, 1)$ and suppose $\xi = \eta$. Let C be a curve defined by $C = \{(x, x) : x \in (0, 1)\}$. Then:

$$\begin{aligned}\mathbb{P}_{(\xi, \eta)}(C) &= \{\omega \in \Omega : (\xi(\omega), \eta(\omega)) \in C\} \\ &= \{\omega \in \Omega : 0 < \xi(\omega) < 1\} \\ &= 1\end{aligned}$$

But now note that no function in two variables can have a positive double integral over a curve. So $\int_C \varphi(s, t) ds dt = 0$ for any $\varphi(s, t)$. So the joint distribution cannot have a pdf. \square

Example 47. Suppose $(\xi, \eta) \sim \varphi(s, t)$ where

$$\varphi(s, t) = \begin{cases} 4st & 0 \leq s \leq 1, 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We'd like to determine pdfs of ξ and η , $f(s)$ and $g(t)$, respectively.

$$\begin{aligned}f(s) &= \int_{-\infty}^{\infty} \varphi(s, t) dt \\ &= \begin{cases} \int_0^1 4st dt & s \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \int_0^1 4st dt & s \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 2s & s \in [0, 1] \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

Also:

$$\begin{aligned}
g(t) &= \int_{-\infty}^{\infty} \varphi(s, t) ds \\
&= \begin{cases} \int_0^1 4st \, ds & t \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \int_0^1 4st \, ds & t \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} 2t & t \in [0, 1] \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Example 48. Going back to the beginning of the section, let

$$\varphi(s, t) = \begin{cases} 8st & 0 \leq s \leq t, 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and suppose $(\xi, \eta) \sim \varphi$. Then from the calculation we've done, we see that:

$$\xi \sim f(s) = \begin{cases} 4s - 4s^3 & 0 \leq s \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\eta \sim g(t) = \begin{cases} 4t^3 & 0 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Notice something different between the examples 47 and 48? In the former, the product of the individual pdfs gave us the joint one, but in the latter it didn't. Sounds like independence must be involved here...

Definition 30. We say that continuous random variables ξ_1, \dots, ξ_n on $(\Omega, \mathcal{F}, \mathbb{P})$ are *independent* if they are jointly continuous with the joint pdf given by

$$\varphi(t_1, \dots, t_n) = f_1(t_1) \cdots f_n(t_n)$$

where $\xi_i \sim f_i$ for each $i = 1, \dots, n$.

So now we see that ξ and η in Ex. 47 are independent, while those in Ex. 48 are not.

Theorem 17. Suppose ξ, η are jointly distributed continuous random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with the pdf of $\mathbb{P}_{(\xi, \eta)}$ being $\varphi(s, t)$. Then:

1. $\xi + \eta$ is a continuous random variable with pdf $\int_{-\infty}^{\infty} \varphi(v, u - v) dv$;

2. if ξ and η are independent, the $\xi + \eta$ has pdf given by $f * g$ where $\xi \sim f$ and $\eta \sim g$.

Before we prove the theorem, we can see that the following things that you proved on the HW, follow from it directly:

Corollary 3. Suppose ξ and η are independent. Then:

1. $\xi \sim n(a_1, \sigma_1^2), \eta \sim n(a_2, \sigma_2^2) \implies \xi + \eta \sim n(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$
2. $\xi \sim g_{(\lambda, t_1)}, \eta \sim g_{(\lambda, t_2)} \implies \xi + \eta \sim g_{\lambda, t_1 + t_2}$

$$3. \xi \sim U(0, 1), \eta \sim U(0, 1) \implies \xi + \eta \sim \begin{cases} t & 0 \leq t \leq 1 \\ 2 - t & 1 \leq t \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

$\xi + \eta$ in (3) is called the *Simpson's triangular* distribution.

Also, before we proceed with the proof, let us review some multivariable calculus: the change of variables and the Jacobian. Suppose we have the transformation

$$T : \mathbb{R}^2 \rightarrow \mathbb{R}^2 : (u, v) \mapsto \begin{cases} x = x(u, v) \\ y = y(u, v) \end{cases}$$

Assume that T maps a region D in \mathbb{R}^2 injectively, except possibly finitely many curves, onto the region R . Then:

$$\iint_R \varphi(x, y) dx dy = \iint_D \varphi(x(u, v), y(u, v)) |x_u y_v - x_v y_u| du dv$$

Recall that $|x_u y_v - x_v y_u| = \left| \det \begin{pmatrix} x_u & x_v \\ y_u & y_v \end{pmatrix} \right| = J$ is the Jacobian of the transformation. An example of the Jacobian is the r when we change from rectangular to polar coordinates. We know that $x = r \cos \theta$ and $y = r \sin \theta$. So the Jacobian is: $\left| \det \begin{pmatrix} x_r & x_\theta \\ y_r & y_\theta \end{pmatrix} \right| = \left| \det \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix} \right| = r$. (Note, however, that the mapping is not bijective, as $(0, \theta) \mapsto (0, 0)$ for all θ .)

Now we can begin the proof:

Proof of 17. To prove (1), observe:

$$\begin{aligned} \mathbb{P}\{\xi + \eta \leq x\} &= \mathbb{P}_{(\xi, \eta)}\{t + s \leq x\} \\ &= \mathbb{P}_{(\xi, \eta)}\{(s, t) : s + t \leq x\} \\ &= \iint_B \varphi(s, t) ds dt & B = \{(s, t) : s + t \leq x\} \\ &= \iint_D \varphi(v, u - v) du dv & \text{change of variables} \\ &= \int_{-\infty}^x \int_{-\infty}^{\infty} \varphi(v, u - v) dv du \end{aligned}$$

We're done since $\int_{-\infty}^{\infty} \varphi(v, u-v)dv$ is the desired pdf. The change of variables we did along the way is as follows:

$$\begin{cases} u = s + t \\ v = s \end{cases} \implies \begin{cases} t = u - v \\ s = v \end{cases} \implies J = \left| \det \begin{pmatrix} s_u & s_v \\ t_u & t_v \end{pmatrix} \right| = \left| \det \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} \right| = 1$$

To prove (2), note that, by the definition of independence, we have $\varphi(s, t) = f(s)g(t)$. So, by (1), if $\xi + \eta \sim h(u)$, we have:

$$\begin{aligned} h(u) &= \int_{-\infty}^{\infty} \varphi(v, u-v)dv \\ &= \int_{-\infty}^{\infty} f(v)g(u-v)dv \\ &= (f * g)(u) \end{aligned}$$

□

3.5 Quotients of Random Variables

1. Let ξ and η be jointly continuous with density $\varphi(s, t)$. Show that ξ/η is also a continuous random variable with probability density function

$$h(u) = \int_{-\infty}^{+\infty} \varphi(uv, v)|v|dv, \quad -\infty < u < +\infty.$$

Hint: imitate the proof for sums and consider the change of variables $u = s/t$ and $v = t$.

Proof. Observe:

$$\begin{aligned} \mathbb{P}\{\xi/\eta \leq x\} &= \mathbb{P}\{\xi/\eta \leq x, -\infty \leq \eta \leq \infty\} \\ &= \mathbb{P}_{(\xi, \eta)}\{(s, t) : s/t \leq x\} \\ &= \iint_B \varphi(s, t)dsdt & B = \{(s, t) : s/t \leq x\} \\ &= \iint_D \varphi(uv, v)|v|dvdu & D = \{(u, v) : u \leq x, v \in \mathbb{R}\} \\ &= \int_{-\infty}^x \int_{-\infty}^{\infty} \varphi(uv, v)|v|dvdu \end{aligned}$$

We're done since $\int_{-\infty}^{\infty} \varphi(uv, v)|v|dv$ is the desired pdf. The change of variables we did along the way is as follows:

$$\begin{cases} u = s/t \\ v = t \end{cases} \implies \begin{cases} s = uv \\ t = v \end{cases} \implies J = \left| \det \begin{pmatrix} s_u & s_v \\ t_u & t_v \end{pmatrix} \right| = \left| \det \begin{pmatrix} v & u \\ 0 & 1 \end{pmatrix} \right| = |v|$$

□

(**Remark**) Strictly speaking, $(\xi/\eta)(\omega)$ is undefined if $\eta(\omega) = 0$. But the latter happens with probability zero. So this is not an issue as long as we define $(\xi/\eta)(\omega) = 0$ whenever $\eta(\omega) = 0$.

2. Suppose that ξ and η are independent and both follow the standard normal distribution $N(0, 1)$. Show that ξ/η has the *Cauchy density*

$$\frac{1}{\pi} \cdot \frac{1}{1 + u^2}, \quad -\infty < u < +\infty.$$

Proof. By part (a), we get that the pdf of ξ/η is as follows:

$$\begin{aligned} h(u) &= \int_{-\infty}^{\infty} \varphi(uv, v) |v| dv \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(uv)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} |v| dv && \text{independent} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{(uv)^2 + v^2}{2}} |v| dv \\ &= 2 \cdot \frac{1}{2\pi} \int_0^{\infty} e^{-\frac{(u^2+1)v^2}{2}} v dv && \text{even function} \\ &= \frac{1}{\pi} \left(\frac{1}{u^2 + 1} - e^{-\frac{(u^2+1)v^2}{2}} \right) \Big|_{v=0}^{v=\infty} \\ &= \frac{1}{\pi} \cdot \frac{1}{u^2 + 1} \end{aligned}$$

□

3. Suppose that ξ and η are independent and $\xi \sim \text{Gamma}(1/2, m/2)$ and $\eta \sim \text{Gamma}(1/2, n/2)$, where m and n are positive integers. Show that ξ/η has the density

$$h(u) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \cdot \Gamma(\frac{n}{2})} \cdot \frac{u^{\frac{m}{2}-1}}{(u+1)^{\frac{m+n}{2}}} & u > 0 \\ 0 & u \leq 0. \end{cases}$$

Proof. Recall that the gamma density is defined as follows:

$$g_{\lambda, r} = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

where $\Gamma(r) = \int_0^{\infty} t^{r-1} e^{-t} dt$ whenever $r > 0$. Then, we just have to compute it using the formula proved in (a):

$$\begin{aligned} h(u) &= \int_{-\infty}^{\infty} g_{\frac{1}{2}, \frac{m}{2}}(uv) \cdot g_{\frac{1}{2}, \frac{n}{2}}(v) \cdot |v| dv \\ &= \int_0^{\infty} g_{\frac{1}{2}, \frac{m}{2}}(uv) \cdot g_{\frac{1}{2}, \frac{n}{2}}(v) \cdot v dv && 0 \text{ whenever } v \leq 0 \end{aligned}$$

Now, let us consider two cases.

Case 1: $u \leq 0$. Then for any $v > 0$, we have $uv \leq 0$, and so $g_{1/2, m/2}(uv) = 0$, by the way the gamma density is defined. Hence the entire integral is 0 as well.

Case 2: $u > 0$. Then:

$$\begin{aligned}
h(u) &= \int_0^\infty \frac{1}{\Gamma(\frac{m}{2})} (1/2)^{m/2} (uv)^{m/2-1} e^{-uv/2} \frac{1}{\Gamma(\frac{n}{2})} (1/2)^{n/2} (uv)^{n/2-1} e^{-v/2} \\
&= \frac{(\frac{1}{2})^{\frac{(m+n)}{2}} u^{m/2-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty v^{\frac{m+n}{2}-1} e^{-\frac{(u+1)}{2}v} dv \\
&= \frac{(\frac{1}{2})^{\frac{(m+n)}{2}} u^{m/2-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \int_0^\infty \left(\frac{2t}{u+1} \right)^{\frac{m+n}{2}-1} e^{-t} \frac{2}{u+1} dt & t = \frac{u+1}{2}v \implies v = \frac{2t}{u+1} \\
&= \frac{u^{m/2-1}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \left(\frac{1}{u+1} \right)^{\frac{(m+n)}{2}} \int_0^\infty t^{\frac{m+n}{2}-1} e^{-t} dt \\
&= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2}) \cdot \Gamma(\frac{n}{2})} \cdot \frac{u^{\frac{m}{2}-1}}{(u+1)^{\frac{m+n}{2}}}
\end{aligned}$$

Thus, we conclude that indeed

$$h(u) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{u^{\frac{m}{2}-1}}{(u+1)^{\frac{m+n}{2}}} & u > 0 \\ 0 & u \leq 0. \end{cases}$$

□

3.6 F and T Distributions

We'll conclude this chapter by discussing two very important continuous distributions in probability theory, the F -distribution and the T -distribution. Their definitions are as follows:

Definition 31.

1. Let ξ and θ be independent random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, if $\xi \sim \text{Gamma}(\frac{1}{2}, \frac{m}{2})$ and $\theta \sim \text{Gamma}(\frac{1}{2}, \frac{n}{2})$, the law of $\frac{\xi/m}{\theta/n}$ is called the F -distribution with parameters m and n , denoted by $F(m, n)$.
2. If ξ and θ are independent and $\xi \sim \mathcal{N}(0, 1)$ and $\theta \sim \text{Gamma}(\frac{1}{2}, \frac{n}{2})$, then the law of $\frac{\xi}{\sqrt{\theta/n}}$ is called the T -distribution with parameter n , denoted by $T(n)$.

Proposition 13. If ξ has pdf f where $f(t) = 0$ for all $t \leq 0$, then $\sqrt{\xi}$ has pdf given by:

$$g(t) = \begin{cases} 2tf(t^2) & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Before we prove this proposition, let us proceed like you did in one of the homeworks – we will compute the cdf and then differentiate using the Fundamental theorem of calculus to get the pdf. Observe:

$$\begin{aligned} \mathbb{P}\{\sqrt{\xi} \leq x\} &= \mathbb{P}\{\xi \leq x^2\} \\ &= \mathbb{P}\{0 \leq \xi \leq x^2\} & f(t) = 0 \text{ when } t \leq 0 \\ &= \int_0^{x^2} f(t)dt \end{aligned}$$

Using the Fundamental Theorem of Calculus and the chain rule for differentiation, we get:

$$\frac{d}{dx} \mathbb{P}\{\sqrt{\xi} \leq x\} = 2xf(x^2)$$

However, this is **not** a proof! The method above only works when f is continuous. If it isn't, then it cannot be differentiable! Since we cannot assume that f is continuous in general, we need to be more careful.

Proof. Note that it is enough to show that $\mathbb{P}\{\sqrt{\xi} \leq x\} = \int_{-\infty}^x g(t)dt$. Since g is piece-wise defined, we need to consider the two cases:

Case $x < 0$: Since square root is always non-negative, we get that $\{\omega \in \Omega : \sqrt{\xi} \leq x\} = \emptyset$. Hence, $\mathbb{P}\{\sqrt{\xi} \leq x\} = 0$ and hence for all $t < 0$, we get:

$$\int_{-\infty}^x g(t)dt = 0$$

Case $x \geq 0$: Observe that in this case,

$$\begin{aligned} \int_{-\infty}^x g(t)dt &= \int_0^x g(t)dt \\ &= \int_0^x 2tf(t^2)dt & y = t^2 \implies dy = 2tdt \\ &= \int_0^{x^2} f(y)dy \\ &= \mathbb{P}\{0 \leq \xi \leq x^2\} \\ &= \mathbb{P}\{\sqrt{\xi} \leq x\} \end{aligned}$$

And we're done! It is very important to note that here, we didn't require the continuity assumption, because the change of variable formula doesn't require the the function to be integrated to be continuous. This is not the case with differentiability – a function is differentiable only if it's continuous! \square

The above proof, in a way, also gives us an "algorithm" of how to proceed when we need to compute a pdf of a function of a random variable. We first need to "guess" the pdf by differentiating the cdf forgetting the differentiability issues and then prove that the guess is correct, as shown above.

Example 49. If ξ is a random variable with pdf f and $a > 0$, find the pdf of $a\xi$.

Proof. By the remark above, we will first guess the pdf of $a\xi$. Proceed by finding the cdf first:

$$\begin{aligned}\mathbb{P}\{a\xi \leq x\} &= \mathbb{P}\left\{\xi \leq \frac{x}{a}\right\} \\ &= \int_{-\infty}^{x/a} f(t)dt\end{aligned}$$

Now, we differentiate:

$$\frac{d}{dx} \int_{-\infty}^{x/a} f(t)dt = \frac{1}{a}f\left(\frac{x}{a}\right)$$

So, our guess is that

$$g(t) = \frac{1}{a}f\left(\frac{t}{a}\right)$$

is the cdf of $a\xi$. Let us now verify that. As before, it suffices to show that $\mathbb{P}\{a\xi \leq x\} = \int_{-\infty}^x g(t)dt$. Observe:

$$\begin{aligned}\int_{-\infty}^x \frac{1}{a}f\left(\frac{t}{a}\right)dt &= \int_{-\infty}^{x/a} f(y)dy & y = t/a \implies dt/a = dy \\ &= \mathbb{P}\{\xi \leq x/a\} \\ &= \mathbb{P}\{a\xi \leq x\}\end{aligned}$$

So, indeed, g is the pdf of $a\xi$. □

It would be a good exercise to check that when we allow a to be negative as well, that is, when $a \neq 0$, then the pdf of $a\xi$ is $g(t) = \frac{1}{|a|}f\left(\frac{t}{a}\right)$.

Theorem 18. The pdf of the $F(m, n)$ is given by:

$$f(t; m, n) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{m^{\frac{m}{2}} t^{\frac{m}{2}-1}}{n^{\frac{m}{2}} (1+\frac{mt}{n})^{\frac{m+n}{2}}} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Proof. We know how to compute the pdf of the quotient of two gamma-distributed random variables. So $\xi/\theta \sim h(t)$ where:

$$h(u) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{t^{\frac{m}{2}-1}}{(t+1)^{\frac{m+n}{2}}} & t > 0 \\ 0 & t \leq 0. \end{cases}$$

Now, simply note that $\frac{\xi/m}{\theta/n} = \frac{n}{m} \cdot \frac{\xi}{\theta}$. Applying 49 with $a = \frac{m}{n}$, we get the desired result. □

Let us consider some of the applications of the F -distribution. Let $X_1, \dots, X_m, Y_1, \dots, Y_n$ be independent and suppose all of them are standard normal random variables, i.e. $X_i \sim \mathcal{N}(0, 1)$ for all $i = 1, \dots, m$ and $Y_j \sim \mathcal{N}(0, 1)$ for all $j = 1, \dots, n$. Consider the following random variable:

$$Z = \frac{\frac{X_1^2 + \dots + X_m^2}{m}}{\frac{Y_1^2 + \dots + Y_n^2}{n}}$$

Recall that we previously said that $X \sim \mathcal{N}(0, 1) \implies X^2 \sim \chi^2 = g_{\frac{1}{2}, \frac{1}{2}}$. Also, recall that for a sum of independent random variables, the pdf is the convolution of their respective pdfs. So, the pdf of $X_1^2 + \dots + X_m^2$ is

$$\underbrace{g_{\frac{1}{2}, \frac{1}{2}} * \dots * g_{\frac{1}{2}, \frac{1}{2}}}_m = g_{\frac{1}{2}, \frac{m}{2}}$$

Similarly, the pdf of $Y_1^2 + \dots + Y_n^2$ is

$$\underbrace{g_{\frac{1}{2}, \frac{1}{2}} * \dots * g_{\frac{1}{2}, \frac{1}{2}}}_n = g_{\frac{1}{2}, \frac{n}{2}}$$

Hence, we conclude Z follows the F -distribution. It is a very important random variable in mathematical statistics and you can learn more about it if you take Math 232.

By similar reasoning, we can see that if ξ, Y_1, \dots, Y_n are independent variables following the standard normal distribution, then the random variable

$$\frac{\xi}{\sqrt{\frac{Y_1^2 + \dots + Y_n^2}{n}}}$$

is T -distributed with parameter n .

It is very important to remark, however, that in the reasoning above, we assumed some things the proof of which is nontrivial. Firstly, when taking the convolution of pdfs, we had to assume that the squares of independent random variables are independent. Moreover, in order to apply the result for the ratio of two random variables, we had to assume that $X_1^2 + \dots + X_m^2$ and $Y_1^2 + \dots + Y_n^2$ are independent. These things are true, and you will probably prove them in Math 232.

Now, let's compute the pdf for the T -distribution. Let us proceed step-by-step:

1. First, observe that by 13, we have that:

$$\sqrt{\theta} \sim \begin{cases} \frac{2}{\Gamma(\frac{n}{2})} \cdot \left(\frac{1}{2}\right)^{\frac{n}{2}} \cdot t^{n-1} \cdot e^{-\frac{t^2}{2}} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

2. Then, by 49, we have that:

$$\frac{1}{\sqrt{n}}\sqrt{\theta} \sim h(t) = \begin{cases} \frac{2\sqrt{n}}{\Gamma(\frac{n}{2})} \cdot \left(\frac{1}{2}\right)^{\frac{n}{2}} (\sqrt{nt})^{n-1} e^{-\frac{nt}{2}} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

3. Finally, we are ready to compute the pdf of the T -distribution by applying the result for the ratio of two random variables:

$$\begin{aligned} \frac{\xi}{\sqrt{\theta/n}} &\sim \int_{-\infty}^{\infty} \mathcal{N}_{0,1}(uv)h(v)|v|dv \\ &= \int_0^{\infty} \mathcal{N}_{0,1}(uv)h(v)v dv \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(uv)^2}{2}} \frac{2\sqrt{n}}{\Gamma(\frac{n}{2})} \cdot \left(\frac{1}{2}\right)^{\frac{n}{2}} (\sqrt{nv})^{n-1} e^{-\frac{nv^2}{2}} v dv \\ &= \frac{1}{\sqrt{2\pi}} \frac{2\sqrt{n}}{\Gamma(\frac{n}{2})} \left(\frac{1}{2}\right)^{\frac{n}{2}} (\sqrt{n})^{n-1} \int_0^{\infty} v^n e^{-\frac{(uv)^2 + nv^2}{2}} dv \\ &= \sqrt{\frac{2}{\pi}} \frac{(\sqrt{n})^n}{\Gamma(\frac{n}{2})} \left(\frac{1}{2}\right)^{\frac{n}{2}} \int_0^{\infty} \frac{2^{n/2}}{(u^2 + n)^{n/2}} t^{n/2} e^{-t} \sqrt{\frac{2}{u^2 + n}} \frac{1}{2\sqrt{2}} dt \quad (*) \\ &= \sqrt{\frac{1}{\pi}} \frac{(\sqrt{n})^n}{\Gamma(\frac{n}{2})} \frac{1}{(u^2 + n)^{1/2}} \left(\int_0^{\infty} t^{n/2-1} e^{-t} dt \right) \frac{1}{(u^2 + n)^{1/2}} \\ &= \sqrt{\frac{1}{n\pi}} \frac{(\sqrt{n})^{n+1}}{\Gamma(\frac{n}{2})} \Gamma\left(\frac{n+1}{2}\right) \frac{1}{(u^2 + n)^{\frac{n+1}{2}}} \\ &= \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \frac{1}{\left(\frac{u^2}{n} + 1\right)^{\frac{n+1}{2}}} \end{aligned}$$

where (*) follows from the substitution:

$$t = \frac{(u^2 + n)v^2}{2} \implies v = \sqrt{\frac{2}{u^2 + n}} \sqrt{t} \implies dv = \sqrt{\frac{2}{u^2 + n}} \frac{1}{2\sqrt{t}} = \frac{1}{\sqrt{u^2 + n} \sqrt{2} \sqrt{t}}$$

4 Limit Theorems

4.1 Poisson Distribution

In this section, we will talk about a very important distribution in probability theory, the Poisson Distribution. It is a discrete distribution, but it comes from a continuous process. Before we do this, let us review the Gamma function we saw earlier in the course. Recall that the Gamma function is defined as

$$\Gamma(r) = \int_0^{\infty} t^{r-1} e^{-t} dt$$

whenever $r > 0$. We would like to compute the general formula for $\Gamma(n)$. First, observe:

$$\Gamma(1) = \int_0^{\infty} e^{-t} dt = -e^{-t} \Big|_0^{\infty} = 1$$

Next, we will compute $\Gamma(r+1)$ in terms of $\Gamma(r)$:

$$\begin{aligned} \Gamma(r+1) &= \int_0^{\infty} t^r e^{-t} dt \\ &= -\frac{t^r}{e^t} \Big|_0^{\infty} + \int_0^{\infty} r e^{-t} t^{r-1} dt \\ &= r\Gamma(r) \end{aligned} \quad \text{L'Hopital } r \text{ times}$$

Finally we can derive an inductive formula for $\Gamma(n)$ as follows:

$$\begin{aligned} \Gamma(2) &= \Gamma(1+1) = \Gamma(1) = 1 \\ \Gamma(3) &= \Gamma(2+1) = 2\Gamma(2) = 2 \cdot 1 \\ \Gamma(4) &= \Gamma(3+1) = 3\Gamma(2) = 3 \cdot 2 \cdot 1 \\ &\dots \\ \Gamma(n) &= (n-1)! \end{aligned}$$

Problem 2. Consider a sequence of random events occurring in time (involving things like: calls at a telephone exchange, radioactive disintegration, goals in soccer matches). Each such occurrence is called an "arrival". The question we would like to ask is: What is the number of arrivals that occur in the time period $[0, t]$. Before we try to answer this question, we need to state all the assumptions:

1. time is continuous;
2. the first arrival is a "memoryless" continuous random variable, i.e. if η_1 is a random variable representing the first arrival, then $\mathbb{P}\{\eta_1 > s\} = \mathbb{P}\{\eta_1 > s+t \mid \eta_1 > t\}$. [***Fun*** fact: it can be proven that the exponential random variable is the only memoryless continuous random variable, so this tells us that $\eta_1 \sim \text{Exp}(\lambda)$];
3. each arrival should behave like the first one – arrivals are mutually independent.

Definition 32. By *arrival process* with parameter $\lambda > 0$, we mean a sequence $\{\eta_n\}$ of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ such that:

1. η_1, \dots, η_n are independent for all $n \in \mathbb{N}$;
2. $\eta_i \geq 0$ for all i ;

3. $\eta_i \sim \text{Exp}(\lambda)$ where the exponential distribution has density given by:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

Definition 33. Define $W_n : \Omega \rightarrow \mathbb{R}$ by:

$$W_n = \begin{cases} 0 & n = 0 \\ \eta_1 + \dots + \eta_n & n \geq 1 \end{cases}$$

So W_n represents the waiting time for the n th arrival.

Let us compute the law of W_n , \mathbb{P}_{W_n} . First, let $f(t) = g_{\lambda,r}(t)$. Recall that the Gamma distribution is defined as:

$$g_{\lambda,r}(t) = \begin{cases} \frac{1}{\Gamma(r)} \lambda^r t^{r-1} e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

where $\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$. Hence, we see that

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

Hence, we see that the exponential distribution is a special case of the Gamma distribution. So, since we proved in the homework that

$$g_{\lambda,r_1} * g_{\lambda,r_2} = g_{\lambda,r_1+r_2}$$

we can conclude that

$$\underbrace{(f * \dots * f)}_n(t) = g_{\lambda,n}(t) = \begin{cases} \frac{1}{\Gamma(n)} \lambda^n t^{n-1} e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases} = \begin{cases} \frac{1}{(n-1)!} \lambda^n t^{n-1} e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

The latter is called the *Erlang* distribution with parameters λ and n . Next, we claim that

$$F_{W_n}(x) = \begin{cases} 1 - e^{-\lambda x} \left[1 + \frac{\lambda x}{1!} + \dots + \frac{(\lambda x)^{n-1}}{(n-1)!} \right] & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Proof. The case when $x \leq 0$ is obvious, so we just need to focus on the case when $x > 0$. Let us go backwards. We need to show:

$$\int_0^x \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} dt = 1 - e^{-\lambda x} \left[1 + \frac{\lambda x}{1!} + \dots + \frac{(\lambda x)^{n-1}}{(n-1)!} \right]$$

Note the LHS by $F_{W_n}(x)$ and denote the RHS by $F(x)$. Equivalently, we need to show:

$$\int_0^x \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} dt = F(x) - F(0)$$

But note that the function is continuous, so by the Fundamental Theorem of Calculus, this is equivalent to:

$$F'(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t}$$

for all $x > 0$. But this can be verified by direct differentiation! Done. \square

Lemma 5. For any $t > 0$ and any $\omega \in \Omega$, if $k \in \{n \in \mathbb{N} : W_n(\omega) \leq t\}$ then $\{0, 1, 2, \dots, k\} \subseteq \{n \in \mathbb{N}^+ : W_n(\omega) \leq t\}$.

Proof. The proof is sort of obvious once you translate the mathematical notation above into English. All that it's saying is that if the time for the k th arrival is at most t , then the time for any arrival before it is also at most t . Mathematically, this means for any $i \leq k$:

$$\eta_1(\omega) + \dots + \eta_i(\omega) \leq \eta_1(\omega) + \dots + \eta_i(\omega) + \dots + \eta_k(\omega) \leq t$$

because of the assumption (2) in 32. \square

Definition 34. For all $t > 0$, define:

$$N_t(\omega) = \begin{cases} k & \{n \in \mathbb{N} : W_n(\omega) \leq t\} = \{0, 1, \dots, k\} \\ -1 & \{n \in \mathbb{N} : W_n(\omega) \leq t\} = \mathbb{N} \end{cases}$$

So we see that $N_t(\omega)$ reveals the number of arrivals in the time period $[0, t]$. Note that

$$N_t(\omega) = k \iff W_k(\omega) \leq t \text{ and } W_{k+1}(\omega) > t$$

Note that the right-hand side means that the waiting time for the k th arrival is at most t and the waiting time for the $(k+1)$ st arrival is greater than t . So, clearly, this means that there are precisely k arrivals in the period $[0, t]$.

Theorem 19. $\mathbb{P}_{N_t} = \sum_{k=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \delta_k$

Before we prove the theorem, let us state a very important definition:

Definition 35. The special case when $t = 1$, $\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \delta_k$ is called the *Poisson Distribution* with parameter λ , denoted by $\text{Poisson}(\lambda)$.

One can show that if $\xi \sim \text{Poisson}(\lambda)$, the $\mathbb{E}\xi = \text{Var}\xi = \lambda$. The conclusion is that **in the arrival process with parameter λ described in 32, the number of arrivals in $[0, t]$ follows $\text{Poisson}(\lambda t)$.**

Proof of 19: Notice:

- $\{N_t = k\} = \{W_k \leq t\} - \{W_{k+1} \leq t\}$
- $\{W_{k+1} \leq t\} \subseteq \{W_k \leq t\}$

Now, we'll first consider the case when $k \geq 0$. Using the first observation above:

$$\begin{aligned}\mathbb{P}\{N_t = k\} &= \mathbb{P}\{W_k \leq t\} - \mathbb{P}\{W_{k+1} \leq t\} \\ &= \left(1 - e^{-\lambda t} \left[1 + \frac{\lambda x}{1!} + \dots + \frac{(\lambda x)^{k-1}}{(k-1)!}\right]\right) - \left(1 - e^{-\lambda t} \left[1 + \frac{\lambda x}{1!} + \dots + \frac{(\lambda x)^k}{k!}\right]\right) \\ &= e^{-\lambda t} \frac{(\lambda t)^k}{k!}\end{aligned}$$

when $k \geq 0$. Now, consider the case when $k = -1$:

$$\begin{aligned}\mathbb{P}\{N_t = -1\} &= 1 - \sum_{k=0}^{\infty} \mathbb{P}\{N_t = k\} \\ &= 1 - e^{-\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \\ &= 1 - 1 = 0\end{aligned}$$

where the last line follows from the fact that the Taylor expansion of e^x is:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

where $x = \lambda t$. □

It's worth noting that original probability space for the Poisson distribution is very complicated, so we don't have the tools to rigorously talk about it (loosely speaking, it is a set of certain step functions). Pursue a PhD in Probability Theory to find out more...

Now let's see how Poisson distribution works in real life. Consider the following FIFA 2010 statistics:

# goals	0	1	2	3	4	5	6	7
# games	7	17	13	14	7	5	0	1

One can check that the average number of goals per game is $\lambda t = 2.265$. Note that the probability that there are no goals in a game is exactly $\frac{7}{64} = 0.109$. Using the Poisson distribution to calculate the same number, we get

$$e^{-\lambda t} \frac{(\lambda t)^0}{0!} = e^{-2.265} = 0.1038$$

which is a pretty close approximation. Let us compare other values:

k	0	1	2	3	4	5	6	7
Frequency	0.109	0.266	0.203	0.219	0.109	0.078	0	0.016
Poisson	0.104	0.236	0.267	0.202	0.114	0.052	0.020	0.006

where we used the Poisson formula

$$e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

for calculating the second row of the table.

4.2 Poisson Limit Theorem

Recall the following important assumptions we put on the arrival process we talked about before:

1. arrivals are independent;
2. each arrival time follows the same memoryless distribution.

Let η_i denote the waiting time between $(i-1)$ st and i th arrival. Recall we defined:

$$W_n = \begin{cases} \eta_1 + \dots + \eta_n & n > 0 \\ 0 & n = 0 \end{cases}$$

Also recall that by N_t we denoted the number of arrivals in $[0, t]$. The below table summarizes the arrival processes we talked about for both continuous and discrete case:

Continuous Time	Expectation (continuous)	Discrete Time	Expectation (discrete)
$\eta_i \sim \text{Exp}(\lambda)$	$\mathbb{E}\eta_i = \frac{1}{\lambda}$	$\eta_i \sim \text{Geo}(p)$	$\mathbb{E}\eta_i = \frac{1}{p}$
$W_n \sim \text{Erlang}(n, \lambda)$	$\mathbb{E}W_n = \frac{n}{\lambda}$	$W_n \sim \text{NB}(n, p)$	$\mathbb{E}W_n = \frac{n}{p}$
$N_t \sim \text{Poisson}(\lambda t)$	$\mathbb{E}N_n = \lambda t$	$S_n \sim \text{Bin}(n, p)$	$\mathbb{E}S_n = np$

Theorem 20. [Poisson Limit Theorem] Suppose that $\lim_{n \rightarrow \infty} np_n = \lambda$. Then $\lim_{n \rightarrow \infty} b(k; n, p_n) = p(k; \lambda)$ where

$$p(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Recall that the the Poisson Distribution law is defined as follows:

$$\text{Poisson}(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \delta_k = \sum_{k=0}^{\infty} p(k; \lambda) \delta_k$$

The above theorem says that as the expectation of binomial distribution goes to λ , the limit of binomial probability mass function approaches the probability mass function of Poisson distribution.

Proof. Let $\lambda_n = np_n$. By assumption, $\lambda_n \rightarrow \lambda$. Note:

$$\begin{aligned} b(k; n, p_n) &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\ &= \frac{n!}{k!(n-k)!} \cdot \frac{(\lambda_n)^k}{n^k} \cdot \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \end{aligned}$$

Notice that:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{n!}{(n-k)! \cdot n^k} &= \lim_{n \rightarrow \infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \\ &= \lim_{n \rightarrow \infty} \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \\ &= 1 \cdot 1 \cdots 1 \\ &= 1 \end{aligned}$$

Moreover, observe that $\lim_{n \rightarrow \infty} \lambda_n^k = \lambda^k$, as $\lim_{n \rightarrow \infty} \lambda_n = \lambda$. Thus, it remains to show that

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda_n}{n}\right)^{n-k} = e^{-\lambda}$$

This is not hard to show:

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda_n}{n}\right)^{n-k} &= \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda_n}{n}\right)^n \\ &= \lim_{n \rightarrow \infty} e^{n \log(1 - \frac{\lambda_n}{n})} \\ &= \lim_{n \rightarrow \infty} e^{n(-\frac{\lambda_n}{n} + \theta_n(\frac{\lambda_n}{n})^2)} && \text{for some } |\theta_n| < 1 \\ &= \lim_{n \rightarrow \infty} e^{-\lambda_n + \theta_n \frac{(\lambda_n)^2}{n}} \\ &= e^{-\lambda} && \lambda_n \rightarrow \lambda \text{ and } \frac{\lambda_n}{n} \rightarrow 0 \end{aligned}$$

Finally, we can conclude that $b(k; n, p_n) \rightarrow p(k; \lambda)$, as desired. \square

Example 50 (A math model in Maxwell-Boltzmann Statistics). Consider the following set-up: n (distinguishable) balls are distributed among r (distinguishable) cells, where each of the r^n outcomes are equally probable. Let $q(k; n, r)$ denote the probability that cell $\boxed{1}$ contains exactly k balls. Note that:

$$q(k; n, r) = \frac{\binom{n}{k} (r-1)^{n-k}}{r^n} = \binom{n}{k} \left(\frac{1}{r}\right)^k \left(1 - \frac{1}{r}\right)^{n-k}$$

This is exactly the probability of getting k successes in a binomial experiment with parameters n and $\frac{1}{r}$, i.e. $q(k; n, r) = b(k; n, \frac{1}{r})$. We will let $r = r_n$ depend on n and let the average

number of balls per cell approach λ , that is $\frac{n}{r_n} \rightarrow \lambda$. These are clearly reasonable assumptions on the model. Then, by 20, we get:

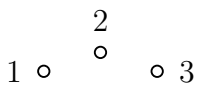
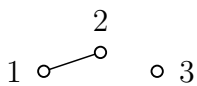
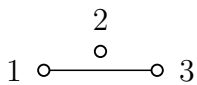
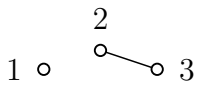
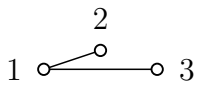
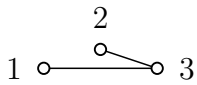
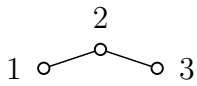
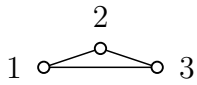
$$\begin{aligned}\lim_{n \rightarrow \infty} q(k; n, r_n) &= p(k; \lambda) \\ &= \frac{\lambda^k}{k!} e^{-\lambda}\end{aligned}$$

Example 51 (Erdos-Renyi Random Graph). Suppose we have a set of n points (called vertices) and any two points have the probability p of being connected by an edge. This probabilistic model is called a *random graph*, which we will denote by $G(n, p) = (\Omega, 2^\Omega, \mathbb{P})$. Note that there are $\binom{n}{2}$ ways to pick a pair of points from the set of n points, and hence there are $\binom{n}{2}$ edges possible in the graph (we assume the graph is simple – no loops or multiple edges between vertices are allowed). Then, our sample spaces must be defined as follows:

$$\Omega = \{(a_1, \dots, a_{\binom{n}{2}}) : a_i \in \{0, 1\}\}$$

where 1 denotes success (an edge) and 0 denotes failure (no edge). It is clear that each graph with exactly k edges has probability of $p^k(1-p)^{\binom{n}{2}-k}$, by the binomial formula.

Now, consider the concrete model, $G(3, \frac{1}{3})$. So, our vertex set consists of three vertices (say, $V = \{\textcircled{1}, \textcircled{2}, \textcircled{3}\}$) and any two vertices have the probability of $\frac{1}{3}$ of being adjacent. Suppose the first term in the sequence correspond to the edge $\{1, 2\}$, the second one corresponds to $\{1, 3\}$ and the third one corresponds to $\{2, 3\}$. Let us draw all possible random graphs satisfying the model and compute the probability of each:

Sequence	Graph	Probability
$(0,0,0)$		$(\frac{2}{3})^3$
$(1,0,0)$		$\frac{1}{3}(\frac{2}{3})^2$
$(0,1,0)$		$\frac{1}{3}(\frac{2}{3})^2$
$(0,0,1)$		$\frac{1}{3}(\frac{2}{3})^2$
$(1,1,0)$		$\frac{2}{3}(\frac{1}{3})^2$
$(0,1,1)$		$\frac{2}{3}(\frac{1}{3})^2$
$(1,0,1)$		$\frac{2}{3}(\frac{1}{3})^2$
$(1,1,1)$		$(\frac{1}{3})^3$

Now, one can ask questions like: what is the probability that $G(3, \frac{1}{3})$ is connected (i.e. there is a path between any two vertices)? Well, only the last four graphs in the table are connected, so by summing their probabilities, one would obtain the desired number.

Next, let $q(k; n, p)$ denote the probability that vertex ① is incident to exactly k edges. Note that each vertex can be adjacent to at most $n - 1$ other vertices, so the probability that exactly k of the adjacencies are present is $\binom{n-1}{k} p^k (1-p)^{n-1-k} = b(k; n-1, p)$. Again, let $p = p_n$ depend on n and $np_n \rightarrow \lambda$. Note that this is a reasonable assumption, as we can think of graphs as, for examples, modeling social networks. If an edge between two vertices (people) represents their friendship, no matter how large the sample size grows, the total number of people you are friends with must still remain the same. Hence, the probability that a random person is your friend must decrease as n gets larger.

So, $np_n \rightarrow \lambda \implies p_n \rightarrow 0 \implies (n-1)p_n = np_n - p_n \rightarrow \lambda$. Using the Poisson Limit Theorem, we get:

$$q(k; n, p_n) \rightarrow p(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$$

When n is sufficiently large, the above two numbers are very close.

4.3 Local Limit Theorem

In this section, we will prove the Local Limit Theorem in several steps.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the infinite coin toss model with parameter $0 < p < 1$, and S_n be the number of cumulative successes in the first n trials. Let us put $b(k; n, p) = \mathbb{P}\{S_n = k\}$, where $0 \leq k \leq n$. We let n and p be fixed.

Lemma 6. $b(k; n, p)$ attains its maximum at $k = m$, where $m = np + \delta$ for some $-q < \delta \leq p$.

Proof. We have

$$\begin{aligned} \frac{b(k; n, p)}{b(k-1; n, p)} &= \frac{\binom{n}{k} p^k q^{n-k}}{\binom{n}{k-1} p^{k-1} q^{n-k}} \\ &= \frac{\frac{n!}{k!(n-k)!} p^k q^{n-k}}{\frac{n!}{(k-1)!(n-(k-1))!} p^{k-1} q^{n-(k-1)}} \\ &= \frac{(n-k+1)p}{kq} \\ &= 1 + \frac{(n+1)p - k}{kq}. \end{aligned}$$

Hence $b(k-1; n, p) < b(k; n, p)$ when $(n+1)p < k$, and $b(k; n, p) < b(k-1; n, p)$ when $(n+1)p > k$. If $(n+1)p$ is an integer, then for $\delta = p$, the function attains its maximum at

$$m = (n+1)p = np + p = np + \delta.$$

Otherwise, it is clear that there is exactly one integer m such that

$$(n+1)p - 1 < m < (n+1)p.$$

In this case, the function attains its maximum at $m = (n+1)p + s$ for some $-1 < s < 0$, and $m = (n+1)p + s = np + p + s = np + \delta$ where $\delta = p + s > p - 1 = -(1-p) = -q$, $\delta = p + s < p + 0 = p$. \square

Let $k > 0$. Show that

$$b(m+k; n, p) = b(m; n, p) \frac{(1-pt_0)(1-pt_1)\cdots(1-pt_{k-1})}{(1+qt_0)(1+qt_1)\cdots(1+qt_{k-1})}, \quad \text{with } t_j = \frac{j+\delta+q}{(n+1)pq}. \quad (3)$$

Proof. We have

$$\begin{aligned}
b(m; n, p) &= \binom{n}{m+k} p^{m+k} q^{n-m-k} \\
&= \frac{n!}{(m+k)!(n-m-k)!} p^{m+k} q^{n-m-k} \\
&= \left(\frac{p^m q^{n-m} n!}{m!(n-m-k)!} \right) \left(\frac{1}{(m+1)(m+2)\cdots(m+k)} \right) \frac{p^k}{q^k} \\
&= \left(\frac{p^m q^{n-m} n!}{m!(n-m)!} \right) \left(\frac{(n-m-k+1)\cdots(n-m-1)(n-m)}{(m+1)(m+2)\cdots(m+k)} \right) \frac{p^k}{q^k} \\
&= b(m; n, p) \frac{(n-m-k+1)\cdots(n-m-1)(n-m)p^k}{(m+1)(m+2)\cdots(m+k)q^k} \\
&= b(m; n, p) \frac{(n-m-k+1)p\cdots(n-m-1)p(n-m)p}{(m+1)q(m+2)q\cdots(m+k)q}.
\end{aligned}$$

For each $j = 0, 1, \dots, k-1$,

$$\begin{aligned}
\frac{p(n-m-j)}{q(m+j+1)} &= \frac{p(n-np-\delta-j)}{q(np+\delta+j+1)} \\
&= \frac{p(nq+q-q-\delta-j)}{q(np+p-p+1+\delta+j)} \\
&= \frac{p((n+1)q-q-\delta-j)}{q(p(n+1)+q+\delta+j)} \\
&= \frac{(n+1)pq\left(1-p\left(\frac{q+\delta+j}{(n+1)pq}\right)\right)}{(n+1)pq\left(1+q\left(\frac{q+\delta+j}{(n+1)pq}\right)\right)}.
\end{aligned}$$

Putting this all together,

$$b(m+k; n, p) = b(m; n, p) \frac{(1-pt_0)(1-pt_1)\cdots(1-pt_{k-1})}{(1+qt_0)(1+qt_1)\cdots(1+qt_{k-1})},$$

where $t_j = \frac{j+\delta+q}{(n+1)pq}$ for all $0 \leq j \leq k-1$. □

Lemma 7. Suppose, in Equation (3), that $t_k < 1/2$. Show that there exists $\theta_k < k^3/(npq)^2$ such that

$$b(m+k; n, p) = b(m; n, p) e^{-(t_0+\cdots+t_{k-1})+\theta_k},$$

using the formula $\ln(1+x) = x + \theta(x)x^2$, where $|\theta(x)| \leq 1$ (Example 12).

Proof. For all $0 \leq j \leq k-1$,

$$\begin{aligned}
\frac{1-pt_j}{1+qt_j} &= \frac{e^{\ln(1-pt_j)}}{e^{\ln(1+qt_j)}} \\
&= \frac{e^{-pt_j+\theta(-pt_j)p^2t_j^2}}{e^{qt_j+\theta(qt_j)q^2t_j^2}} \\
&= e^{-(p+q)t_j+t_j^2(\theta(-pt_j)p^2-\theta(qt_j)q^2)} \\
&= e^{-t_j+t_j^2(\theta(-pt_j)p^2-\theta(qt_j)q^2)}
\end{aligned}$$

where $|\theta(-pt_j)| \leq 1$ and $|\theta(qt_j)| \leq 1$. This is justified because $|-pt_j| < 1/2 < 1$ and $qt_j < 1/2 < 1$. Thus

$$\begin{aligned}
b(m+k; n, p) &= b(m; n, p) \frac{(1-pt_0)(1-pt_1)\cdots(1-pt_{k-1})}{(1+qt_0)(1+qt_1)\cdots(1+qt_{k-1})} \\
&= b(m; n, p) e^{-(t_0+t_1+\cdots+t_{k-1})+\sum_{j=0}^{k-1} t_j^2(\theta(-pt_j)p^2-\theta(qt_j)q^2)} \\
&= b(m; n, p) e^{-(t_0+t_1+\cdots+t_{k-1})+\theta_k}
\end{aligned}$$

if we let $\theta_k = \sum_{j=0}^{k-1} t_j^2(\theta(-pt_j)p^2 - \theta(qt_j)q^2)$. Notice

1. $-1 \leq \theta(-pt_j) \leq 1 \Rightarrow p^2\theta(-pt_j) \leq p^2$,
2. $-1 \leq \theta(qt_j) \leq 1 \Rightarrow 1 \geq -\theta(qt_j) \geq -1 \Rightarrow -q^2\theta(qt_j) \leq q^2$, and
3. $p^2 + q^2 \leq 1 \Leftrightarrow p^2 + q^2 + 2pq \leq 2pq + 1 \Leftrightarrow 1 = (p+q)^2 \leq 2pq + 1$.

Also, it is easy to see that for all j , $t_j^2 \leq t_{k-1}^2$. Hence

$$\begin{aligned}
\theta_k &\leq \sum_{j=0}^{k-1} t_j^2(p^2 + q^2) \\
&\leq \sum_{j=0}^{k-1} t_j^2 \\
&\leq kt_{k-1}^2 \\
&= k \left(\frac{k-1+\delta+q}{(n+1)pq} \right)^2 \\
&\leq k \left(\frac{k-1+p+q}{(n+1)pq} \right)^2 \\
&= \frac{k^3}{((n+1)pq)^2} < \frac{k^3}{(npq)^2}.
\end{aligned}$$

□

Lemma 8. Show that

$$b(m+k; n, p) = b(m; n, p) e^{-k^2/(2npq)+\rho_k}, \quad \text{with } |\rho_k| < \frac{k^3}{(npq)^2} + \frac{2k}{npq}.$$

Proof. We show

$$\left| \frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq} - \frac{\frac{k^2}{2}}{npq} \right| < \frac{2k}{npq}.$$

First, If the first term greater than the second, then

$$\begin{aligned} \frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq} - \frac{\frac{k^2}{2}}{npq} &= \frac{\frac{nk(k-1)}{2} + nk(\delta + q) - \frac{(n+1)k^2}{2}}{n(n+1)pq} \\ &< \frac{\frac{nk^2}{2} - \frac{nk}{2} + nk - \frac{nk^2}{2} - \frac{k^2}{2}}{n(n+1)pq} \\ &= \frac{\frac{1}{2}(nk - k^2)}{n(n+1)pq} \\ &< \frac{k(n-k)}{n(n+1)pq} \\ &< \frac{k(n+1)}{n(n+1)pq} \\ &= \frac{k}{npq} < \frac{2k}{npq}. \end{aligned}$$

Otherwise,

$$\begin{aligned} \frac{\frac{k^2}{2}}{npq} - \frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq} &= \frac{\frac{nk^2}{2} + \frac{k^2}{2} - \frac{nk(k-1)}{2} - nk(\delta + q)}{n(n+1)pq} \\ &< \frac{\frac{1}{2}(k^2 + kn)}{n(n+1)pq} \\ &< \frac{k(k+n)}{n(n+1)pq} \\ &< \frac{2nk}{n(n+1)pq} \\ &= \frac{2k}{(n+1)pq} < \frac{2k}{npq}. \end{aligned}$$

Let

$$\rho_k = \theta_k + \frac{k^2}{2npq} - \frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq}.$$

By what we showed above,

$$\begin{aligned} b(m+k; n, p) &= b(m; n, p)e^{-(t_0 + \dots + t_{k-1}) + \theta_k} \\ &= b(m; n, p)e^{-\frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq} + \theta_k} \\ &= b(m; n, p)e^{-\frac{k^2}{2npq} + \frac{k^2}{2npq} - \frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq} + \theta_k} \\ &= b(m; n, p)e^{-k^2/(2npq) + \rho_k} \end{aligned}$$

where

$$\begin{aligned} |\rho_k| &\leq \theta_k + \left| \frac{k^2}{2npq} - \frac{\frac{k(k-1)}{2} + k(\delta + q)}{(n+1)pq} \right| \\ &< \frac{k^3}{(npq)^2} + \frac{2k}{npq}. \end{aligned}$$

□

Lemma 9. For any $|\delta| < 1$ and $n \geq 2$, there exists $|\theta| < 1$ such that

$$\left(1 + \frac{\delta}{n}\right)^n = e^{\delta + \frac{\theta}{n}}.$$

Proof. Consider $n \ln(1 + \delta/n)$. Since $|\delta| < 1$ and $n \geq 2$, $|x| = |\delta/n| < 1/2$, and by Example 12, there exists a θ_x such that $|\theta_x| \leq 1$ and

$$n \ln(1 + x) = n(x + \theta_x x^2) = \delta + \frac{\theta_x \delta^2}{n}.$$

Since $|\theta_x| \leq 1$ and $|\delta| < 1 \Rightarrow \delta^2 < 1$, $|\theta_x \delta^2| < 1$, so choose $\theta = \theta_x \delta^2$. Then

$$\left(1 + \frac{\delta}{n}\right)^n = e^{n \ln(1 + \frac{\delta}{n})} = e^{\delta + \frac{\theta}{n}}.$$

□

Lemma 10. Using the Stirling formula, we can obtain an estimate of θ_n for which we have

$$\sqrt{2\pi npq} \cdot b(m; n, p) = \sqrt{\frac{np}{m}} \cdot \sqrt{\frac{nq}{n-m}} \cdot \left(\frac{np}{m}\right)^m \cdot \left(\frac{nq}{n-m}\right)^{n-m} \cdot e^{\theta_n}.$$

Proof. Let $n \geq 2$. Then

$$\begin{aligned} \sqrt{2\pi npq} \cdot b(m; n, p) &= \sqrt{2\pi npq} \cdot \frac{n!}{m!(n-m)!} \cdot p^m q^{n-m} \\ &= \sqrt{2\pi npq} \cdot \frac{\sqrt{2\pi n}(n/e)^n e^{\theta_n}}{\sqrt{2\pi m}(m/e)^m e^{\theta_m} \sqrt{2\pi(n-m)}((n-m)/e)^{n-m} e^{\theta_{n-m}}} \cdot p^m q^{n-m} \\ &= \sqrt{2\pi npq} \cdot \frac{\sqrt{2\pi n}}{\sqrt{2\pi m} \sqrt{2\pi(n-m)}} \cdot \frac{n^m n^{n-m}}{m^m (n-m)^{n-m}} \cdot e^{\theta_n - \theta_m - \theta_{n-m}} \cdot p^m q^{n-m} \\ &= \sqrt{\frac{np}{m}} \cdot \sqrt{\frac{nq}{n-m}} \cdot \left(\frac{np}{m}\right)^m \cdot \left(\frac{nq}{n-m}\right)^{n-m} \cdot e^{\theta_n - \theta_m - \theta_{n-m}} \end{aligned}$$

where $0 < \theta_n < \frac{1}{12n}$, $0 < \theta_m < \frac{1}{12m}$, and $0 < \theta_{n-m} < \frac{1}{12(n-m)}$ by Stirling's Formula. Thus,

$$-\left(\frac{1}{12n} + \frac{1}{12(n-m)}\right) < \theta_n - \theta_m - \theta_{n-m} < \frac{1}{12n} < \frac{1}{12n} + \frac{1}{12(n-m)},$$

so $|\theta_n - \theta_m - \theta_{n-m}| < \frac{1}{12} + \frac{1}{12(n-m)}$.

□

Lemma 11. $\lim_n \sqrt{2\pi npq} \cdot b(m; n, p) = 1$.

Proof. For any $n \geq 2$, for the $-q < \delta \leq p$ such that $m = np + \delta$, denote $\delta_n = \delta$. We have

$$\begin{aligned} \lim_{n \rightarrow \infty} n - m &= \lim_{n \rightarrow \infty} n - (np + \delta_n) \\ &\geq \lim_{n \rightarrow \infty} n - (np - q) \\ &= \lim_{n \rightarrow \infty} n(1 - p) + q = \infty, \end{aligned}$$

so $n - m \rightarrow \infty$ as $n \rightarrow \infty$. For any $n \geq 2$,

$$\begin{aligned} \left(\frac{np}{m}\right)^m &= \left(\frac{m - \delta_n}{m}\right)^m \\ &= \left(1 + \frac{-\delta_n}{m}\right)^m \end{aligned}$$

and

$$\begin{aligned} \left(\frac{nq}{n - m}\right)^{n - m} &= \left(\frac{n(1 - p)}{n - m}\right)^{n - m} \\ &= \left(\frac{n - np}{n - m}\right)^{n - m} \\ &= \left(\frac{n - (m - \delta_n)}{n - m}\right)^{n - m} \\ &= \left(\frac{n - m + \delta_n}{n - m}\right)^{n - m} \\ &= \left(1 + \frac{\delta_n}{n - m}\right)^{n - m}. \end{aligned}$$

By previous lemmas, there are $|\Theta_{n,m}| < 1, |\Theta_{n,n-m}| < 1$ such that

$$\left(1 + \frac{-\delta_n}{m}\right)^m = e^{-\delta_n + \frac{\Theta_{n,m}}{n}}$$

and

$$\left(1 + \frac{\delta_n}{n - m}\right)^{n - m} = e^{\delta_n + \frac{\Theta_{n,n-m}}{n - m}}.$$

Therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} \sqrt{2\pi npq} \cdot b(m; n, p) &= \lim_{n \rightarrow \infty} \sqrt{\frac{np}{m}} \cdot \sqrt{\frac{nq}{n - m}} \cdot \left(\frac{np}{m}\right)^m \cdot \left(\frac{nq}{n - m}\right)^{n - m} \cdot e^{\theta_n} \\ &= \lim_{n \rightarrow \infty} \sqrt{\frac{np}{np + \delta_n}} \cdot \sqrt{\frac{nq}{nq - \delta_n}} \cdot e^{-\delta_n + \frac{\Theta_{n,m}}{n}} \cdot e^{\delta_n + \frac{\Theta_{n,n-m}}{n - m}} \cdot e^{\theta_n} \\ &= \lim_{n \rightarrow \infty} \sqrt{\frac{1}{1 + \frac{\delta_n}{np}}} \cdot \sqrt{\frac{1}{1 - \frac{\delta_n}{nq}}} \cdot e^{\frac{\Theta_{n,m}}{n} + \frac{\Theta_{n,n-m}}{n - m}} \cdot e^{\theta_n} = 1, \end{aligned}$$

where $e^{\theta_n} \rightarrow 1$ as $n \rightarrow \infty$. □

Theorem 21. Let k_n be a sequence of positive real numbers such that $\lim_n k_n^3/n^2 = 0$ (for instance, $k_n = \sqrt{n}$). Combine what we developed in Part (I) and (II) to show that for any $\epsilon > 0$ there exists $N > 0$ such that for any integers $n > N$ and $0 \leq k \leq k_n$ we have

$$\left| b(m+k; n, p) \cdot \sqrt{2\pi npq} \cdot e^{k^2/(2npq)} - 1 \right| < \epsilon.$$

Proof. Since we only care about small values of ϵ , we might as well take $\epsilon < 1$. For any n , for any $0 \leq k \leq n$,

$$\begin{aligned} \left| b(m+k; n, p) \cdot \sqrt{2\pi npq} \cdot e^{k^2/(2npq)} - 1 \right| &= \left| b(m; n, p) \cdot \sqrt{2\pi npq} \cdot e^{\rho_k} - 1 \right| \\ &\leq b(m; n, p) \sqrt{2\pi npq} |e^{\rho_k} - 1| + \left| b(m; n, p) \sqrt{2\pi npq} - 1 \right|, \end{aligned}$$

where the equality follows from one of the previous lemmas and the inequality is gotten by adding and subtracting $b(m; n, p) \sqrt{2\pi npq}$ in the middle and applying the Triangle Inequality. Let's make a few observations.

- There exists an $N_1 > 0$ such that for any $n > N_1$, $|b(m; n, p) \sqrt{2\pi npq} - 1| < \epsilon/2$.
- Since $\lim_n k_n^3/n^2 = 0$, $\lim_n \frac{k_n^3}{(npq)^2} + \frac{2k_n}{npq} = 0$ as well. Thus, there exists $N_2 > 0$ such that for any $n > N_2$,

$$\frac{k_n^3}{(npq)^2} + \frac{2k_n}{npq} < \frac{\epsilon}{8}.$$

- We have

$$\sqrt{2\pi npq} \cdot b(m; n, p) < 1 + \frac{\epsilon}{2} < 2$$

by the first bullet and because $\epsilon < 1$.

- Let $N = \max\{N_1, N_2\}$. For any $n > N$, for any $0 \leq k \leq k_n$,

$$|\rho_k| < \frac{k^3}{(npq)^2} + \frac{2k}{npq} \leq \frac{k_n^3}{(npq)^2} + \frac{2k_n}{npq} < \frac{\epsilon}{8} < 1.$$

Hence by Example 13,

$$|e^{\rho_k} - 1| \leq 2|\rho_k| < 2 \cdot \frac{\epsilon}{8} = \frac{\epsilon}{4}.$$

We now put this all together. Choose $N = \max\{N_2, N_2\}$. Let $n > N$ and $0 \leq k \leq k_n$ be given. Then by our work above,

$$\begin{aligned} \left| b(m+k; n, p) \cdot \sqrt{2\pi npq} \cdot e^{k^2/(2npq)} - 1 \right| &= \left| b(m; n, p) \cdot \sqrt{2\pi npq} \cdot e^{\rho_k} - 1 \right| \\ &\leq b(m; n, p) \sqrt{2\pi npq} |e^{\rho_k} - 1| + \left| b(m; n, p) \sqrt{2\pi npq} - 1 \right| \\ &< 2 \cdot \frac{\epsilon}{4} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

□

4.4 Central Limit Theorem

Theorem 22 (Central Limit Theorem). let $a, b \in \mathbb{R}$ such that $a < b$ and $(\Omega, \mathcal{F}, \mathbb{P})$ be the coin toss model with parameter $0 < p < 1$. Then:

$$\lim_{n \rightarrow \infty} \mathbb{P}\left\{\omega \in \Omega : a < \frac{S_n(\omega) - np}{\sqrt{npq}} < b\right\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt$$

Note that the right-hand side of the above expression is simply the integration of the standard normal density from a to b . This means the new random variable $S^* = \frac{S_n - \mathbb{E}S_n}{\sqrt{\text{Var}(S_n)}} \sim \mathcal{N}(0, 1)$. Let us postpone the proof for a little bit, and see how the theorem can be applied.

Suppose, again, we throw a die 12,000 times and ask for the probability that the number of times we get a 6 is between 1800 and 2100. The actual probability would be calculated as follows:

$$\sum_{\ell=1800}^{2100} b(\ell; 12,000, 1/6)$$

Clearly, this number would be a pain to compute by hand or even using a calculator, and you shouldn't do it, unless you like pain. Instead, we'll use the Central Limit Theorem to obtain an approximation. First, note that $\sqrt{npq} \approx 40.825$. Also, $a = \frac{1800-2000}{40.825} = -4.898$ and $b = \frac{2100-2000}{40.825} = 2.449$. Hence, using the Central Limit Theorem, we obtain:

$$\begin{aligned} \mathbb{P}\{1800 < S_n < 2100\} &= \mathbb{P}\left\{-4.898 < \frac{S_n - 2000}{40.825} < 2.449\right\} \\ &= \int_{-4.898}^{2.449} \varphi(t) dt \\ &= \int_{-\infty}^{2.449} \varphi(t) dt - \int_{-\infty}^{-4.898} \varphi(t) dt \\ &= \Phi(2.449) - \Phi(-4.898) \\ &> 0.9929 - 0.0014 \\ &> 0.99 \end{aligned}$$

Note that we used the fact that $\Phi(-x) = 1 - \Phi(x)$ for all x , which can be easily deduced from noting that the bell curve is symmetric around 0. Note that the above tells us that with 99% chance, the number of 6's will belong to the specified range! (This is somewhat surprising, but is in fact quite natural, because of how the normal curve behaves – values are extremely concentrated close to the mean).

Example 52. Another example of applying the Central Limit Theorem would be in the setting of the one-dimensional random walk. That is, suppose that a particle is moving in one dimension in a discrete manner (that is, we start at 0 and can only move one integer to the right or one integer to the left in one step); the probability of moving right is $1/2$ and probability of moving left is $1/2$ as well. We're interested in determining the position

of the particle after n steps. Clearly, it has to be in the interval $[-n, n]$. Can we analyze the behavior better? Yes – by translating this model into the infinite coin toss model. That is, we can consider the infinite coin toss model $(\Omega, \mathcal{F}, \mathbb{P})$ with $p = \frac{1}{2}$. A step right can be treated as 1 (or success) and a step left can be treated as 0 (or failure). Let $L_n(\omega)$ denote the *position* of a particle at epoch n .

Theorem 23.

$$\lim_{n \rightarrow \infty} \mathbb{P}\left\{\omega : a < \frac{L_n(\omega)}{\sqrt{n}} < b\right\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Before we prove this, note that from the z -table we can deduce that $\int_{-2}^2 \varphi dt > 95\%$. This means that with high probability the particle will be in the interval $[-2\sqrt{n}, 2\sqrt{n}]$. Now let us prove 23.

Proof. First, note that:

$$L_n(\omega) = S_n(\omega) - (n - S_n(\omega)) = 2S_n(\omega) - n$$

since L_n is obtained by subtracting the number of left steps from the number of right steps. Now notice that:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}\left\{\omega : a < \frac{L_n(\omega)}{\sqrt{n}} < b\right\} &= \lim_{n \rightarrow \infty} \mathbb{P}\left\{\omega : a < \frac{S_n(\omega) - \frac{n}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} < b\right\} \\ &= \lim_{n \rightarrow \infty} \mathbb{P}\left\{\omega : a < \frac{S_n(\omega) - np}{\sqrt{npq}} < b\right\} \\ &= \int_a^b \varphi(t) dt \quad \text{CLT} \end{aligned}$$

□

It is now time to prove the Central Limit Theorem. Before giving a direct proof, we need to do some analysis-style scratch work. Let $\varepsilon > 0$ be given. Recall that what we want to show is that for any real numbers $a < b$ in the infinite coin toss model, we have:

$$\underbrace{\lim_{n \rightarrow \infty} \mathbb{P}\left\{\omega \in \Omega : a < \frac{S_n(\omega) - np}{\sqrt{npq}} < b\right\}}_{\text{LHS}} = \underbrace{\frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt}_{\text{RHS}}$$

For simplicity of notation, denote the left-hand side of the expression above (without limit) by LHS and the right-hand side by RHS. Equivalently, we must show that for any $\varepsilon > 0$ there is some $N \in \mathbb{N}$ such that for all $n > N$, we have $|\text{LHS} - \text{RHS}| < \varepsilon$. The following lemma, which you will prove on your next homework, will be useful:

Lemma 12.

$$\lim_{n \rightarrow \infty} \underbrace{\mathbb{P}\left\{\omega \in \Omega : a < \frac{S_n(\omega) - m}{\sqrt{npq}} < b\right\}}_{\text{NLHS}} = \lim_{n \rightarrow \infty} \text{LHS}$$

where $m \in \mathbb{Z} \cap (np - q, np + p]$.

Let $\text{NLHS} = \mathbb{P}\left\{\omega \in \Omega : a < \frac{S_n(\omega) - m}{\sqrt{npq}} < b\right\}$. By Lemma 12, it suffices to show that:

$$|\text{NLHS} - \text{RHS}| < \varepsilon$$

By triangle inequality:

$$\left| \text{NLHS} - \sum_{\frac{a}{h} < k < \frac{b}{h}} h\varphi(hk) \right| + \left| \sum_{\frac{a}{h} < k < \frac{b}{h}} h\varphi(hk) - \text{RHS} \right| < \varepsilon \quad (4)$$

where h is defined as in the remark in the beginning of the section. So, now, it suffices to show that each of the terms is less than $\varepsilon/2$. We claim that the right term is easy to deal with because of the Reimann sum approximation of integrals, so we will do so in the actual proof later. Let us see how we shall deal with the left absolute value term. Note:

$$\begin{aligned} \text{NLHS} &= \mathbb{P}\left\{\omega : \frac{a}{h} < S_n(\omega) - m < \frac{b}{h}\right\} \\ &= \sum_{\substack{m + \frac{a}{h} < \ell < m + \frac{b}{h} \\ \ell \in \mathbb{Z}}} b(\ell; n, p) \\ &= \sum_{\substack{\frac{a}{h} < k < \frac{b}{h} \\ k \in \mathbb{Z}}} b(m + k; n, p) \\ &\approx \sum_{\substack{\frac{a}{h} < k < \frac{b}{h} \\ k \in \mathbb{Z}}} h\varphi(hk) + h\varepsilon \quad (*) \\ &= \left[\sum_{\substack{\frac{a}{h} < k < \frac{b}{h} \\ k \in \mathbb{Z}}} h\varphi(hk) \right] + h \cdot \varepsilon \cdot \frac{(b - a)}{h} \end{aligned}$$

where $(*)$ follows from the Local Limit Theorem, i.e. we get some N such that for all $n > N$:

$$|b(m + k; n, p) - h\varphi(hk)| < h\delta$$

where δ is the number that makes our previous reasoning work. Note that LLT works for any δ , so, in our case, in order to make the left expression in (2) less than $\varepsilon/2$, we want to let $\delta = \frac{\varepsilon}{2(b-a)}$. OK, seems like we're now ready to write down the proof!

Actual Proof of CLT. Let $\varepsilon > 0$ be given and let $h = \frac{1}{\sqrt{npq}}$. Using the same notation as above, applying the Lemma 12, it suffices to show that there exists $N > 0$ such that for any $n > N$:

$$|\text{NLHS} - \text{RHS}| < \varepsilon$$

By triangle inequality, it suffices to show:

$$\left| \text{NLHS} - \sum_{\frac{a}{h} < k < \frac{b}{h}} h\varphi(hk) \right| + \left| \sum_{\frac{a}{h} < k < \frac{b}{h}} h\varphi(hk) - \text{RHS} \right| < \varepsilon \quad (5)$$

So, we need to show that each of the terms above is less than $\varepsilon/2$. notice that:

$$\begin{aligned} \text{NLHS} &= \mathbb{P}\left\{\omega : \frac{a}{h} < S_n(\omega) - m < \frac{b}{h}\right\} \\ &= \sum_{k \in \mathbb{Z} \cap (\frac{a}{h}, \frac{b}{h})} b(m+k; n, p) \end{aligned}$$

By the Local Limit Theorem, letting $k_n = C\sqrt{n}$ where $C = \max(|a|, |b|)$, we get that for the chosen ε , there is some $N_1 \in \mathbb{N}$ such that for all $n > N_1$ and all $|k| < C\sqrt{n}$, we have:

$$|b(m+k; n, p) - h\varphi(kh)| < \frac{\varepsilon}{2(b-a)} \cdot h$$

Then, we get that:

$$\left| \text{NLHS} - \sum_{\frac{a}{h} < k < \frac{b}{h}} h\varphi(hk) \right| < \frac{b-a}{h} \cdot \frac{\varepsilon}{2(b-a)} \cdot h = \varepsilon/2 \quad (6)$$

On the other hand, using the definition of the Riemann integral, for the chosen ε , there is some $h_0 > 0$ such that for all $0 < h' < h_0$, we have:

$$\left| \sum_{\frac{a}{h} < k < \frac{b}{h}} h'\varphi(h'k) - \text{RHS} \right| < \varepsilon/2$$

Let $N_2 = \left(\frac{1}{h_0}\right)^2 \cdot \frac{1}{pq}$. Then, for all $n > N_2$, we have that $h = \frac{1}{\sqrt{npq}} < h_0$ and hence:

$$\left| \sum_{\frac{a}{h} < k < \frac{b}{h}} h\varphi(hk) - \text{RHS} \right| < \varepsilon/2 \quad (7)$$

Letting $N = \max\{N_1, N_2\}$ and combining (4) and (5), we get that for any $n > N$:

$$|\text{NLHS} - \text{RHS}| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

and this concludes the proof. \square

5 Markov Chains (Optional)

Roughly speaking, a Markov chain is a probabilistic model describing a sequence of events in which one can make predictions about the future based on the present state, without knowing the full history. In fact, you've already seen concrete examples of Markov chains

in your homework, where you had to find the formula for p_n , given a recursive definition. Recall the example from Homework 5 where we flipped a coin n times. In the first flip, the probability of getting a head was c and from the second flip on, each flip had probability p of showing the same face as the previous one did. So, to predict the outcome in the n th flip, we only needed to know what happened in the $(n-1)$ st.

Definition 36. By a *finite Markov model* of length n , we mean a probability space $(\Omega, 2^\Omega, \mathbb{P})$ defined by:

1. a finite set X , referred to as *state space*;
2. non-negative functions $p_0 : X \rightarrow \mathbb{R}$ and $p_i : X \times X \rightarrow \mathbb{R}$, where $i = 1, \dots, n$ such that for $\Omega = \{\omega = (x_0, \dots, x_n) : x_i \in X\}$, we have $\mathbb{P}(\{\omega\}) = p_0(x_0)p_1(x_0x_1)\dots p_n(x_{n-1}x_n)$ and moreover $\sum_{x \in X} p_0(x) = 1$ and $\sum_{y \in X} p_i(x, y) = 1$ for all $x \in X$, $i = 1, \dots, n$.

We call p_0 the initial probability and each p_i the i th transition probability. If moreover $p_1 = p_2 = \dots = p_n$, the model is called *homogeneous*.

Definition 2. In a Markov model $(\Omega, 2^\Omega, \mathbb{P})$ above, define $\xi_i(\omega) = x_i$. Then, $(\xi_0, \xi_1, \dots, \xi_n)$ is called a *Markov chain*.

let us consider some examples now. We will view the coin toss experiment from Homework 5 as a one-dimensional random walk where heads = step right and tails = step left. So, $X = \{H, T\}$ and $\Omega = \{(x_0, \dots, x_n) : x_i \in X\}$. Then, $p_0(H) = c$ and $p_0(T) = 1 - c$. We can represent it as a vector $\begin{pmatrix} c & 1-c \end{pmatrix}$. Similarly, we can observe that the assumptions of the problem translate to having $p_i(H, H) = p$, $p_i(T, T) = p$, $p_i(H, T) = q$ and $p_i(T, H) = q$ for all $i = 1, \dots, n$. This can be also represented by what we refer to as *stochastic matrix*:

$$\begin{pmatrix} p_i(H, H) & p_i(H, T) \\ p_i(T, H) & p_i(T, T) \end{pmatrix} = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$$

Note that the sum in each row is 1, which is exactly the condition imposed on p_i in the definition 36. Hence, by the same definition, we can compute the probability of obtaining heads (and tails) at the time 1:

$$\begin{pmatrix} c & 1-c \end{pmatrix} \cdot \begin{pmatrix} p & q \\ q & p \end{pmatrix} = \begin{pmatrix} pc + q(1-c) & qc + p(1-c) \end{pmatrix}$$

Similarly, we can obtain respective probabilities for any time n . However, while calculating the n th power of a matrix may be a complicated task, a useful fact to remember would be that for any diagonalizable matrix $A = P\Lambda P^{-1}$, we have $A^n = P\Lambda^n P^{-1}$. Of course, not all square matrices are diagonalizable, but the one in our example is, so we'll use this fact. For time n , we obtain:

$$\begin{aligned} \begin{pmatrix} c & 1-c \end{pmatrix} \cdot \begin{pmatrix} p & q \\ q & p \end{pmatrix}^n &= \begin{pmatrix} c & 1-c \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (2p-1)^n \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \\ &= \begin{pmatrix} c & 1-c \end{pmatrix} \begin{pmatrix} \frac{1}{2} + \frac{1}{2}(2p-1)^n & \frac{1}{2} - \frac{1}{2}(2p-1)^n \\ \frac{1}{2} - \frac{1}{2}(2p-1)^n & \frac{1}{2} + \frac{1}{2}(2p-1)^n \end{pmatrix} \end{aligned}$$

One can check that the above goes to $\left(\frac{1}{2} \quad \frac{1}{2}\right)$ as $n \rightarrow \infty$. Plugging this in, we obtain:

$$(c \quad 1-c) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

We call this the equilibrium state of the model.