MATH 6605: Probability Theory

Joe Tran

Winter 2025

Preface

These are the first edition of these lecture notes for MATH 6605 (Probability Theory). Consequently, there may be several typographical errors, missing exposition on necessary background, and more advanced topics for which there will not be time in class to cover. Future iterations of these notes will hopefully be fairly self-contained provided one has the necessary background. If you come across any typos, errors, omissions, or unclear expositions, please feel free to contact me so that I may continually improve these notes.

Contents

Chapter 1

The Need For Measure Theory

This introductory section is directed primarily to those who have some familiarity with undergraduate level probability theory, and who may be unclear as to why it is necessary to introduce measure theory and other mathematical difficulties in order to study probability theory in a rigorous manner.

We attempt to illustrate the limitations of undergraduate-level probability in two different ways: the restrictions on the kinds of random variables it allows, and the question of what sets can probabilities defined on them.

1.1 Various Kinds of Random Variables

The reader with undergraduate level probability will be comfortable with the statement like, "Let X be a random variable which has the Poisson(5) distribution." The reader will know that this means that X takes as its value a "random" nonnegative integer such that the integer $k \geq 0$ chosen with probability

$$P(X = k) = \frac{e^{-5}5^k}{k!}$$

The expected value of, say, X^2, can then be computed as

$$\mathcal{E}(X^2) = \sum_{k=0}^{\infty} x^2 \frac{e^{-5}5^k}{k!}$$

X is an example of a *discrete random variable.*

Similarly, the reader will be familiar with the statement like, "Let Y be a random variable which has the Normal$(0, 1)$ distribution." This means that the probability that Y lies between two real numbers $a < b$ is given by the integral

$$P(a \leq Y \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

The expected value, of say, Y^2, can then by computed as

$$\mathcal{E}(Y^2) = \int_{-\infty}^{+\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$$

Y is an example of an *absolutely continuous random variable*, or simply, *continuous random variable.*

But now suppose we introduce a new random variable, Z as follows: We let X and Y be as above, and then flip an (independent) fair coin. If the coin comes up heads, then we set $Z = X$, while if it comes up tails, we set $Z = Y$. In symbols,

$$P(Z = X) = P(Z = Y) = \frac{1}{2}$$

Then what sort of random variable is Z? It is not discrete since it can take on an uncountable number of different values, and it is not absolutely continuous for certain values of z, in particular, when z is a nonnegative integer, we have $P(Z = z) > 0$. So how can we study the random variable Z, and how can we compute the expect value of Z^2.

The correct response to this question, is that the division of random variables into discrete versus absolutely continuous is artificial. Instead, measure theory allows us to give a common definition of expected value, which applies equally well to discrete random variables to a continuous random variables to combinations of them, and other kinds of random variables not yet imagined.

1.2 The Uniform Distribution and Non-Measurable Sets

In undergraduate-level probability, continuous random variables are often studied in detail. However, a closer examination suggests that perhaps such random variables are not completely understood at all.

To take the simplest case, suppose that X is a random variable which has the uniform distribution on the unit interval $[0, 1]$. In symbols, $X \sim \text{Uniform}[0, 1]$. What does this precisely mean.

This means that

$$\begin{aligned} P(0 \le X \le 1) &= 1 \\ P(0 \le X \le 1/2) &= \frac{1}{2} \\ P(3/4 \le X \le 7/8) &= \frac{1}{8} \end{aligned}$$

and so on. In general, this means

$$P(a \le X \le b) = b - a$$

whenever $0 \le a \le b \le 1$ with the same formula holding if $\le$ is replaced with $<$. Indeed,

$$P([a, b]) = P((a, b]) = P([a, b)) = P((a, b)) = b - a \tag{1.1}$$

for all $0 \le a \le b \le 1$. In other words, the probability that X lies in any interval contained in $[0, 1]$ is simply the length of the interval. Similarly, this means

$$P([1/4, 1/2] \cup [2/3, 5/6]) = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

and in general, if A and B are disjoint subsets of $[0, 1]$, then

$$P(A \cup B) = P(A) + P(B) \tag{1.2}$$

Equation (1.2) is called *finite additivity.*

Similarly, to allow for countable operations (such as limits, which are extremely important in probability theory), we could extend (1.2) to the case of a countably infinite number of disjoint subsets: if $(A_n)_{n=1}^{\infty}$ is a collection of disjoint subsets of $[0,1]$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \tag{1.3}$$

Equation (1.3) is called *countable additivity.*

Note that we do not extend (1.3) to *uncountable additivity.* Indeed, if we did, then we would expect that

$$P([0,1]) = \sum_{x\in[0,1]} P(\{x\})$$

which is absurd, since the left side is equal to 1, but the right side is equal to 0, but note there is no contradiction to (1.3) since $[0,1]$ is not even countable.

Similarly, to reflect the fact that X is "uniform" on the interval $[0,1]$, the probability that X lies in some subset should be unaffected by "shifting" (with wrap-around) the subset by a fixed amount. That is, if for each subset $A \subset [0,1]$, we define the *r-shift of A* by

$$A \oplus r = \{a+r : a \in A, a+r \leq 1\} \cup \{a+r-1 : a \in A, a+r > 1\} \tag{1.4}$$

then we should have

$$P(A \oplus r) = P(A) \tag{1.5}$$

for all $0 \leq r \leq 1$.

Now suppose we ask, what is the probability that X is rational. What is the probability that X^n is rational for some positive integer n? What is the probability that X is *algebraic*, i.e. the solution to some polynomial with integer coefficients? Can we compute these things? More fundamentally, are all probabilities such as these necessarily even *defined*? That is, does $P(A)$ (i.e. the probability that X lies in the subset A) even make sense for every possible subset $A \subset [0,1]$?

The answer to the last question is no, as the following proposition shows.

Proposition 1.2.1

There does not exist a definition of $P(A)$, defined for all subsets $A \subset [0,1]$ satisfying (1.1), (1.3) and (1.5).

Proof.

Suppose for a contradiction that $P(A)$ could be defined for any subset $A \subset [0,1]$. Define an equivalence relation "$\sim$" on $[0,1]$ as follows: $x \sim y$ if and only if $y - x \in \mathbb{Q}$. This relation partitions the interval $[0,1]$ into a disjoint union of equivalence classes. This relation partitions the interval $[0,1]$ into a disjoint union of equivalence classes. Let H be a subset of $[0,1]$ consisting of precisely one element from each equivalence class (such H exists by the Axiom of Choice). For definiteness, assume that $0 \notin H$ (say, if $0 \in H$, replace by $1/2$).

Now since H contains an element of each equivalence class, we see that each point in $(0,1]$ is contained in the union $\bigcup_{r\in[0,1)\cap\mathbb{Q}}(H \oplus r)$ of shifts of H. Furthermore, since H contains just *one* point from each equivalence class, we see that the sets $H \oplus r$ for $r \in [0,1)$ rational, are all disjoint.

But then by (1.3), we have

$$P((0,1]) = \sum_{r \in [0,1) \cap \mathbb{Q}} P(H \oplus r)$$

Shift-invariance (1.5) implies that $P(H \oplus r) = P(H)$ and so

$$1 = P((0,1]) = \sum_{r \in [0,1) \cap \mathbb{Q}} P(H)$$

which is absurd, since a countably infinite sum of the same quantity repeated can only equal to 0, ∞, or $-\infty$, but it can never equal to 1. ■

This proposition says that if we want our probabilities to satisfy reasonable properties, then we *cannot* define them for all possible subsets of $[0,1]$. Rather, we must *restrict* their definition to certain "measurable" sets. This is the motivation for the next chapter.

Remark 1.2.2

The existence of problematic sets like H above turns out to be equivalent to the axiom of choice. In particular, we can never define such sets explicitly—only implicitly by the axiom of choice as in the proof in Proposition 1.2.1.

Chapter 2

Probability Triples

In this section, we consider the probability triples and how to construct them. In light of the previous chapter, we see that to study probability theory properly, it will be necessary to keep track of which subsets A have a probability $P(A)$ defined for them.

2.1 Basic Definitions

Definition 2.1.1

We define a *probability triple* or (*probability*) *measure space* or *probability space* to be a triple $(\Omega, \mathcal{A}, P)$, where Ω is the *sample space* that is any nonempty set, $\mathbb{F}$ is a σ*-algebra* which is a collection of subsets of Ω, containing Ω itself and the empty set $\emptyset$, and closed under the formation of complements and countable unions and countable intersections, and the *probability measure* P is a mapping from $\mathcal{A}$ to $[0, 1]$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$, such a P is countably additive as in (1.3)

This definition will be in constant use throughout the text. Furthermore, it contains a number of subtle points. Thus, we pause to make a few observations.

The σ-algebra $\mathcal{A}$ is the collection of all *events* or *measurable sets*. These are the subsets $A \subset \Omega$ for which $P(A)$ is well-defined. From Proposition 1.2.1 that in general $\mathcal{A}$ may not contain all subsets of Ω, though we still expect it to contain most of the subsets that come up naturally.

To say that $\mathcal{A}$ is closed under the formation of complements and countable unions and countable intersections means

(i) For all $A \in \mathcal{A}$, if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

(ii) For any countable collection $(A_n)_{n=1}^{\infty}$ of Ω, if for all $n \in \mathbb{N}$, $A_n \in \mathcal{A}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

(iii) For any countable collection $(A_n)_{n=1}^{\infty}$ of ω, if for all $n \in \mathbb{N}$, $A_n \in \mathcal{A}$, then $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$.

Like for countable additivity, the reason why we require $\mathcal{A}$ to be closed under countable operations is to allow for taking limits, etc., when studying probability theory. Also, like for additivity, we cannot extend the definition to require that $\mathcal{A}$ is closed under uncountable unions; in this case, for the example in Section 1.2, $\mathcal{A}$ would contain every subset A, since every subset can be written as $A = \bigcup_{x \in A} \{x\}$ and since the singleton sets $\{x\}$ are all in $\mathcal{A}$.

There are some redundancy in the definition above. For example, it follows from De Morgan's Law that if $\mathcal{A}$ is closed under complement and countable unions, then it is automatically closed under countable intersections. Similarly, it follows from countable additivity that we must have $P(\emptyset) = 0$ and that (once we know that $P(\Omega) = 1$ and $P(A) \geq 0$ for all $A \in \mathcal{A}$), we must have $P(A) \leq 1$.

More generally, from additivity, we have $P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1$, whence

$$P(A^c) = 1 - P(A) \tag{2.1}$$

a fact that will be used often. Similarly, if $A \subset B$, then since $B = A \sqcup (B \setminus A)$, where $\sqcup$ means disjoint union, we have that

$$P(B) = P(A) + P(B \setminus A) \geq P(A)$$

and thus,

$$P(A) \leq P(B) \tag{2.2}$$

whenever $A \subset B$, which is the *monotonicity property of probability measures.*

Also, if $A, B \in \mathcal{A}$, then

$$\begin{aligned} P(A \cup B) &= P((A \setminus B) \sqcup (B \setminus A) \sqcup (A \cap B)) \\ &= P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &= P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

the *principle of inclusion-exclusion.*

Finally, if $A_0 = \emptyset$ and for any sequence $(A_n)_{n=1}^{\infty}$ of $\mathcal{A}$ (disjoint or not), we have by the countable additivity and monotonicity that

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigsqcup_{n=1}^{\infty} (A_n \setminus A_{n-1})\right) = \sum_{n=1}^{\infty} P(A_n \setminus A_{n-1}) \leq \sum_{n=1}^{\infty} P(A_n)$$

which is the *countable subadditivity* property of probability measures.

2.2 Constructing Probability Triples

We clarify the definition of Section 2.1 with a simple example.

Example 2.2.1

Let us again consider Poisson(5) distribution that was mentioned in Section 1.1. In this case, the sample Ω would consist of all nonnegative integers: $\Omega = \mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Also, the σ-algebra $\mathcal{A}$ would consist of all subsets of $\mathbb{N}_0$. Finally, the probability measure P would be defined as, for any $A \in \mathcal{A}$,

$$P(A) = \sum_{a \in A} \frac{e^{-5} 5^a}{a!}$$

It is elementary to check that $\mathcal{A}$ is a σ-algebra, as it contains all subsets of $\mathbb{N}_0$, so it is closed under any set operations, and it is also elementary to check that P is a probability measure defined on $\mathcal{A}$ (the additivity following since if A and B are disjoint, then $\sum_{x\in A\cup B}$ is the same as $\sum_{x\in A}+\sum_{x\in B}$).

So in the case of Poisson(5), we see that it is entirely straightforward to construct an appropriate triple $(\mathbb{N}_0, \mathcal{A}, P)$. The construction is similarly straightforward for any discrete probability space, i.e. any space for which Ω is finite or countable. We record this as follows.

Theorem 2.2.2

Let Ω be a finite or countable nonempty set. Let $p : \Omega \to [0,1]$ be any function satisfying $\sum_{\omega\in\Omega} p(\omega) = 1$. Then there exists a valid probability triple $(\Omega, \mathcal{A}, P)$ where $\mathcal{A}$ is the collection of all subsets of Ω, and for $A \in \mathcal{A}$, $P(A) = \sum_{a\in A} p(a)$.

Example 2.2.3

Let Ω be any finite nonempty set, $\mathcal{A}$ be the collection of all subsets of Ω, and P be the probability measure defined as

$$P(A) = \frac{|A|}{|\Omega|}$$

for all $A \in \mathcal{A}$, where $|A|$ denotes the cardinality of the set A. Then $(\Omega, \mathcal{A}, P)$ is a valid probability triple, called the *uniform distribution on* Ω, denoted by Uniform(Ω).

However, if a sample space is not countable, then the situation is considerably more complex, as seen in Section 1.2. How can we formally define a probability triple $(\Omega, \mathcal{A}, P)$ which corresponds to say, the Uniform($[0,1]$) distribution?

It is clear that we should choose $\Omega = [0,1]$, but what about $\mathcal{A}$? We know from Proposition 1.2.1 that $\mathcal{A}$ cannot contain all intervals of Ω, but it should contain all the intervals $[a,b]$, $[a,b)$, etc. That is, we must have $\mathcal{A} \supset \mathcal{B}$, where

$$\mathcal{B} = \{\text{all intervals contained in } [0,1]\}$$

and where "intervals" is understood to include the open, closed, half-open, singleton, and empty intervals.

Exercise 2.2.4

Prove that the above collection $\mathcal{B}$ is a *semi-algebra* of subsets of Ω. That is, show

(i) $\Omega, \emptyset \in \mathcal{B}$

(ii) $\mathcal{B}$ is closed under finite intersection

(iii) The complement of any element of $\mathcal{B}$ is equal to a finite disjoint union of elements of $\mathcal{B}$.

Since $\mathcal{B}$ is only a semi-algebra, how can we create a σ-algebra? As a first try, we might consider

$$\mathcal{C} = \{\text{all finite unions of elements of } \mathcal{B}\} \tag{2.3}$$

After all, by additivity, we already know how to define P on $\mathcal{C}$. However, $\mathcal{C}$ is not a σ-algebra.

Exercise 2.2.5

(a) Prove that $\mathcal{C}$ is an *algebra* (or, *field*) of subsets of Ω, that is,

(i) $\mathcal{C}$ contains Ω and $\emptyset$

(ii) $\mathcal{C}$ is closed under the formation of complements and of finite unions and intersections.

(b) Prove that $\mathcal{C}$ is not a σ-algebra.

As a second try, we might even consider

$$\mathcal{D} = \{\text{all finite or countable unions of elements of } \mathcal{B}\} \tag{2.4}$$

Unfortunately, $\mathcal{D}$ is not a σ-algebra.

Thus, the construction of $\mathcal{A}$ and of P presents serious challenges. To deal with them, we next prove a very general theorem about constructing probability triples.

2.3 The Extension Theorem

The following theorem is of fundamental importance in constructing complicated probability triples. Recall the definition of *semi-algebra* from Exercise 2.2.4.

Theorem 2.3.1: The Extension Theorem

Let $\mathcal{B}$ be a semi-algebra of subsets of Ω. Let $P : \mathcal{B} \to [0,1]$ with $P(\emptyset) = 0$ and $P(\Omega) = 1$ satisfying the finite superadditivity property that for all $A_1, \ldots, A_n \in \mathcal{B}$ such that $\bigcup_{i=1}^n A_i \in \mathcal{B}$, and $(A_i)_{i=1}^n$ are disjoint, we have

$$\sum_{i=1}^n P(A_i) \leq P\left(\bigcup_{i=1}^n A_i\right) \tag{2.5}$$

and also the countable monotonicity property that

$$P(A) \leq \sum_{n=1}^\infty P(A_n) \tag{2.6}$$

for all $A, A_1, A_2, \ldots \in \mathcal{B}$ with $A \subset \bigcup_{n=1}^\infty A_n$. Then there exists a σ-algebra $\mathcal{C} \supset \mathcal{B}$ and a countably additive probability measure P^* on $\mathcal{C}$ such that $P^*(A) = P(A)$ for all $A \in \mathcal{B}$ (that is, $(\Omega, \mathcal{C}, P^*)$ is a valid probability triple, which agrees with our previous probabilities on $\mathcal{B}$).

Remark 2.3.2

Of course, the conclusions of Theorem 2.3.1 imply that (2.5) must actually hold with *equality*. However, (2.5) need only be verified as an *inequality* to apply Theorem 2.3.1.

However, it is not clear how to even start proving this theorem. Indeed, how could we begin to define $P(A)$ for all A in a σ-algebra? The key is given by *outer measure* P^* defined by

$$P^*(A) = \inf\left\{\sum_{n=1}^{\infty} P(A_n) : A_1, ..., A_n \in \mathcal{B} \text{ and } A \subset \bigcup_{n=1}^{\infty} A_n\right\} \tag{2.7}$$

That is, we define $P^*(A)$ for any subset $A \subset \Omega$ to be the infimum of the sums of $P(A_n)$, where $(A_n)_{n=1}^{\infty}$ is any countably collection of the elements of the original semi-algebra $\mathcal{B}$ whose union contains A. In other words, we use the values of $P(A)$ for $A \in \mathcal{B}$ to help us define $P^*(A)$ for any $A \subset \Omega$. Of course, we know that P^* will not necessarily be a proper probability measure for all $A \subset \Omega$; for example, this is not possible for Uniform($[0,1]$) by Proposition 1.2.1. However, it is still useful that $P^*(A)$ is at least *defined* for all $A \subset \Omega$. We shall eventually show that P^* is indeed a probability measure on some σ-algebra $\mathcal{C}$, and that P^* is an extension of P.

To continue, we note a few simple properties of P^*. Firstly, we clearly have $P^*(\emptyset) = 0$; indeed, we can simply take $A_n = \emptyset$ for all $n \in \mathbb{N}$ in (2.7). Secondly, P^* is *monotone*; indeed, if $A \subset B$, then the infimum of (2.7) for $P^*(A)$ includes all choices of $(A_n)_{n=1}^{\infty}$ which work for $P^*(B)$ plus many more besides, so that $P^*(A) \leq P^*(B)$. We also have the following lemma.

Lemma 2.3.3

P^* is an extension of P, i.e. $P^*(A) = P(A)$ for all $A \in \mathcal{B}$.

Proof.

Let $A \in \mathcal{B}$ be arbitrary. From (2.6), that $P^*(A) \geq P(A)$. On the other hand, let $A_1 = A$ and $A_n = \emptyset$ for all $n \in \mathbb{N}_{\geq 2}$ in (2.7) shows that $P^*(A) \leq P(A)$. ■

Lemma 2.3.4

P^* is countably subadditive; that is, for all $(B_n)_{n=1}^{\infty}$ of Ω,

$$P^*\left(\bigcup_{n=1}^{\infty} B_n\right) \leq \sum_{n=1}^{\infty} P^*(B_n)$$

Proof.

Let $(B_n)_{n=1}^{\infty}$ be arbitrary. From (2.7), we see that for any $\varepsilon > 0$, there exists a subsequence $(B_{n_k})_{k=1}^{\infty}$ for each $n \in \mathbb{N}$ with $B_{n_k} \in \mathcal{B}$ such that $B_n \subset \bigcup_{k=1}^{\infty} B_{n_k}$ and $\sum_{k=1}^{\infty} P(B_{n_k}) \leq P^*(B_n) + \varepsilon 2^{-n}$. But then the overall collection $(B_{n_k})_{k=1}^{\infty}$ contains $\bigcup_{n=1}^{\infty} B_n$. It follows that $P^*\left(\bigcup_{n=1}^{\infty} B_n\right) \leq \sum_{n,k=1}^{\infty} P(B_{n_k}) \leq \sum_{n=1}^{\infty} P^*(B_n) + \varepsilon$. Since this is true for any $\varepsilon > 0$, we have that

$$P^*\left(\bigcup_{n=1}^{\infty} B_n\right) \leq \sum_{n=1}^{\infty} P^*(B_n)$$

as claimed. ■

Now we set

$$\mathcal{C} = \{A \subset \Omega : P^*(A \cap E) + P^*(A^c \cap E) = P^*(E), E \subset \Omega\} \tag{2.8}$$

That is, $\mathcal{C}$ is the collection of all subsets A with the property that P^* is additive by the union of $A \cap E$ and $A^c \cap E$ for all $E \subset \Omega$. Note that it follows by subadditivity that

$$P^*(A \cap E) + P^*(A^c \cap E) \geq P^*(E)$$

so (2.8) is equivalent to

$$\mathcal{C} = \{A \subset \Omega : P^*(A \cap E) + P^*(A^c \cap E) \le P^*(E), E \subset \Omega\} \tag{2.9}$$

which is sometimes helpful. Furthermore, P^* is countably additive on $\mathcal{C}$.

Lemma 2.3.5

If $(A_n)_{n=1}^{\infty} \in \mathcal{C}$ are disjoint, then $P^*\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P^*(A_n)$.

Proof.

If A_1 and A_2 are disjoint with $A_1 \in \mathcal{C}$, then

$$P^*(A_1 \cup A_2) = P^*(A_1 \cap (A_1 \cup A_2)) + P^*(A_1^c \cap (A_1 \cup A_2)) = P^*(A_1) + P^*(A_2)$$

Hence, by induction, the lemma holds for any finite collection of A_n.

Then with countably many disjoint $A_n \in \mathcal{C}$, we see that for any $m \in \mathbb{N}$,

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

where the inequality follows from monotonicity. Since this holds for any $m \in \mathbb{N}$, we have that

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

On the other hand, by subadditivity, we have

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

So the lemma holds for countably many A_n as well. ■

The plan now is to show that $\mathcal{C}$ is a σ-algebra containing $\mathcal{B}$. We break up the proof into several lemmas.

2.4 Constructing the $\mathrm{Uniform}[0, 1]$ Distribution

2.5 Corollaries of the Extension Theorem

2.6 Coin Tossing and Other Measures

Chapter 3

Further Probabilistic Foundations

3.1 Random Variables

3.2 Independence

3.3 Continuity of Probabilities

3.4 Limit Events

3.5 Tail Fields

Chapter 4

Expected Values