SM-4335: Advanced Probability

Dr. Haziq Jamil

Assistant Professor in Statistics

`haziq.jamil@ubd.edu.bn`

https://haziqj.ml

`DRAFT - document last updated: 2022-07-16`

Semester I 2022/23

Module Description

Welcome to SM-4335! This is a course where we study the attempt in making probabilities rigorous, in the mathematical sense. At the outset I should tell you that there certainly will be an abundance of discussions surrounding abstract concepts in this course, particularly building on some fundamental ideas in foundational mathematics such as sequences, limits, functions or maps, and other notions from number theory.

Modern probability theory, as we will discover, branches off from a section of mathematics called measure theory. Using measure-theoretic ideas, we build a robust framework that allows us to confidently define and measure probabilities of events and perform calculus involving random variables. What this course will not focus on is probability calculations (for example, involving known probability distributions) like the one you are used to in SM-2205 or earlier.

At this point you might be wondering why is it, exactly, that you need such heavy machinery for computing probabilities that you already know how to compute. Understandably, it lends greatly to your decision whether or not to take this module! In the next few paragraphs, I'll try to make the case for why measure theory is important for probability theory.

Probability distributions

Undoubtedly, you've come across probabilities before, and perhaps you are currently taking some other courses in probability or statistics. A classical example is that of a die roll–if we represent the (random) outcome of the die roll by X, then we say that $\Pr(X = x) = \frac{1}{6}$, for $x \in \{1, \ldots, 6\}$. This is an example of a *discrete* probability function.

We know that we treat *continuous* random variables differently: for instance, let Y represent the height of adult males in Brunei. Y could be modelled by the normal distribution centred around μ with some variability σ^2 (the variance). If we wanted to compute, say, the probability of individual being less than or equal to 1.65 m in height, we would do $\Pr(Y \leq 1.65) = \int_{-\infty}^{1.65} f_Y(y)\, dy$. Here, f_Y is the probability density function.

Reliance on pdf leaves us wanting

In either of the above cases, the central object of discussion is the *probability distribution*, encapsulated by the probability mass function (pmf) or probability density function (pdf) for discrete and continuous random variables respectively. It turns out that these pmf/pdf objects can be quite limiting, being inadequate under certain circumstances. There are two examples that come to mind:

(a) Let X and Y be as defined, and now consider a new random variable Z, to be defined as follows. Flip a fair coin: if heads, set $Z = X$ (the outcome of the die roll); if tails, set $Z = Y$ (the normal random variable). What kind of random variable is this? Is it discrete or continuous? What would the graph of the probability function look like? How do we compute the expectation of Z, $E(Z) = \int z f_Z(z)\,dz$, especially if the pdf f_Z is not absolutely continuous?

(b) If I asked you to pick a random number between 0 and 1, the first thing you'll ask is what is its distribution? If it's uniform, then in theory any number is equally likely, but the point is you can easily conceptualise the notion of picking a random number. What if, instead, you were interested in picking a *random function*, say from $C[0, 1]$. How would you do this? What would the pdf be? Of note, this is highly of interest in statistical finance (think stock prices behaving like random functions).

Outcomes and events

As a first step in rectifying the issue, we have to shift our focus away from pdfs. Instead, we have to define the probability for every event in our random "experiment". In probability theory and statistics, an experiment (or trial) is a procedure that can be repeated and has a well-defined set of possible outcomes (known as the sample space Ω). Events are thought of as being subsets of Ω, while probabilities are merely a mapping from some *event space* $\mathcal{F}$ to $[0, 1]$.

To make this idea concrete, for the die roll example, $\Omega = \{1, \ldots, 6\}$, while an event could be $E = \{2, 4, 6\} \subset \Omega$ (getting an even number). The probability of the event E occurring is $\Pr(E) = \frac{1}{2}$–so it indeed behaves like a function, taking input some event and spitting out a number between 0 and 1.

Note here that $\mathcal{F}$ is not Ω–it has to be bigger than Ω as we're not just interested in singleton outcomes. A good starting point would be $\mathcal{F} = \mathcal{P}(\Omega)$, the set of all subsets of Ω, which should contain all possible events constructed from the set of outcomes.

Rules of probability

Having abstracted the notion of Ω and $\mathcal{F}$, we should also define some rules that the probability function $\Pr : \mathcal{P}(\Omega) \to [0, 1]$ must follow. Let us list down a few:

i. $\Pr(E) \geq 0, \forall E$;
ii. $\Pr(\varnothing) = 0$ and $\Pr(\Omega) = 1$;[1]
iii. If $E_1 \cap E_2 = \varnothing$, then $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$; and
iv. If $E_1, E_2, \ldots$ are mutually disjoint events, then $\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i)$.

Indeed, these are quite logical impositions that ensure we don't end up with nonsensical probabilities. For instance, by ii. and iii., modelling a (biased) coin toss by $\Pr(H) = 0.7$ necessitates $\Pr(T) = 0.3$ and not anything else, e.g. $\Pr(T) = 0.5$.

The need for measure theory

We've managed to come up with probability rules so far without the need for measure theory, so what's the problem? The problem is that in the way that we've described it, this is actually too much to ask! There will be instances where this whole framework fails and we can't assign probabilities properly, especially when we need it the most.

[1] $\varnothing = \{\}$ is the empty set.

For instance, with all these demands, we can't even define the uniform random variable on $\Omega = [0, 1]$! That is, no mapping $\Pr : \mathcal{P}([0, 1]) \to [0, 1]$ exists such that $\Pr([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$ that satisfies all of the rules i. to iv. listed above. For a proof, see the appendix.

Evidently some concession has to be made, and the probability map must be constructed more carefully. The answer lies in measure theory.

The beauty in the use of measure theory to define probabilities in this abstract sense is that it can apply to a variety of settings. To resolve (a), the hybrid random variable case, measure theory affords us a unifying concept of integration. To resolve (b), the issue of random functions, we may apply the theory to $\Omega = C([0, 1])$ which generates a valid event space (specifically, it's called a Borel set–more details later!).

As a final note, we don't have to worry about describing the entire sample space and event space *every time* we encounter a probability problem. Often times, the object of interest is a random variable anyway, which is seen as a (measurable) map from the sample space to the real line. Again, measure theory is useful here!

Module Contents and Learning Outcomes

In the introduction we've alluded to a couple of notions, especially from set theory and measure theory, that we will definitely delve deeper in this course. In 14 weeks, we aim to cover the following topics:

- Set theory, algebra of sets, and limit sets
- Measure theory & probability measures
- Measurable functions & random variables
- Integrals & expectations
- Convergence of random variables

Here are the learning outcomes of this course (which also serves as a useful checklist of things you should know when studying for the exam).

1. Set theory
    - Construct sets and perform set operations including unions, intersections, differences and complements.
    - Discern whether a set is finite or infinite, countable or uncountable, and compare cardinalities.
    - Take limits of bounded sets by taking countable unions or intersections, and be able to describe limit sets in everyday words.
2. Algebra of sets
    - Define an algebra of sets, and know how to construct the smallest algebra generated by a collection of sets.
    - Define a σ-algebra of sets, and know how to construct the smallest σ-algebra generated by a collection of sets.
    - Know how to check whether a given collection is the smallest (σ-)algebra containing that collection.
    - Prove that the σ-algebra generated by two different collection of sets is the same.
3. (Probability) measure
    - Define a measure, and in particular, a probability measure, and how to check if a set-theoretic function is a probability measure.
    - Know and be able to prove the properties of a probability measure.

- State Caratheodory's extension theorem, and use it in order to decide whether the extension of a probability defined on an algebra to a given σ-algebra is unique.
- Construct counter-examples to disprove uniqueness of probability measures, i.e. two probability measures that are the same on the given algebra, but different on the given σ-algebra.

4. Random variables

   - Define a random variable as a measurable map, and check whether a given function is a random variable/measurable.
   - Know that a random variable induces a probability space, in particular how to obtain the smallest σ-algebra generated by a random variable.
   - Construct the Lebesgue integral for measurable functions, and understand how it is used to compute expectations of random variables.
   - Define the expectations of indicator, simple random variables, and non-negative random variables, and how to use these as building blocks for arbitrary random variables.
   - Be familiar with the Monotone Convergence Theorem and the Dominated Convergence Theorem, and be able to use them to compute expectations of random variables that are limits of a sequence of simpler random variables.

5. Convergence

   - State and prove both Borel-Cantelli lemmas, and use them in order to compute probabilities of limit sets.
   - Know the definition of "almost sure convergence" and "convergence in probability".
   - Prove Markov's inequality and use it in order to prove convergence in probability.
   - State and apply the Law of Large Numbers.

Readings

> Some thirty years ago it was still possible, as Loève so ably demonstrated, to write a single book in probability theory containing practically everything worth knowing in the subject. The subsequent development has been explosive, and today a corresponding comprehensive coverage would require a whole library.
>
> —Olav Kallenberg in *Foundations of Modern Probability, 1997*

There will not be an *official* textbook for this course. Different textbooks will cover probability theory differently, due to the vast knowledge space. My lectures will be curated from various sources, one that I think is suitable to be run as a fourth-year module in 14 weeks. However, the following is a list of books that I will occasionally refer. This should also serve as a suggestive reading list.

- David Williams. *Probability with martingales.* Cambridge University Press, 1991
- Patrick Billingsley. *Probability and measure.* John Wiley & Sons, 2008
- Jeffrey S Rosenthal. *A first look at rigorous probability theory.* World Scientific Publishing Company, 2006
- John B Walsh. *Knowing the odds: An introduction to probability.* Vol. 139. American Mathematical Society, 2012
- Geoffrey Grimmett and David Stirzaker. *Probability and random processes.* Oxford university press, 2020
- John Christopher Taylor. *An introduction to measure and probability.* Springer Science & Business Media, 2012

Supplementary reading list

If you find yourself struggling with fundamental concepts in mathematics, I urge you to pick up these books.

- Richard H Hammack. *Book of proof.* Richard Hammack Virginia, 2013. URL: https://www.people.vcu.edu/~rhammack/BookOfProof/
- Ian Stewart and David Tall. *The foundations of mathematics.* OUP Oxford, 2015
- Sheldon M Ross. *A first course in probability.* Pearson Boston, 2019

Class Format

See the end of the document for the full schedule

There are two official time slots that are time-tabled for this module:

1. Wednesday 11.50am–1.40pm
2. Saturday 7.50am–9.40am

In general, lectures will be delivered on Wednesdays, and Saturday slots will be tutorial sessions. As such, there will be new material learnt every week. However, tutorials will cover exercise sheets pertaining to the material of the previous week, giving you time to prepare.

Lectures will be face-to-face sessions in a classroom (room TBC). I plan to lecture off of PDF slides, and will supplement the slides with some annotations (either on the PDF slides themselves or on the whiteboard).

Tutorial sessions are a great opportunity for discussing the problem sets. While the solutions themselves are not assessed, your participation is. This means that the more you contribute to discussions, offer ideas, and volunteer to present, the more points you gain. The rubric is found in the next section.

Assessment

Formative assessment

- Diagnostic test
- Exercise sheets
- Topical quizzes (on Canvas)—TBC

The diagnostic test is given at the beginning of Week 1. It is a simple test intended to gauge your level of readiness for this course. The topical quizzes are multiple choice questions on Canvas designed to test your understanding of the concepts that we've covered.

Summative assessment

- **60% examination**: The exam paper will consist of tutorial-style questions, i.e. problem sets that require written solutions. You will be required to answer a total of 3 questions in the span of 2 hours.

  The scheduled date for the exam is **DAY DATE TBC 2022 @ TIME TBC**.

- **30% class tests**: Two short class tests are scheduled: one in week 7 and one in week 14. Again these will be tutorial-style questions. The test will be one hour in length.
- **10% participation**: I will implement a "token" system in order to incentivise participation during class (lectures and tutorials). You will each have the opportunity to earn weekly points for your participation. Points are awarded for

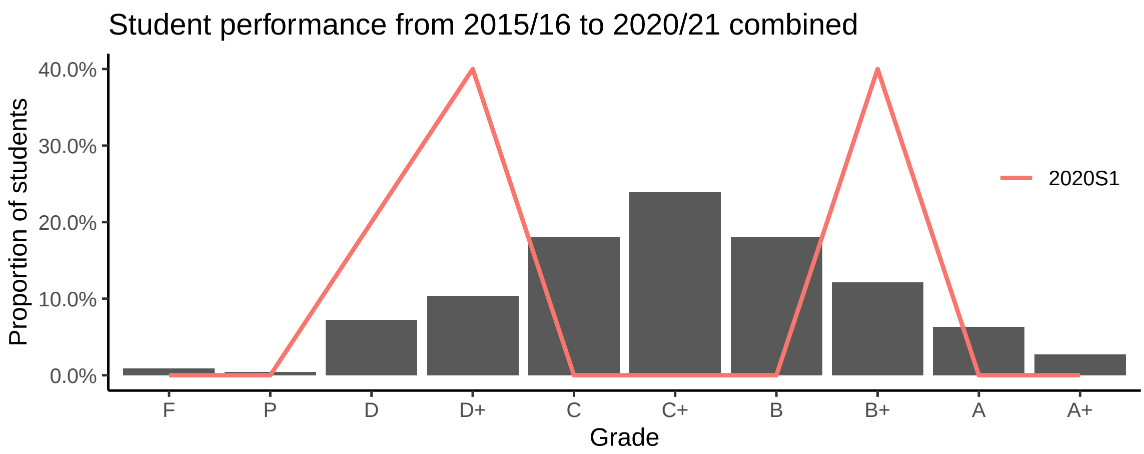
1. Answering questions posed by me (or others).
2. Offering ideas during problem-solving discourses.
3. Volunteering to present ideas or solutions on the board.
4. Asking beneficial, course-related questions.

As long as your participation is related to the course, you will earn the points. Points will be tallied weekly and recorded on Canvas for you to keep track.

- **5% BONUS**: Submit your written notes to me and I will award you bonus marks (up to 5%) based on the completeness, tidiness, and usability of your notes.

Key Data

- Past class sizes: 2015S1 = 23, 2016S1 = 78, 2017S1 = 62, 2018S1 = 49, 2019S1 = 5, 2020S1 = 5 (average: 37)
- SFE grade average: 3.7 / 5.0



Module Advice

Here are some advice that I can offer you to make this course bearable.

1. You must write your own notes. My slides will only be 70% of the way there, and you should aim to supplement the knowledge with your own writing.
2. Form study groups, whether you're ahead or behind on the materials. Talking about problems among peers is a great way to learn.
3. When you get stuck (and you will), take your mind off the problem by doing some other activities. Fermentation of ideas is a necessary step in mathematical problem solving.
4. Complete all tasks early, especially exercise sheets. This gives you ample opportunity to reflect on the problem set. **Do not come to tutorials empty handed.**
5. Study for the tests. Some students instead like to focus solely on the exam. That would definitely be the wrong strategy.
6. Peruse the reading list. Each lecture will be accompanied by some suggested reading. In addition, there are some great lectures on YouTube about measury theory that you might find helpful as well.
7. Come to my office hours. From time to time I will announce office hours so that you may come and see me to discuss any issues you face about the course. Please do make use of these opportunities.

Class Schedule

Here is the class schedule for the August 2022/23 semester.

Week 01, 01/08 – 07/08

- Lecture: Introduction
- Complete diagnostic test

Week 02, 08/08 – 14/08

- Lecture: Elementary set theory
- Lecture: Algebra of sets

Week 03, 15/08 – 21/08

- Lecture: σ-algebra
- Tutorial: Exercise 1 (Sets)

Week 04, 22/08 – 28/08

- Lecture: Limit sets
- Tutorial: Exercise 2 (Algebras and σ-algebras)

Week 05, 29/08 – 04/09

- Lecture: The Cantor set
- Tutorial: Exercise 3 (Limit sets)

Week 06, 05/09 – 11/09

- Lecture: Probability measure
- Tutorial: Exercise 4 (Uncountable sets and their sizes)

Week 07, 12/09 – 18/09

- Tutorial: Exercise 5 (Probability measures)
- Class test
- Submit notes for marking

19/09 – 25/09: Mid-semester Break

No classes. Take a break!

Week 08, 26/09 – 02/10

- Lecture: Properties of probability measure & extension theorem
- Tutorial: Class test review

Week 09, 03/10 – 09/10

- Lecture: Random variables
- Lecture: Lebesgue measure
- Tutorial: Exercise 6 (The Cantor distribution)

Week 10, 10/10 – 16/10

- Lecture: Lebesgue integral
- Tutorial: Exercise 7 (Extension theorem and random variables)

Week 11, 17/10 – 23/10

- Lecture: Expectations
- Tutorial: Exercise 8 (Lebesgue integral)

Week 12, 24/10 – 30/10

- Lecture: Convergence
- Tutorial: Exercise 9 (Expectations)

Week 13, 31/10 – 06/11

- Lecture: Independence
- Tutorial: Exercise 10 (Convergence of random variables)

Week 14, 07/11 – 13/11

- Tutorial: Exercise 11 (Independence)
- Class test

14/11 – 20/11: Revision Week

- Review of class test

Exam

- Date TBC

Appendix

An unmeasurable set

As mentioned, we are unable to define a uniform probability measure on the unit interval, given by

$$\Pr([a,b]) = b - a$$

that satisfies all the probability rules listed in i. to iv. earlier. On the face of it, all the rules themselves are satisfied: $\Pr(\Omega) = \Pr([0,1]) = 1$, $\Pr(\varnothing) = \Pr([a,a]) = 0$ (for any $a \in [0,1]$), and certainly probabilities of disjoint subsets of $[0,1]$ are just the sum of the lengths of the intervals.

These are all great properties to have, so we must concede instead on the domain of the probability function, i.e. the event space. The proof of the proposition below in instructive, in that it illustrates the existence of a “non-measurable” set. That is, there are such events (subsets in $[0,1]$) for which we are unable to assign probabilities to.

Proposition 1. *There does not exist a definition of* $\Pr : \mathcal{P}([0,1]) \to [0,1]$ *satisfying* $\Pr([a,b]) = b - a$ *and i. to iv. (as listed earlier).*

Proof. All we need to show is the existence of one such subset of $[0,1]$ whose measure is undefined. The set we are about to construct is called the Vitali set[2], after Giuseppe Vitali who described it in 1905.

Before proceeding, we introduce some notation. For a uniform measure on $[0,1]$, one expects that the measure of some subset $A \subseteq [0,1]$ to be unaffected by “shifting” (with wrap-around) of that subset by some fixed amount $r \in [0,1]$. Define the *r-shift* of $A \subseteq [0,1]$ by

$$A \oplus r := \{a + r \mid a \in A, a + r \le 1\} \cup \{a + r - 1 \mid a \in A, a + r > 1\}.$$

Then we should have

$$\Pr(A \oplus r) = \Pr(A).$$

For example, $\Pr([0.7, 0.9] \oplus 0.2) = \Pr([0.9, 1] \cup [0, 0.1]) = 0.2$.

Now, define an equivalence relation on $[0,1]$ by the following:

$$x \sim y \Rightarrow y - x \in \mathbb{Q}$$

That is, two real numbers x and y are deemed to be similar if their difference is a rational number. The intent is to segregate all the real numbers $x \in [0,1]$ by this equivalence relation, and collect them into groups called equivalence classes, denoted by $[x]$. Here, $[x]$ is the set $\{y \in [0,1] \mid x \sim y\}$. For instance,

- The equivalence class of 0 is the set of real numbers x such that $x \sim 0$, i.e. $[0] = \{y \in [0,1] \mid y - 0 \in \mathbb{Q}\}$, which is the set of all rational numbers in $[0,1]$.
- The equivalence class of an irrational number $z_1 \in [0,1]$ is clearly not in $[0]$, thus would represent a different equivalence class $[z_1] = \{y \in [0,1] \mid y - z_1 \in \mathbb{Q}\}$.
- Yet another irrational number $z_2 \notin [z_1]$ would exist, i.e. a number $z_2 \in [0,1]$ such that $z_2 - z_1 \notin \mathbb{Q}$, and thus would represent a different equivalence class $[z_2]$.
- And so on...

[2]https://en.wikipedia.org/wiki/Vitali_set

The equivalence classes may therefore be represented by $[0], [z_1], [z_2], \dots$ where z_i are all irrational numbers that differ by an irrational number, and there are uncountably many such numbers, and therefore classes.

Construct the Vitali set V as follows: Take precisely one element from each equivalent class, and put it in V. As a remark, such a V must surely exist by the Axiom of Choice[3].

Consider now the union of shifted Vitali sets by some rational value $r \in [0, 1]$,

$$\bigcup_r (V \oplus r)$$

As a reminder, the set of rational numbers is countably infinite[4]. We make two observations:

1. **The equivalence relation partitions the interval $[0, 1]$ into a disjoint union of equivalence classes.** In other words, the sets $(V \oplus r)$ and $(V \oplus s)$ are disjoint for any rationals $r \neq s$, such that $r, s \in [0, 1]$. If they were not disjoint, this would mean that there exists some $x, y \in [0, 1]$ with $x + r \in (V \oplus r)$ and $y + s \in (V \oplus s)$ such that $x + r = y + s$. But then this means that $x - y = s - r \in \mathbb{Q}$ so x and y are in the same equivalent class, and this is a contradiction. Importantly,
$$\Pr\left(\bigcup_r (V \oplus r)\right) = \sum_r \Pr(V \oplus r) = \sum_r \Pr(V) \tag{1}$$

2. **Every point in $[0, 1]$ is contained in the union $\bigcup_r (V \oplus r)$.** To see this, fix a point x in $[0, 1]$. Note that this point belongs to some equivalent class of x, and in this equivalence class there exists some point α which belongs to V as well by construction. Hence, $\alpha \sim x$, and thus $x - \alpha = r \in \mathbb{Q}$, implying that x is a point in the Vitali set V shifted by r. Therefore,
$$[0, 1] \subseteq \bigcup_r (V \oplus r).$$
and we may write
$$1 = \Pr([0, 1]) \leq \Pr\left(\bigcup_r (V \oplus r)\right) \leq 1,$$
since the measure of any set contained in another must have smaller or equal measure (a relation implied by property iii.[5]) as well as all probabilities are less than equal to 1[6]. We see that
$$\Pr\left(\bigcup_r (V \oplus r)\right) = 1. \tag{2}$$

Equating (1) and (2) together, we find a contradiction: A countably infinite sum of a constant value can only equal 0, $+\infty$ or $-\infty$, but never 1. ■

[3] Given a collection of non-empty sets, it is always possible to construct a new set by taking one element from each set in the original collection. See https://brilliant.org/wiki/axiom-of-choice/

[4] https://www.homeschoolmath.net/teaching/rational-numbers-countable.php

[5] Let A and B be such that $A \subseteq B$. Then we may write $B = A \cup (B \setminus A)$ where the sets A and $B \setminus A$ are disjoint. Hence, $\Pr(B) = \Pr(A) + \Pr(B \setminus A)$, and since probabilities are non-negative, we have that $\Pr(B) \geq \Pr(A)$.

[6] For any A, $\Pr(\Omega) = \Pr(A \cup A^c) = \Pr(A) + \Pr(A^c) = 1$, so $\Pr(A) \leq 1$.