Statistics 210: Probability I

Eric K. Zhang
ekzhang@college.harvard.edu

Fall 2020

Abstract

These are notes for Harvard's *Statistics 210*, a graduate-level probability class providing foundational material for statistics PhD students, as taught by Joe Blitzstein[1] in Fall 2020. It has a history as a long-running statistics requirement at Harvard. We will focus on probability topics applicable to statistics, with a lesser focus on measure theory.

Course description: Random variables, measure theory, reasoning by representation. Families of distributions: Multivariate Normal, conjugate, marginals, mixtures. Conditional distributions and expectation. Convergence, laws of large numbers, central limit theorems, and martingales.

Contents

[1] With teaching fellows: Ambarish Chattopadhyay, Louis Cammarata, Franklyn Wang, Michael Isakov, Mike Bao

1 September 3rd, 2020

We start with an overview of the course. The class has roughly 80 students, ranging from first-year PhD students in statistics and other areas to advanced undergraduates. We will cover many aspects of probability from a rigorous standpoint.

1.1 Course Logistics

Since the class will be held virtually, a lot of the interaction will be on the Ed discussion board. Regular office hours and sections will be organized every week, and you can feel free to attend as many (or as few) sessions as you wish.

We'll start with a bit of measure theory, to provide a foundation for the probability in this course. The goal is to provide *enough* foundations to understand a paper using σ-algebras or other machinery, but not to have a course *dominated* by measure theory. This way, you can do things in sufficient generality, but you don't have to spend hundreds of hours proving pedantic things (for example, showing measurability).

Like Stat 110 (the undergraduate partner course), we will cover various distributions, their properties, and useful machinery. However, we generally try to allow deeper, more general analysis with less assumptions. For example:

- Multivariate normal distributions, but also mixtures of Gaussians.
- Central limit theorem, including variants that don't assume i.i.d. variables.

The course material is structured from Joe Blitzstein and Carl Morris's forthcoming textbook, *Probability for Statistical Science* [BM20]. Key philosophies of the course: conditioning distributions and balancing "coin-flipping" intuition versus analysis.

1.2 Breakout Puzzle: Random Walks

Joe gives us a "brain-teaser" puzzle that we can work on in breakout rooms.

Exercise 1.1 (Simple symmetric random walk)**.** Suppose that you have a simple, symmetric random walk on the real line, moving either $+1$ or -1 on each step with independent probability $\frac{1}{2}$. If you start at 0, what is the expected number of times you reach 10^{100} before returning to 0?

Proof. The answer is 1. Let $b = 10^{100}$, and we can proceed in either of a couple of ways:

- Let p be the probability that we reach 10^{100} at least once before returning to 0.[2] Then, the distribution of the number of visits N to 10^{100} before returning is

$$[N \mid N \geq 1] \sim FS(p),$$

  where FS is the "first-success" geometric distribution with success probability p. Therefore,

$$\mathbf{E}[N] = \Pr(N \geq 1) \cdot \mathbf{E}[FS(p)] = p \cdot \frac{1}{p} = 1.$$

- Imagine that during your random walk, you decide to write down an infinite sequence of letters: 'A' whenever you hit the number 0, and 'B' whenever you hit the number b. This

[2] We can actually compute that $p = \frac{1}{2 \cdot 10^{100}}$ with martingales, but this is irrelevant.

creates some long string $AAAAAABBBBAA\ldots$. For symmetry, we can start from the point $b/2$ and generate this string. Since the random walk is memoryless, we simply want to know the expected number of B's we hit between any two adjacent A's.

By symmetry, the expected number of A's and B's in any finite subsegment is equal. Since every B (except a finite number) is between a pair of A's with high probability, we have that

$$\mathbf{E}[N] = \lim_{n\to\infty} \frac{\#(\text{number of } B\text{'s})}{\#(\text{number of } A\text{'s})} = 1.$$

□

1.3 Representations of Distributions

We'll introduce a couple of neat distributions. It turns out that dealing with the *representations* of distributions can be a very powerful tool. We learned about the exponential distribution in Stat 110, but it turns out that the following distribution is more common.

Definition 1.1 (Weibull distribution)**.** The *Weibull distribution* is given by the power X^c of an exponential random variable X.

Joe notes that entire books have been written on the Weibull distribution. Here's another interesting distribution.

Definition 1.2 (Cauchy distribution)**.** The *Cauchy distribution* has probability density function

$$C : p(x) = \frac{1}{\pi(1+x^2)}.$$

Example 1.3. There are several interesting properties of the Cauchy distribution in terms of *representations* by other distributions:

- If $z_1, z_2 \sim \mathcal{N}(0,1)$ are independent, then $z_1/z_2 \sim C$.
- If $z \sim C$, then $1/z \sim C$ (corollary of above).
- If $x, y \sim C$, then $\frac{1+x}{1+y} \sim C$.

Another neat fact is that if $x, y \sim \mathcal{N}(0,1)$, then $x+y$ is independent from $x-y$![3]

Finally, we'll give some intuition for our forays into measure theory, starting next lecture.

Example 1.4 (Banach-Tarski Paradox)**.** Assuming the axiom of choice, there exists a way to decompose a 3-ball B^3 into two separate, yet congruent balls. However, *at least one of the sets must not be measurable.*

In some sense, the intuition of measure theory allows you to rigorously define an intuitive concept of *mass*. This can also help axiomatize concepts to get at the core of problems. We'll see that measure theory lets us unify many proofs for different distributions into a single general proof.

[3] This is a special property of the normal distribution, not a general fact.

2 September 8th, 2020

Today is our first real lecture, where we introduce measure theory and its applications to continuous distributions.

2.1 Measure Theory

The motivation here is to rigorously define what it means to be *measurable*, so that we can talk about continuous random variables in a reasonable way.

Example 2.1. Suppose that we had a continuous random variable X varying uniformly on $[0, 1]$. Then, how can we calculate

$$\Pr(X = 0 \mid X \in \{0, 1\})?$$

We would expect, intuitively, for the answer to be $\frac{1}{2}$. However, if we naively apply the definition of conditional probability, we get something like

$$\Pr(X = 0 \mid X \in \{0, 1\}) = \frac{\Pr(X = 0)}{\Pr(X = 0 \cup X = 1)} = \frac{0}{0}.$$

This is not well-defined, so we are unhappy.

Similarly, we want "fundamental" laws of probability like Bayes' Rule to be formalized over continuous probability distributions like this. The core concept that will allow this to be possible is called a σ-algebra.

Definition 2.2 (σ-algebra)**.** Given a set X, a *σ-algebra* on X is a collection $\Sigma \subset 2^{\Sigma}$, which satisfies the following axioms:

- $X \in \Sigma$,
- If $A \in \Sigma$, then $A^c = X \setminus A \in \Sigma$,
- If $A_1, \ldots, A_n \in \Sigma$, then $A_1 \cup \cdots \cup A_n \in \Sigma$.

Unlike a typical set algebra, which is a collection of subsets that is closed under *finite* unions and intersections, a σ-algebra is closed under *countable* unions and intersections (hence the letter σ). The important takeaway from this definition is that it's *fine* enough to talk about probability in a reasonable way, but *coarse* enough so that we don't have Banach-Tarski and friends.

Now we can define the core concept of a *probability measure*.

Definition 2.3 (Probability measure)**.** Let Ω be a set of *samples*, and $\mathcal{F}$ a σ-algebra on Ω, called the *events*. A function $P : \mathcal{F} \to [0, 1]$ is called a *probability function* if it satisfies the following axioms:

- For any countable collection $A_1, A_2, \ldots \in \mathcal{F}$ of pairwise disjoint sets,

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$$

- $P(\Omega) = 1$.

Note. Since this isn't a measure theory course (Math 114), we don't usually care about measures in general. A general *measure space* is defined the same way as a *probability space*, except we call it (X, Σ, μ) instead of $(\Omega, \mathcal{F}, P)$ by convention, and we also do not require the last axiom $P(\Omega) = 1$. Indeed, probability measures are the special case where the *total measure is finite*.

Let's do a couple of examples to visualize σ-algebras.

Example 2.4 (Finite σ-algebra)**.** Suppose that you partition Ω into four disjoint subsets,

$$\Omega = A \coprod B \coprod C \coprod D.$$

Then, the σ-algebra generated by $\{A, B, C, D\}$ has 16 elements, and can be written as

$$\begin{aligned}\mathcal{F} = \{&\emptyset, A, B, C, D,\\ &A \cup B, A \cup C, A \cup D, B \cup C, B \cup D, C \cup D,\\ &A \cup B \cup C, A \cup B \cup D, A \cup C \cup D, B \cup C \cup D,\\ &A \cup B \cup C \cup D\}.\end{aligned}$$

It turns out that all finite σ-algebras basically look like this. They all have a power-of-two size, and they consist of all subcollections of some finite collection of events.

Essentially, finite σ-algebras are uninteresting because they're too coarse, but it does help lend some intuition for the general case. We can think of σ-algebras as offering some kind of *information* about the events that we have observed so far. This lends itself to the following definition:

Definition 2.5 (Filtration)**.** Given a probability space $(\Omega, \mathcal{F}, P)$, a *filtration* is a sequence of sub σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ where for all $k \le \ell$,

$$\mathcal{F}_k \subseteq \mathcal{F}_\ell \subseteq F.$$

2.2 Uncountable σ-Algebras

Here's an proof-based exercise to work on in breakout rooms.

Exercise 2.1. Show that any infinite σ-algebra is uncountable.

Proof. The sketch of the proof looks as follows. Suppose for the sake of contradiction that you had some countably infinite σ-algebra consisting of subsets $\mathcal{F} = \{A_1, A_2, \ldots\}$, where A_i is indexed by each natural number $i \in \mathbb{N}$. Then define the *atoms* of $\mathcal{F}$ to be sets B_x for each $x \in \Omega$, where

$$B_x = \bigcap_{A_i \ni x} A_i.$$

In other words, B_x is the smallest measurable set containing x. We claim that all distinct atoms are disjoint. In other words, if $B_x \cap B_y \neq \emptyset$, then there exists some $z \in B_x \cap B_y$, so $B_z \subseteq B_x \cap B_y$.

Assume for the sake of contradiction that $x \notin B_z$. Then, $B_x \setminus B_z$ is a subset containing x but not containing z. However, this implies that $B_x \setminus B_z \subseteq B_x$, so $z \notin B_x$, which is a contradiction. Therefore $x \in B_z$, and by symmetry $y \in B_z$ as well, so $B_x = B_y = B_z$.

Finally, consider the set of all atoms $\{B_x\}_{x \in X}$. If this set is finite, then $\mathcal{F}$ must be finite as well, which is a contradiction. Therefore there must be at least countably many distinct atoms $B_1, B_2, \ldots$. We can define an injective map $f : 2^{\mathbb{N}} \to \mathcal{F}$ by

$$f(\{n_1, n_2, \ldots\}) = B_{n_1} \cup B_{n_2} \cup \cdots,$$

so $\#(\mathcal{F}) \ge \#(2^{\mathbb{N}}) = 2^{\aleph_0}$. □

Now we know the axioms of probability, and everything starts from here!

3 September 10th, 2020

Last lecture, we broke off after defining the foundations of probability: two axioms that define everything from basics to modern research. We won't cover too much more about this, as that is the topic of measure theory classes (Math 114, Statistics 212). Instead we'll shift gears and start defining higher-level concepts.

3.1 The Borel Measure

We're going to start working with the reals soon, so it's useful to define a measure on the reals. To do that, we'll first need a bit of machinery.

Proposition 3.1 (Intersection of σ-algebras)**.** *If $\mathcal{A}, \mathcal{B} \subseteq 2^\Omega$ are σ-algebras on Ω, then their intersection $\mathcal{A} \cap \mathcal{B}$ is also a σ-algebra. This also holds for infinite intersections.*

Proof. Straightforward, verify the σ-algebra properties directly. □

Be careful! The above proposition does not work for *unions* of σ-algebras.

Definition 3.2 (σ-algebra generated by subsets)**.** Given a collection of subsets $\mathcal{A} \subseteq 2^\Omega$, we define the *$\sigma$-algebra generated by $\mathcal{A}$* to be the smallest σ-algebra containing $\mathcal{A}$, i.e.,

$$\sigma(\mathcal{A}) = \bigcap_{\substack{\mathcal{A} \subseteq \mathcal{F} \subseteq \Omega \\ \mathcal{F} \text{ is a } \sigma-\text{algebra}}} \mathcal{F}.$$

With this machinery, we can now define the *Borel measure* on the real numbers.

Definition 3.3 (Borel sets)**.** Consider the set of closed intervals $[a, b] \subset \mathbb{R}$. The *Borel sets* are members of the σ-algebra generated by closed intervals.

Note. We can actually construct a stratified *Borel hierarchy* as follows. Start from $\mathcal{F}_0$, the set of closed intervals in $\mathbb{R}$. Then, let $\mathcal{F}_1$ be the collection of all sets formed as countable unions or intersections of sets in $\mathcal{F}_0$, or their complements. This is already very complex, but we can similarly let $\mathcal{F}_2$ be the collection of all sets formed as countable unions, intersections, or complements of sets in $\mathcal{F}_1$. It turns out that $\mathcal{F}_0 \subsetneq \mathcal{F}_1 \subsetneq \mathcal{F}_2 \subsetneq \cdots$, and even the limit $\mathcal{F}_\omega$ is not a σ-algebra. You have to keep going up to the *first uncountable ordinal*, and then you reach the *Borel σ-algebra* $\mathcal{B} = \mathcal{F}_{\omega_1}$.

Definition 3.4 (Lebesgue measurable sets)**.** These exist and are more general than the Borel sets, but we won't talk too much more about them.

Note. These definitions are really general, which begs the question: are there sets that are not measurable? The answer is *yes* (assuming the axiom of choice), for example, the *Vitali sets*.

3.2 Random Variables

Intuitively, we all have some idea of what a *random variable* is — it varies randomly! However, we need to be slightly careful if we want to define this notion rigorously.

Definition 3.5 (Measurable function)**.** Given a set X equipped with a given σ-algebra $\Sigma \subseteq 2^X$, a function $f : X \to \mathbb{R}$ is called *measurable* if for all Borel sets[4] B,

$$f^{-1}(B) \in \Sigma.$$

[4] Technically, we can define this more generally for sets in the Lebesgue measure, but the difference is unimportant.

Note. For the rest of this course, we may implicitly assume that functions are measurable if not specified. It is *extremely difficult* to construct a non-measurable function, and they almost never occur in practice.

Definition 3.6 (Random variable)**.** A *random variable* X is a measurable function $X : \Omega \to \mathbb{R}$.

Random variables are so useful that we give them special notation. In particular, suppose that you have a random variable X, and you want to know the probability that its value lies between 1 and 3. We could write this rigorously in terms of events, i.e.,

$$P(X^{-1}([1,3])) = P(\{\omega \in \Omega \mid X(\omega) \in [1,3]\}).$$

However, this is a bit cumbersome, so we use the notation "$X \in B$" to mean the same thing as $X^{-1}(B)$. We can then write the above as

$$P(X^{-1}([1,3])) = P(X \in [1,3]) = P(1 \leq X \leq 3),$$

which seems much more natural to read.

Note. Joe philosophically comments that the reason *why statistics is a field* is because of random variables. They give us a common framework to talk about all kinds of events in completely different probability spaces, no matter their topology or other outlook. This unifies probability across a very broad range of disciplines. In contrast, fields like social network modeling have very fragmented theories with different journals and conventions, but they all draw upon ideas like *distributions* from statistics.

With that comment in mind, let's rigorously define the idea of a distribution.

Definition 3.7 (Distribution)**.** Given a random variable X, the *distribution* of X is a function $\mathcal{L}(X)$ that sends $B \mapsto P(X \in B)$. You can also write this compositionally as $\mathcal{L}(X) = P \circ X^{-1}$. The important property is that $\mathcal{L}(X)$ is a probability measure on the real line.

Going back to the philosophical note, this lets us define arbitrary probability spaces on complex events $(\Omega, \mathcal{F}, P)$. A random variable is just a means of projecting these complex world states into the real line, which creates a probability space $(\mathbb{R}, \mathcal{B}, \mathcal{L}(X))$. This is much easier to analyze!

3.3 Properties of Random Variables

We will state some important properties of random variables that will be useful to us.

Proposition 3.8. *If $X \sim Y$, and $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, then $g(X) \sim g(Y)$.*

Proof. For any Borel set $B \in \mathcal{B}$, observe that

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

□

In addition to being a useful, intuitive theorem (measurable functions preserve equality of measures), the proof gives us an instructive method of attack for proving similar theorems of this nature. Essentially, *preimages are really powerful*, especially when we have measurable functions.

Next we will prove an essential uniqueness theorem for probability theory. Although this won't let us prove *existence*,[5] it will still give us a powerful tool for some problems.

[5] If you also want to prove existence, Carathéodory's extension theorem works really well.

Proposition 3.9 (Dynkin's π-λ theorem)**.** *Call a collection of subsets $P \subseteq 2^\Omega$ a* π-system *if it is closed under set intersection. If two measures P, Q agree on a π-system P, then they also agree on all subsets in $\sigma(P)$, the σ-algebra generated by P.*

Proof. This involves some complicated analysis wizardry. See Section 2.10 of the book. □

Corollary 3.9.1 (CDFs are all you need)**.** *Any distribution is uniquely determined by its cumulative distribution function $F(x) = P(X \leq x)$.*

4 September 15th, 2020

Today is our final day focused primarily on measure theory, before we move on to random variables and representations.

4.1 A Couple Notes on Proofs

Note. Joe first gives us a reminder about the goal of proofs. They should be mathematically rigorous, but should not be encumbered by "obvious" details that mask the main idea. Writing mathematical proofs with clarity is a skill that you can develop with time.

Note. Also, Joe mentions that we should not be intimidated by his use of the words *trivial* or *obvious* in class. These words indicate that the ideas are simple enough to not require further justification once you understand them, not that you should feel bad if you don't immediately see the justification.

4.2 Working with π-λ

One of the most common ways we interact with random variables is through their CDF, which gives the measure of the variable on sets of the form $(-\infty, x] \in \mathbb{R}$. To illustrate the main idea of Proposition 3.9, let's provide some details. First, to get a taste, we prove a slightly more fundamental fact related to the uniqueness of CDFs.

Proposition 4.1. *If an function $X : \Omega \to \mathbb{R}$ satisfies $X^{-1}((-\infty, x]) \in \mathcal{F}$ for all $x \in \mathbb{R}$, then X is a random variable.*

Proof. Let $X : \mathcal{F} \to \mathcal{B}$ be an arbitrary function. Let $\mathcal{A}$ be the set of all Borel sets B such that

$$\mathcal{A} = \{B \in \mathcal{B} \mid X^{-1}(B) \in \mathcal{F}\}.$$

We know that $(-\infty, x] \in \mathcal{A}$ for all $x \in \mathbb{R}$. The key observation is that $\mathcal{A}$ is a σ-algebra, which we can directly verify by checking the three properties and mapping them back to properties of $\mathcal{F}$. Therefore, $\mathcal{A} = \mathcal{B}$. □

Definition 4.2 (Random vector)**.** A *random vector* is a collection of n random variables, which may or may not be independent. You can also see it as a measurable function $X : \Omega \to \mathbb{R}^n$, which defines the *joint distribution* of these variables. The *marginal distribution* of each variable is simply the composition of X with the projection map.

Marginal distributions give us some information, but this is lossy. Only the joint distribution gives us the full story of a random vector.

Now let's go back to talking about π-λ. What does the letter λ mean?

Definition 4.3 (λ-system)**.** A collection of subsets $L \subseteq 2^\Omega$ is called a λ-system if it satisfies the following properties:

- (Whole set) $\Omega \in L$,
- (Complement and difference) $A, B \in L, A \subseteq B \implies B \setminus A \in L$,
- (Restricted countable union) If $A_1, A_2, \ldots \in L$ with $A_1 \subseteq A_2 \subseteq \cdots$, then $\bigcup_{k=1}^{\infty} A_k \in L$.

It turns out that there is only one example of a λ-system that we really care about.

Example 4.4. Let P_1, P_2 be probability measures on $(\Omega, \mathcal{F})$. Let

$$L = \{A \in \mathcal{F} \mid P_1(A) = P_2(A)\}.$$

Then, L is a λ-system.

The above example can be easily checked by verifying the axioms; Joe skips this justification in class. Anyway, with this context, we can provide the general statement of Dynkin's theorem.

Lemma 4.5. *A family of sets is a σ-algebra if and only if it is both π and λ.*

Proposition 4.6 (Dynkin's π-λ, full form)**.** *If S is a π-system and L is a λ-system, and $S \subseteq L$, then $\sigma(S) \subseteq L$.*

Proof. Once again, the same tricky proof. We'll outline it in the next lecture. □

Some intuition for π-λ is that you can take a finite non π-system such as $S = \{\{1, 2\}, \{2, 3\}\}$, and this is not enough to guarantee uniqueness on the σ-algebra generated by S, which includes sets like $\{2\}, \{1, 2, 3\}$. But, at least in the countable case, you can use the π-system property to do disjointification/partitioning on Ω, which finishes the proof.

5 September 17th, 2020

We'll first go through the proof of π-λ, then finally begin talking about distributions.

5.1 Proof of the π-λ Theorem

Joe mentions this is one of his favorite proofs, so let's grind through it in all of its technical detail.

Proof of Proposition 4.6. Without loss of generality, let L be the smallest λ-system containing S.[6] The key idea will be to show that L is a σ-algebra, by showing that it is a π-system. In other words, it suffices to show that for all $A, B \in L$, we have $A \cap B \in L$.

To prove this result, we will rely on the following key claim. For some fixed $A_0 \in L$, we define a collection of sets $L(A_0) = \{B \in L \mid A_0 \cap B \in L\}$. Then $L(A_0)$ is a λ-system for any A_0.

The proof of the above claim is completely mechanical; just verify the axioms. Then, by the assumption that S is a π-system, we know that $S \subseteq L(A_0)$ whenever $A_0 \in S$, and since L is the smallest λ-system containing S, we in fact have $L(A_0) = L$. This means that whenever $A \in S$ and $B \in L$, we can conclude $A \cap B \in L$.

With this stronger fact, we can apply the lemma once again to get the stronger result that $S \subseteq L(A_0)$ whenever $A_0 \in L$. Then, applying the same logic again, this means that $L(A_0) = L$ for any $A_0 \in L$, as desired. □

Lemma 5.1. *Recall that the definition of two random variables X, Y being independent is that for all Borel sets $A, B \in \mathcal{B}$, you have $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$. You can show that this is equivalent to, for all $x, y \in \mathbb{R}$,*

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y).$$

Proof. Apply π-λ twice, judiciously. This can be generalized to n random variables. □

Lemma 5.2. *If g, h are measurable functions, then $X \perp\!\!\!\perp Y \implies g(X) \perp\!\!\!\perp h(Y)$.*

Proof. This is immediate by preimages. For any $A, B \in \mathcal{B}$,

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

□

5.2 Representations of Random Variables

If there are many distributions that might be useful in the general measure-theoretic framework, then why are there only a couple dozen of them that are commonly used and have names? Joe claims that this is because *representations* of distributions give us ways to model very complex spaces, and they're all connected.

Definition 5.3 (Bernoulli distribution)**.** The simplest distribution is the *Bernoulli*, which models a weighted coin toss. If $0 \leq p \leq 1$ and $Y \sim \text{Bern}(p)$, then $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$.

The expected value of the Bernoulli distribution is p, while the variance is $p(1 - p)$.

[6] This is valid because λ-systems are closed under intersection.

Definition 5.4 (Rademacher distribution)**.** The *Rademacher* distribution takes values $\{-1, +1\}$ with equal probabilities $\frac{1}{2}$ each.

If $Y \sim \text{Bern}(1/2)$, then you can also represent a Rademacher random variable by $R = 2Y - 1$. This immediately tells us that $\mathbf{E}[R] = 0$ and $\mathbf{Var}[R] = 1$.

Example 5.5. The position of a random walk on the real line, after n steps, can be modeled as a sum of n i.i.d. Rademacher random variables.

Definition 5.6 (Binomial distribution)**.** The *binomial* distribution $\text{Bin}(n, p)$ is the sum of n independent and identically distributed $\text{Bern}(p)$ random variables.

The mean of a binomial distribution is np, while the variance is $np(1-p)$.

Definition 5.7 (Uniform distribution)**.** The *uniform* distribution, written as $U \sim \text{Unif}$, is the distribution of equal density on the unit interval $[0, 1]$. It has the property that $P(U \in [a, b]) = b - a$ whenever $0 \le a \le b \le 1$. It can be represented in terms of i.i.d. $Y_1, Y_2, \ldots \sim \text{Bern}(1/2)$ by

$$U = \sum_{k=1}^{\infty} \frac{Y_k}{2^k}.$$

For brevity, we omit the measure theoretic details that the above dyadic construction is valid. Note that many sources represent uniform distributions on intervals as $\text{Unif}(a, b)$ instead, but Joe prefers to write $(b-a)U + a$. The uniform distribution satisfies $\mathbf{E}[U] = \frac{1}{2}$ and $\mathbf{Var}[U] = \frac{1}{12}$.

Definition 5.8 (Exponential distribution)**.** The *exponential* distribution is the distribution of random variables $X \sim \text{Expo}$ represented by $X = -\log U$, where $U \sim \text{Unif}$. Typically people also write $\text{Expo}(\lambda)$ as the exponential distribution with *rate* λ, which we choose to write as $\frac{1}{\lambda} \cdot \text{Expo}$.

The mean and variance are both 1. Note that in the above definition, we're using the syntactic convention of doing arithmetic on a distribution. This actually means that we draw random variables from that distribution, and do arithmetic on the values. Although unambiguous in most cases, Joe mentions that we should not write things like $\text{Expo} + \text{Expo}$, where the joint distribution is unclear.

Definition 5.9 (Gamma distribution)**.** The *gamma* distribution is the sum of independent exponentially distributed random variables. We call r the integer *rate parameter*. Then, $\text{Gamma}(r)$ is the distribution of

$$X_1 + X_2 + \cdots + X_r,$$

where X_j are i.i.d. and drawn from Expo. The probability density function can be written as

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

where $x > 0$, and $\Gamma : \mathbb{R}^+ \to \mathbb{R}^+$ is the gamma function.

6 September 22nd, 2020

Today we discuss reasoning by representation in more depth, and we introduce a fair number of useful, common distributions.

6.1 Probability Integral Transform

We previously defined the uniform distribution, which has measure on any interval proportional to the length of the interval. This is pretty simple. One important fact from statistics is that *there is no uniform distribution on* $\mathbb{R}$, for clear reasons. It turns out that if you know any distribution's CDF, you can generate it from a uniform distribution with the following theorem.

Definition 6.1 (Quantile function)**.** The *quantile function* of a distribution with CDF F is

$$F^{-1}(p) = \min\{x \mid F(x) \geq p\}.$$

When F is continuous and strictly increasing, F^{-1} is identical to the inverse. Otherwise, it serves as a sort of proxy that skips over regions with zero probability.

Proposition 6.2 (Probability integral transform)**.** *Let F be any CDF, with quantile function F^{-1}. If we sample $U \sim$ Unif, then it follows that $F^{-1}(U) \sim F$.*[7]

Proof. Note that $u \leq F(y)$ is the same as $F^{-1}(u) \leq y$, since F is a non-decreasing function. Therefore, the events $U \leq F(y)$ and $F^{-1}(U) \leq y$ are the same for any $y \in \mathbb{R}$, so

$$P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y).$$

□

Notice that this reminds us of the exponential distribution, which is in fact *defined* in a manner similar to the probability integral transform, as a function of a uniform random variable.

Example 6.3. To generate a Bernoulli random variable with probability p, we can generate a uniform random variable and pass it through the quantile function

$$F(u) = \begin{cases} 0 & \text{if } u \leq 1-p, \\ 1 & \text{if } u > 1-p. \end{cases}$$

This is consistent with our intuition about the uniform distribution.

It's worth mentioning that the uniform distribution is not necessarily special. Using a variant of the probability integral transform, we can generate a uniform from any continuous probability distribution, and by extension, we can generate any probability distribution from any continuous probability distribution.

Note. We can generate a normal distribution this way as well, by taking $\text{erf}^{-1}(U)$ for $U \sim$ Unif. However, this is not terribly appealing because the error function is not expressible in terms of elementary functions.

[7] This is also sometimes called *universality of the uniform.*

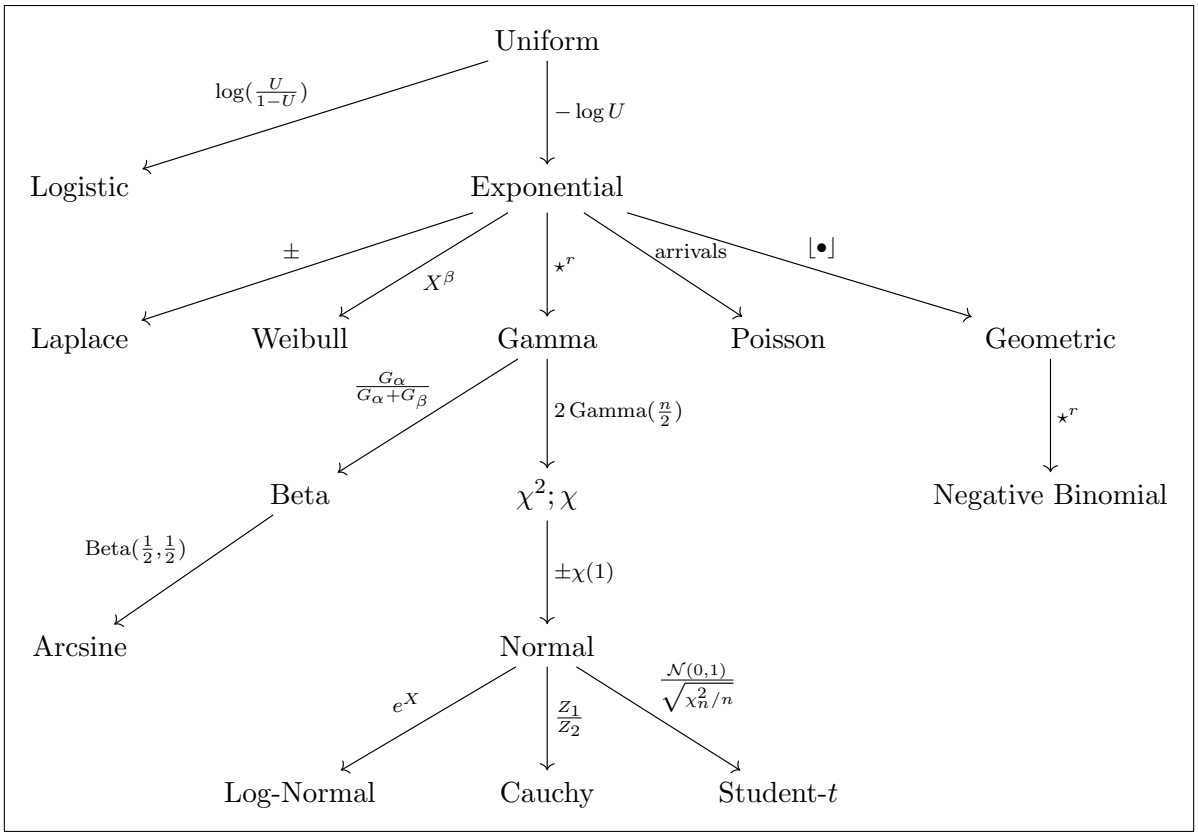

Figure 1: Derivation path of distribution representations.

Example 6.4 (Logistic distribution)**.** The *logistic distribution* has representation $\log(\frac{U}{1-U})$, where we sample $U \sim \mathrm{Unif}$. The quantile function of the distribution is called the *logit* function, which is

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

This maps a probability $(0,1) \mapsto \mathbb{R}$, and it can be thought of as the *log-odds* of a probability. For example, you can imagine predicting logits with a linear model (logistic regression), or a neural network (softmax and cross entropy). The CDF is the *sigmoid function*,

$$\sigma(y) = \mathrm{logit}^{-1}(y) = \frac{e^y}{1+e^y},$$

which can also be used as a nonlinearity in neural networks!

6.2 Reasoning By Representation

Now we can start introducing more distributions by strategy of representation. See Fig. 1 for a graphical overview of how we'll proceed, which can also serve as a useful reference.

Example 6.5. A useful fact about the gamma function from last time is that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Definition 6.6 (Chi-square distribution)**.** We define the *chi-square distribution* $\chi^2_n \sim 2\,\mathrm{Gamma}(\frac{n}{2})$. This is clearly a special case of the gamma distribution, but it's worth noting that you can interpret it as the sum of the squares of n i.i.d. standard normal random variables. The *chi distribution* with n degrees of freedom is defined similarly, by $\chi_n \sim \sqrt{\chi^2_n}$.

Finally, in a somewhat roundabout manner, we finally arrive at a definition of the normal distribution from the χ^2 distribution!

Definition 6.7 (Normal distribution)**.** The celebrated *standard normal distribution* is defined by $\mathcal{N}(0,1) \sim S \cdot \chi_1$, where $S \sim \text{Rad}$. We can scale this standard normal distribution to define a family of distributions with various means and variances, which we denote $\mathcal{N}(\mu, \sigma^2) \sim \sigma\mathcal{N}(0,1) + \mu$.

Example 6.8. $\chi_2^2 \sim Z_1^2 + Z_2^2$, where Z_1, Z_2 are i.i.d. $\sim \mathcal{N}(0,1)$. Also, $\chi_2^2 \sim 2\,\text{Expo}$.

Finally, here's our last collector's item today, which is often used in hypothesis testing.

Definition 6.9 (Student's t-distribution)**.** The *t-distribution* with $n-1$ degrees of freedom is represented by $t_n \sim \frac{Z}{\sqrt{V/n}}$, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi_n^2$. You can think of the denominator as the distribution of the empirical variance in a sample of size n.

Definition 6.10 (Cauchy distribution)**.** The *Cauchy distribution* is defined by $C \sim T_1 \sim Z_1/Z_2$, where Z_1, Z_2 are i.i.d. $\sim \mathcal{N}(0,1)$.

We now work on an exercise in breakout rooms. Joe mentions that this exercise is very difficult to solve with calculus, involving messy integrals, but it is surprisingly elegant when you attack it by means of representations!

Example 6.11. Find the expected value of $|T|$, where $T \sim T_n$.

Proof. We can write the representation of T as

$$|T| \sim \left|\frac{\mathcal{N}(0,1)}{\sqrt{\chi_n^2/n}}\right| \sim \frac{\chi_1}{\chi_n}\sqrt{n}.$$

From here, since the numerator and denominator are independent (sorry for the sloppy notation), we end up with a simpler expression:

$$\mathbf{E}\left[|T|\right] = \mathbf{E}\left[\chi_1\right] \cdot \mathbf{E}\left[\frac{1}{\chi_n}\right] \cdot \sqrt{n}.$$

This is still kind of messy, but it's broken down into much more manageable parts. For example, we can find $\mathbf{E}\left[\chi_1\right]$ by a quick search on Wikipedia. □

Definition 6.12 (Beta distribution)**.** The *beta distribution* with shape parameters $\alpha, \beta > 0$ is supported on $[0,1]$. Its representation is $\text{Beta}(\alpha,\beta) \sim \frac{G_\alpha}{G_{\alpha+\beta}}$, where $G_\alpha \sim \text{Gamma}(\alpha)$ and $G_{\alpha+\beta} \sim \text{Gamma}(\alpha+\beta)$, independently.

The beta distribution is often used as a conjugate prior for an unknown probability parameter. Its probability density function is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$, and we'll see some nice properties connecting it to the gamma distribution next lecture.

7 September 24th, 2020

We continue where we left off, with the beta distribution, and we also talk about basic properties of the normal distribution.

7.1 The Beta-Gamma Calculus

Here's an interesting fact that turns out to be a key property of the beta distribution.

Proposition 7.1 (Beta-Gamma)**.** *Suppose that you have independent random variables* $G_\alpha \sim \text{Gamma}(\alpha)$ *and* $G_\beta \sim \text{Gamma}(\beta)$*. Then by representations, we have that* $G_\alpha + G_\beta \sim \text{Gamma}(\alpha+\beta)$*, and* $\frac{G_\alpha}{G_\alpha+G_\beta} \sim \text{Beta}(\alpha, \beta)$*. The interesting fact is that*

$$\frac{G_\alpha}{G_\alpha + G_\beta} \perp\!\!\!\perp G_\alpha + G_\beta.$$

Proof. This fact comes from a straightforward calculation with Jacobians. Alternatively, you can also reason about this by relating both variables to a Poisson process and order statistics, which might help provide additional intuition. □

Note. Surprisingly, this fact actually completely characterizes the beta and gamma distributions, though nontrivial. This was formalized and proven in a theorem of Lukacs.

Joe remarks that he wants to emphasize the "Choose Your Favorite (CYF)" methodology. Whenever you start working on a new problem about distributions, try to pick whichever representation of said distributions gives you the most salient properties.

Example 7.2. Let $B_1 \sim \text{Beta}(\alpha, \beta)$ and $B_2 \sim \text{Beta}(\alpha+\beta, \delta)$, such that $B_1 \perp\!\!\!\perp B_2$. Using Proposition 7.1, we can choose the following construction for B_1 and B_2:

- Select independent $G_\alpha \sim \text{Gamma}(\alpha), G_\beta \sim \text{Gamma}(\beta), G_\delta \sim \text{Gamma}(\delta)$.
- Let $B_1 = \frac{G_\alpha}{G_\alpha+G_\beta}$ and $B_2 = \frac{G_\alpha+G_\beta}{G_\alpha+G_\beta+G_\delta}$.

We can verify in the above representation that indeed, $B_1 \perp\!\!\!\perp B_2$. Therefore the fractions cancel, and we can write $B_1 B_2 = \frac{G_\alpha}{G_\alpha+G_\beta+G_\delta} \sim \text{Beta}(\alpha, \beta+\delta)$.

Example 7.3. Let's try to compute the mean of the beta distribution. Note that independent random variables are *uncorrelated*, so by definition

$$\frac{G_\alpha}{G_\alpha + G_\beta} \perp\!\!\!\perp G_\alpha + G_\beta \implies \mathbf{E}\left[\frac{G_\alpha}{G_\alpha + G_\beta}\right] \mathbf{E}\left[G_\alpha + G_\beta\right] = \mathbf{E}\left[G_\alpha\right].$$

Rearranging this equation yields

$$\mathbf{E}\left[\frac{G_\alpha}{G_\alpha + G_\beta}\right] = \frac{\mathbf{E}\left[G_\alpha\right]}{\mathbf{E}\left[G_\alpha + G_\beta\right]} = \frac{\alpha}{\alpha+\beta}.$$

7.2 The Normal Distribution and Box-Muller

Recall some basic properties of the normal distribution. If $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $Z_1 \perp\!\!\!\perp Z_2 \implies Z_1 + Z_2 \sim \mathcal{N}(\mu_1 + \mu_1, \sigma_1^2 + \sigma_2^2)$. Other useful properties are that the normal distribution is invariant under rotations (e.g., $Z_1 + Z_2 \perp\!\!\!\perp Z_1 - Z_2$), and it is symmetric.

Proposition 7.4 (Box-Muller transform)**.** *If $U_1, U_2 \sim$ Unif and $U_1 \perp\!\!\!\perp U_2$, then define*

$$Z_1 = \sqrt{-2\log U_1}\cos(2\pi U_2),$$
$$Z_2 = \sqrt{-2\log U_1}\sin(2\pi U_2).$$

It follows that Z_1, Z_2 are i.i.d. $\sim \mathcal{N}(0,1)$.

Proof. Note that (Z_1, Z_2) has support on $\mathbb{R}^2$. Since the multivariate normal distribution is centrally symmetric, we can sample the angle $\theta \sim 2\pi\,\text{Unif}$, which is what U_2 is used for. Meanwhile, to get the radius, observe that $Z_1^2 + Z_2^2 \sim \chi_2^2 \sim 2\,\text{Gamma}(1)$, which motivates the use of $\sqrt{-2\log U_1}$. □

This transformation gives us an efficient way to sample i.i.d. normal random variables in the special case of a parallel processor (SIMD or GPU). However, the Ziggurat algorithm, a variant of rejection sampling, is more efficient on common processors. Taking NumPy's implementation as an example, see the current Ziggurat version, or the old Box-Muller version.

In addition to being computationally nice in some cases (avoiding code branches), the Box-Muller transform is also useful as a representation, which transforms many problems about the normal distribution into ones about trigonometric functions.

Example 7.5. If $U \sim$ Unif, then $\tan(2\pi U) \sim$ Cauchy.

7.3 Order Statistics

Order statistics generalize properties like the minimum, maximum, median, and quartiles.

Definition 7.6 (Order statistics)**.** Given a joint family of random variables $X_1, \ldots, X_n$, we call their *k-th order statistic* $X_{(k)}$ a variable reflecting to the k-th smallest of the values. In other words, the order statistics together are a rearrangement of the variables in increasing order:

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

Order statistics tend to be of general use in many cases. For example, insurance companies will often care about the worst of several events, and the probability of that happening.[8] We will first consider a case of interest: when X_i are i.i.d. exponential r.v.s.

Proposition 7.7 (Rényi representation)**.** *If $X_1, \ldots, X_n$ are i.i.d. $\sim$ Expo, then their order statistics are jointly distributed as*

$$\begin{aligned}
X_{(1)} &\sim \frac{1}{n}Y_1,\\
X_{(2)} &\sim \frac{1}{n}Y_1 + \frac{1}{n-1}Y_2,\\
X_{(3)} &\sim \frac{1}{n}Y_1 + \frac{1}{n-1}Y_2 + \frac{3}{n-2}Y_3,\\
&\vdots\\
X_{(n)} &\sim \frac{1}{n}Y_1 + \frac{1}{n-1}Y_2 + \frac{3}{n-2}Y_3 + \cdots + Y_n,
\end{aligned}$$

[8] Order statistics are also deeply related to a fallacy known as the *optimizer's curse*.

where $Y_1, Y_2, \ldots, Y_n$ *are i.i.d.* $\sim$ Expo.

Proof. This follows from induction and the memoryless property of the exponential distribution. □

One other interesting case to consider is when the distributions are uniform. In some sense, these are the two nicest order statistics to work with.

Proposition 7.8 (Uniform order statistics)**.** *If* $U_1, \ldots, U_n$ *are i.i.d.* $\sim$ Unif*, then their order statistics are jointly distributed as*

$$U_{(j)} = \frac{X_1 + \cdots + X_j}{X_1 + \cdots + X_{n+1}},$$

where $X_1, \ldots, X_{n+1}$ *are i.i.d.* $\sim$ Expo*. It immediately follows that the marginal distributions of the order statistics are* $U_{(j)} \sim \text{Beta}(j, n+1-j)$*.*

Proof. Joe notes that there's a nice proof of this due to Franklyn Wang, when viewed as related to the Rényi representation. Essentially, you map this to a transformed Poisson process. See the textbook for details. □

When dealing with order statistics for exponential distributions with different rates $\lambda_1, \ldots, \lambda_n$, the first order statistic is nice.[9] However, all of the other order statistics are unfortunately messy.

[9] It's not hard to show that this is distributed according to $\frac{1}{\lambda_1 + \cdots + \lambda_n}$ Expo.

8 September 24th, 2020

Today we formally introduce Poisson distribution and related Poisson process.

8.1 Poisson Processes

We introduce the *Poisson distribution*, which is a discrete probability distribution, as follows.

Definition 8.1 (Poisson distribution)**.** The Poisson distribution $\text{Pois}(\lambda)$ with rate parameter λ, supported on $\{0, 1, 2, \ldots\}$, is defined by the probability mass function

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

This distribution is deeply connected to the exponential and gamma distributions.

Definition 8.2 (Poisson process)**.** The *Poisson process* refers to the sequence of *arrival times* $T_1, T_2, \ldots \geq 0$, where the successive time differences $X_1 = T_1, X_2 = T_2 - T_1, X_3 = T_3 - T_2, \ldots$ are i.i.d. $\sim \lambda^{-1}\,\text{Expo}$. The marginal distribution of arrival times is

$$T_n = X_1 + X_2 + \cdots + X_n \sim \lambda^{-1}\,\text{Gamma}(n).$$

Furthermore, if $N_t = \#(\text{arrivals in } [0, t])$, then the two events $\{N_t \geq n\} = \{T_n \leq t\}$ are equivalent. This holds for general arrival processes, and we sometimes call this *count-time duality.*[10]

Proposition 8.3. *For any t in the Poisson process, $N_t \sim \text{Pois}(\lambda t)$.*

Proof. Observe from count-time duality that

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

Both of these latter probabilities can be expressed as a CDF of the gamma distribution. Although the incomplete gamma function is messy, applying integration by parts cracks the problem:

$$\begin{aligned} P(T_k \leq t) - P(T_{k+1} \leq t) &= \frac{1}{\Gamma(k)} \int_0^{\lambda t} e^{-x} x^{k-1}\,\mathrm{d}x - \frac{1}{\Gamma(k+1)} \int_0^{\lambda t} e^{-x} x^k\,\mathrm{d}x \\ &= \frac{1}{\Gamma(k)} \int_0^{\lambda t} e^{-x} x^{k-1}\,\mathrm{d}x + \frac{1}{\Gamma(k+1)} e^{-\lambda t}(\lambda t)^k - \frac{k}{\Gamma(k+1)} \int_0^{\lambda t} e^{-x} x^{k-1}\,\mathrm{d}x \\ &= \frac{e^{-\lambda t}(\lambda t)^k}{k!}. \end{aligned}$$

□

Corollary 8.3.1. *Given any fixed time interval of length t, the number of Poisson arrival events in that interval is distributed $\sim \text{Pois}(\lambda t)$. Furthermore, given two disjoint time intervals of any lengths, the number of Poisson arrival events in those intervals are independent.*

Proof. Use the memoryless property of the exponential distribution. □

Previously, we mentioned Poisson processes before through a connection with the order statistics of the uniform distribution. We formalize this below.

[10] This is Joe's invented name for the fact.

Proposition 8.4 (Conditional arrival times)**.** *If $T_{n+1} = t$, then the conditional joint distribution of $(T_1, T_2, \ldots, T_n)$ are the order statistics of i.i.d. uniform random variables multiplied by t, i.e.,*

$$[(T_1, \ldots, T_n) \mid T_{n+1} = t] \sim (tU_{(1)}, \ldots, tU_{(n)}),$$

where $U_1, \ldots, U_n \sim$ Unif.

Proof. This stems from distribution representations and the Beta-Gamma calculus. Observe that

$$\frac{T_k}{T_{n+1}} = \frac{X_1 + X_2 + \cdots + X_k}{X_1 + X_2 + \cdots + X_{n+1}}.$$

The right-hand side is precisely the representation from Proposition 7.8 for the joint distribution of uniform order statistics $U_{(k)}$. □

The above fact also yields a nice proof that $\text{Beta}(1, 1) \sim \text{Unif}$.

8.2 The Broken Stick Problem

Exercise 8.1. Cut a stick of unit length at n randomly chosen points. This will produce $n + 1$ segments. What is the distribution of the length of the shortest segment?

Proof. Use the order statistics of the uniform distribution. This tells us that in the joint distribution, the k-th cut point can be represented as

$$\frac{X_1 + \cdots + X_k}{X_1 + \cdots + X_{n+1}},$$

where $X_1, \ldots, X_{n+1}$ are i.i.d. $\sim$ Expo. Then, apply the Rényi representation of the exponential distribution, which tells us that

$$X_{(1)} \sim \frac{1}{n+1} Y_1; \quad X_{(2)} - X_{(1)} \sim \frac{1}{n} Y_2; \quad \ldots; \quad X_{(n+1)} - X_{(n)} \sim Y_{n+1};$$

where $Y_1, \ldots, Y_n$ are also i.i.d. $\sim$ Expo. Finally, we can conclude that the length of the shortest segment is simply distributed as

$$\frac{X_{(1)}}{X_{(1)} + \cdots + X_{(n+1)}} = \frac{\frac{1}{n+1} Y_1}{Y_1 + \cdots + Y_{n+1}} = \frac{1}{n+1} \text{Beta}(1, n).$$

This has mean $\frac{1}{(n+1)^2}$. □

Note. By a slight modification of the above argument, using linearity of expectation, we can see that the expected value of the length of the k-th largest segment is simply

$$\frac{1}{n+1} \left(\frac{1}{n+1} + \cdots + \frac{1}{k} \right).$$

In the next lecture, we will begin discussing expected value through Lebesgue integration!

9 October 1st, 2020

Today we will continue discussing Poisson processes and some of their nice properties. Then, we introduce the notion of expected value, which is defined in Chapter 4 of the textbook.

9.1 General Poisson Point Processes

We can generalize Poisson processes to general measure spaces under some weak assumptions.[11]

Definition 9.1 (Poisson point process)**.** A *Poisson point process* on a measure space (X, μ) with rate λ has the property that the number of points in a bound region $U \subset X$ is distributed according to a Poisson random variable with parameter $\lambda\mu(U)$.

Note that with this definition, we lose the interpretation of a Poisson process as having exponential arrival times. This only works for Poisson point processes on $\mathbb{R}^+$, which is what we have been working with so far. When we take Poisson processes over other measure spaces, there is no longer any notion of arrival time.

Example 9.2 (Poisson process on a circle)**.** We can define a Poisson process with rate λ on the unit circle S^1. Over any angle θ of the circle, written in radians, the number of points in that arc is distributed according to $\text{Pois}(\lambda\theta)$. The expected total number of points on the circle is $2\pi\lambda$.

Example 9.3 (2D Poisson process)**.** Consider the special case where X is a compact subset of $\mathbb{R}^2$, and μ is the Lebesgue (or Borel) measure. Then, we call this a *2D Poisson process*, and it has the property that the number of points in any two separate regions are independent, and the mean is proportional to the area of those regions.

We can arrive at an approximation for a 2D Poisson point process by subdividing our region into many small squares, then giving each square a finite and i.i.d. Bernoulli probability of having a point. As the number of squares gets larger, and each square gets smaller, our approximation gets closer to a true Poisson process.

9.2 Properties of the Poisson Distribution

Here, we introduce 3 key properties about the Poisson distribution, and we justify them through connection with a Poisson process. This might help glean some intuition for the deep connections between these two topics.[12]

Lemma 9.4 (#1)**.** *If* $X \sim \text{Pois}(\lambda_1)$*,* $Y \sim \text{Pois}(\lambda_2)$*, and* $X \perp\!\!\!\perp Y$*, then* $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$*.*

Proof. Consider a Poisson point process with rate λ_1, and another Poisson point process with rate λ_2. Then we can simply *superimpose* these processes together into a single process, combining the arrival times from both. It's easy to see that X is the number of arrivals in $[0, 1]$ for the first process, Y is the number of arrivals in $[0, 1]$ for the second process, and $X + Y$ becomes the number of arrivals in the superimposed process, which has rate $\lambda_1 + \lambda_2$. □

Lemma 9.5 (#2)**.** *If* $X \sim \text{Pois}(\lambda_1)$*,* $Y \sim \text{Pois}(\lambda_2)$*, and* $X \perp\!\!\!\perp Y$*, then the conditional distribution of* X *on* $X + Y = n$ *is given by* $\text{Bin}(n, \frac{\lambda_1}{\lambda_1+\lambda_2})$*.*

[11] Technically, these need to be a Radon measure for mathematical reasons.

[12] Joe calls these his "favorite" properties, particularly #3.

Proof. This is equivalent to the following fact about a Poisson process. Given a Poisson process with rate λ, the distribution of the arrival times $T_1, \dots, T_N$, conditioned on $N \sim \text{Pois}(\lambda t)$ equaling the number of point events in the interval $[0, t]$, is equivalent to the order statistics of N i.i.d. uniform random variables multiplied by t. □

Lemma 9.6 (#3)**.** *Consider the chicken-egg story, where a chicken lays $N \sim \text{Pois}(\lambda)$ eggs, which each hatch independently with probability p. Let X be the number of eggs that hatch, and let Y be the number of eggs that do not hatch. Then, $X \perp\!\!\!\perp Y$, $X \sim \text{Pois}(\lambda p)$, and $Y \sim \text{Pois}(\lambda(1-p))$.*

Proof. The proof of this result comes from LOTP, where we compute

$$P(X = x, Y = y) = P(X = x, Y = y \mid N = x + y) \cdot P(N = x + y).$$

After some algebraic manipulation, this eventually shows independence of $X \perp\!\!\!\perp Y$. As an interpretation in the corresponding Poisson point process, you can imagine starting with a process of rate λ, then *thinning* the process by *coloring* each point independently with probability p. The colored and uncolored points then form their own, independent, Poisson processes with rates λp and $\lambda(1-p)$. This can be seen as the reverse of superposition. □

9.3 Defining Integration and Expectation

When people typically define expected value, they usually do it separately either for discrete random variables (as a summation), or for continuous random variables (as an integral). This is technically okay, related to some analysis tricks such as the Lebesgue decomposition. However, it makes sense to have a more general definition of expected value that works for any distribution, even if it is partly discrete and partly continuous.

Definition 9.7 (Riemann integral)**.** Recall that the *Riemann integral* of $f : [a, b] \to \mathbb{R}$ is defined as a limit of *Riemann sums*

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{n\to\infty} \sum_{i=0}^{n-1} f(t_i)(x_{i+1} - x_i),$$

where $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, and for each i, $t_i \in [x_i, x_{i+1}]$.

This definition of Riemann integral clearly does not work when you have a discrete distribution, which does not have a finite PDF. The Riemann sums will not converge in this case, so we need something slightly more powerful.

Definition 9.8 (Riemann-Stieltjes integral)**.** The *Riemann-Stieltjes integral* of $f : [a, b] \to \mathbb{R}$ with respect to a non-decreasing *integrator* function $g : [a, b] \to \mathbb{R}$ is

$$\int_a^b f(x)\,\mathrm{d}g(x) = \lim_{n\to\infty} \sum_{i=0}^{n-1} f(t_i)(g(x_{i+1}) - g(x_i)),$$

where $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, and for each i, $t_i \in [x_i, x_{i+1}]$.

Note that this coincides with the ordinary Riemann integral when $g(x) = x$. This integral has the property that it works for computing the expected value of discrete distributions, since you can simply plug in the CDF (which is well-defined) as the integrator. It's also fairly easy to compute by hand, which makes it useful in practice. However, the strongest and most general integral, which is often used in proofs, is as follows.[13]

[13] We will only define the Lebesgue integral for random variables in probability spaces here, but you can generalize the definition to other functions on measure spaces.

Definition 9.9 (Lebesgue integral)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable. Then the *expected value* of X, denoted $\mathbf{E}[X]$ is defined by the following three-step construction:[14]

1. For indicator random variables, which are simply 1 on a bounded measurable set $S \in \mathcal{F}$ and 0 otherwise. Their expectation is the measure $P(S)$.

2. Extending to non-negative weighted sums of indicator random variables, called *simple random variables*. We do this by linearity of expectation.

3. Defining for non-negative random variables by taking the supremum over all dominated simple random variables X^*,
$$\mathbf{E}[X] = \sup_{X^* \leq X} \mathbf{E}[X^*].$$

4. Extending to general signed random variables by taking a partition $X = X^+ - X^-$ into positive and negative parts, and computing the integral for each separately.

We omit the remaining details, but these are the key ideas of the Lebesgue integral construction.

[14] Joe refers to this as *InSiPoD*, short for Indicator-Simple-Positive-Difference.

10 October 6th, 2020

Last week we introduced the Riemann-Stieltjes and Lebesgue integrals, for the purpose of defining what a random variable is. Today we'll continue by discussing them in more detail.

10.1 Riemann-Stieltjes and Lebesgue Integration

Recall our two main integral definitions, shown below:

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x \, \mathrm{d}F(x), \qquad \mathbf{E}[X] = \int_{\Omega} X(\omega) P(\mathrm{d}\omega).$$

The first is a mild generalization of our familiar Riemann integral from high school Calculus, while the second is the venerable Lebesgue integral, which is general enough to work on any measurable domain (not necessarily just the reals!). In general, when an integral is written, you can choose whichever definition as they are consistent where defined.

Example 10.1 (Indicator of $\mathbb{Q}$)**.** Consider the indicator function $I_{\mathbb{Q}} : \mathbb{R} \to \{0, 1\}$, which is 1 on all the rationals and 0 everywhere else. This function is not Riemann integrable (non-convergent) on any nonzero interval of the reals, yet it is Lebesgue integrable. In fact, because the rationals are countable,

$$\int_{\mathbb{R}} I_{\mathbb{Q}}(x) \lambda(\mathrm{d}x) = 0,$$

where λ is the Lebesgue measure.[15]

The single most important property of expectation is *linearity*.

Proposition 10.2 (Linearity of expectation)**.** *For r.v.s X and Y, $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.*

This is not an obvious statement, and proving it requires some work. We can also generalize to countably infinite sums of random variables, as linearity still holds under some mild regularizing assumptions. Joe uses this as an example of the difference between the Riemann and Lebesgue definitions of expected value. Compare the statement of linearity in both senses:

$$\int_{-\infty}^{\infty} t f_{X+Y}(t) \, \mathrm{d}t = \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x + \int_{\infty}^{\infty} y f_Y(y) \, \mathrm{d}y,$$
$$\int_{\Omega} (X + Y)(\omega) P(\mathrm{d}\omega) = \int_{\Omega} X(\omega) P(\mathrm{d}\omega) + \int_{\Omega} Y(\omega) P(\mathrm{d}\omega).$$

Either statement requires a formal mathematical proof, but the second statement (in terms of the Lebesgue integral) is much more intuitive to read, as the integrating factor ω is the same.

Example 10.3 (Simple random variables)**.** Consider a simple random variable $X = \sum_{j=1}^{n} a_j I_{A_j}$. We can usually write such a variable in *canonical form* by assuming that each subset A_j is distinct from the rest, which makes X essentially a collection of disjoint positive rectangles over the sample space Ω.

[15] Joe notes that we won't care much about weird cases like $\mathbb{Q}$, as they don't come up in practice. For example, in the real world, *all* of your measurements will be in $\mathbb{Q}$ due to finite precision. Here's a pun: "In this course, we care about hard work, not $I_{\mathbb{Q}}$."

For an additional clarification about Definition 9.9, consider the following equivalent description of the nonnegative case. This isn't written in the book yet, but there's a really clean formula for the simple random variables approximating any nonnegative random variable X. We can just take a monotone sequence of random variables:

$$X_n = \min\left(n, 2^{-n}\lfloor 2^n X \rfloor\right).$$

It's not hard to show that this is equivalent to the step in the definition of the Lebesgue integral that uses a supremum over simple random variables. Basically, all this does is cut off the values of X at n, then quantize it to the first n digits of its binary representation. However, this definition can be much easier to use in an actual computation.

Example 10.4 (Darth Vader rule)**.** For any nonnegative random variable Y, the following formula for the expectation holds:

$$\mathbf{E}[Y] = \int_0^\infty P(Y > y)\,\mathrm{d}y.$$

Proof. First we will show this for the Lebesgue definition of expected value. If Y is an indicator random variable I_A, then the right-hand side integral just becomes $P(A)$, which follows immediately. Next, if Y is simple, then we proceed simply by breaking up the variable into its canonical form and writing a double sum. After some manipulation (swapping the order of sums), this works. Finally, we can generalize to all nonnegative random variables by using the monotone convergence theorem, which lets us swap the order of lim and $\mathbf{E}$.

For completeness, we also sketch the proof when the left-hand side has $\mathbf{E}[Y]$ defined according to the Riemann-Stieltjes definition. Recall by definition that

$$\mathbf{E}[Y] = \int_{-\infty}^\infty y\,\mathrm{d}F(y) = \int_{-\infty}^\infty \int_0^y \mathrm{d}x\,\mathrm{d}F(y).$$

Writing it in this form, it's clear that this statement just becomes a consequence of Fubini's theorem. Swapping the order of the integrals yields our desired result:

$$\int_{-\infty}^\infty \int_0^y \mathrm{d}x\,\mathrm{d}F(y) = \int_0^\infty \int_x^\infty \mathrm{d}F(y)\,\mathrm{d}x = \int_0^\infty P(Y > y)\,\mathrm{d}y.$$

□

10.2 Convergence Theorems in Analysis

We will now cover some common convergence from analysis. Oftentimes in many branches of mathematics, we want to ask questions of the following commutative form: can I swap two operators in an expression? Oftentimes we draw this as a commutative square. In this case, we're concerned with the commutative square shown below:

$$\begin{array}{ccc} \{X_j\} & \xrightarrow[j\to\infty]{\lim} & X \\ \Big\downarrow{\scriptstyle \mathbf{E}} & & \Big\downarrow{\scriptstyle \mathbf{E}} \\ \{\mathbf{E}[X_j]\} & \xrightarrow[j\to\infty]{\lim} & \mathbf{E}[X]. \end{array}$$

In other words, when does $\lim_{j\to\infty} \mathbf{E}[X_j] = \mathbf{E}[\lim_{j\to\infty} X_j]$? It turns out that this intuitive statement holds most of the time, but there are counterexamples where it fails to hold.

Example 10.5 (Failure of convergence)**.** Consider the sequence of discrete random variables X_n, where X_n is n^2 with probability $1/n^2$ and 0 otherwise. By the Borel-Cantelli lemma, note that

$$\sum_{n=1}^{\infty} P(X_n > 0) = \sum_{n=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6} < \infty,$$

so with probability 1, at most finitely many of the X_n will be nonzero. This means that $X_n \to X$ almost surely, where $X = 0$ is in a Dirac delta distribution. However, now we have a counterexample, as $\mathbf{E}[X_n] = 1$ for all n, yet $\mathbf{E}[X] = 0$.

Despite this pessimistic example, under some mild assumptions we can prove that expected values and limits commute, using three so-called *convergence theorems*.

Proposition 10.6 (Monotone convergence theorem)**.** *If $0 \leq X_1 \leq X_2 \leq \cdots$, and $X_1, X_2, \ldots \to X$ in probability, then*

$$\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X].$$

Proposition 10.7 (Dominated convergence theorem)**.** *If there exists a random variable W such that $|X_n| \leq W$ for all n, $\mathbf{E}[W] < \infty$, and $X_1, X_2, \ldots \to X$ in probability, then*

$$\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X].$$

Corollary 10.7.1 (Bounded convergence theorem)**.** *If $|X_n| \leq c$ for some c, and $X_1, X_2, \ldots \to X$ in probability, then*

$$\lim_{n\to\infty} \mathbf{E}[X_n] = \mathbf{E}[X].$$

See Section 4.6 of the book for proofs of each of these theorems.

11 October 8th, 2020

Today we review convergence theorems a bit, for the purpose of providing us some analysis intuition.

11.1 Proof of Bounded Convergence

Recall the bounded convergence theorem from last lecture, i.e., Corollary 10.7.1. This is a theorem about random variables that converge *in probability*. It might be useful to have the definitions of various convergence types explicitly written down.

Definition 11.1 (Convergence in probability)**.** If $X_1, X_2, \ldots$ and X are random variables, then we say that $X_1, X_2, \ldots \to X$ *in probability* if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

Definition 11.2 (Almost sure convergence)**.** If $X_1, X_2, \ldots$ and X are random variables, then we say that $X_1, X_2, \ldots \to X$ *strongly*, or *almost surely* converges, if

$$P\Big(\lim_{n\to\infty} X_n = X\Big) = 1.$$

In general, almost sure convergence is a stronger condition than convergence in probability, which is likewise stronger than convergence in distribution. The bounded convergence theorem works generally for random variables that converge in probability. Let's walk through the proof.

Proof of Corollary 10.7.1. In the statement of the bounded convergence theorem, we assumed that $|X_n| \le c$ for all n. Let's first try to bound X as well. To do this, we will take a strategy of adding some slack to the variable.[16] For any n and $\epsilon > 0$, note that by a union bound,

$$\begin{aligned} P(|X| > c + \epsilon) &\le P(|X_n| > c \vee |X_n - X| > \epsilon) \\ &\le P(|X_n| > c) + P(|X_n - X| > \epsilon) \\ &= P(|X_n - X| > \epsilon). \end{aligned}$$

However, as $n \to \infty$, this probability just goes immediately to zero. We can view this as "taking the limit" on both sides, except the left-hand side doesn't actually contain the variable n in it. Therefore, by the squeeze theorem,

$$P(|X| > c + \epsilon) = \lim_{n\to\infty} P(|X| > c + \epsilon) \le \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0,$$

where the last step follows from the definition of convergence in probability. For the next part of our proof, consider $\mathbf{E}[|X_n - X|]$ for varying n. By the triangle inequality, since $|X_n|, |X| \le c$, we must have $|X_n - X| \le 2c$. Then,

$$\begin{aligned} \mathbf{E}[|X_n - X|] &\le 2cP(X_n - X > \epsilon) + \epsilon P(X_n - X \le \epsilon) \\ &\le 2cP(X_n - X > \epsilon) + \epsilon. \end{aligned}$$

For any ϵ, as $n \to \infty$, the first term on the right-hand side approaches zero. Therefore,

$$\limsup_{n\to\infty} \mathbf{E}[|X_n - X|] = 0.$$

Therefore, $\mathbf{E}[X_1], \mathbf{E}[X_2], \ldots \to \mathbf{E}[X]$. □

[16] Joe calls this technique *GSAS*: Give Some Arbitrary Slack.

Note that we choose to present the proof above, instead of the more general dominated convergence theorem, as that proof requires using machinery such as Fatou's lemma.

Exercise 11.1. Does the bounded convergence theorem still hold if we replace "converges in probability" with "converges in distribution" for $X_1, X_2, \ldots$? Joe mentions that he's not sure if this true, but he can't think of an easy counterexample at the moment.

11.2 Conditional Expectation

In undergraduate-level probability classes like Stat 110, we often define conditional expectation in the following form:

$$\mathbf{E}[Y \mid X = x] = g(x).$$

This focuses on the concrete values of X, rather than the underlying object. However, this kind of definition leads to some common fallacies like Borel's paradox. However, in graduate-level probability classes like this one, we will often condition based on a *random variable* instead:

$$\mathbf{E}[Y \mid X] = g(X).$$

The connection between these definitions is that $\mathbf{E}[Y \mid X = x] = \mathbf{E}[Y \mid X](x)$. Actually, what we're doing here is conditioning based on the σ-algebra *generated by* X, or in other words, the coarsest *filtration* of our underlying σ-algebra $\mathcal{F}$ that determines the value of X.[17]

Note. To be very rigorous about definitions, assume that $Y : \Omega \to \mathbb{R}$ is a random variable, and $\mathcal{G} \subseteq \mathcal{F}$ is a σ-subalgebra. Then $\mathbf{E}[Y \mid \mathcal{G}]$ is also a function $\Omega \to \mathbb{R}$, defined in terms of an *averaging operator* across all atomic sets in $\mathcal{G}$. In other words, we already have that Y is a $\mathcal{F}$-measurable function, but by applying a certain averaging map, we can make it $\mathcal{G}$-measurable, which is a stronger condition because $\mathcal{G}$ is coarser than $\mathcal{F}$. Mathematically, we have for all $G \in \mathcal{G}$ that

$$\int_G \mathbf{E}[Y \mid \mathcal{G}] \, \mathrm{d}P = \int_G Y \, \mathrm{d}P.$$

Therefore, the equation that $\mathbf{E}[Y \mid X] = g(X)$ is actually somewhat of an abuse of notation according to this σ-algebra definition, but it makes the definition much easier to think about!

For more intuition about conditional expectation, you can also think of it as a form of *projection*. This is reflected in the (albeit nonconstructive) definition below.

Definition 11.3 (Conditional expectation)**.** The *conditional expectation* $\mathbf{E}[Y \mid X]$ is the (almost surely) unique function $g(X)$ that *uncorrelates* $Y - g(X)$ from $h(X)$ for all all bounded, measurable functions $h : \mathbb{R} \to \mathbb{R}$. In other words,

$$\mathbf{E}[(Y - g(X))h(X)] = 0.$$

This makes sense intuitively, as you can pushforward the Lebesgue integral from the underlying σ-algebra $\mathcal{F}$ to the σ-subalgebra $\sigma(X) \subset \mathcal{F}$, which makes $h(X)$ a constant and $Y - g(X)$ zero. Anyway, this is the property we'd really like for conditional expectation to have. Let's now see if this definition is actually valid, i.e., showing existence and uniqueness! For what follows, let's assume (for the sake of convenience) that all r.v.s have finite variance.

[17] Therefore, you may see some papers writing this with the notation $\mathbf{E}[Y \mid \sigma(X)]$.

Definition 11.4 (Hilbert space)**.** A *Hilbert space* is a real or complex inner product space that is also a complete metric space with respect to its norm.

We care about this completeness condition because in function spaces, which are infinite-dimensional real vector spaces, we don't actually get completeness for free. Anyway, we could spend more time talking about Hilbert and Banach spaces, but that's the content of Math 114. Instead, we'll just state the theorem.

Proposition 11.5. *Zero-centered random variables, i.e., such that* $\mathbf{E}[X] = 0$*, form a Hilbert space under the covariance inner product*

$$\langle X, Y \rangle = \operatorname{Cov}(X, Y) = \mathbf{E}[XY].$$

This assumes that we consider two random variables to be equivalent if they are almost surely equal.

It's a well-known fact that quotient Hilbert spaces exist. Using some kind of argument along this form, you can essentially show with relative ease that conditional expectations exist and are unique. The details are omitted here in the lecture, as it's all measure theory.

Proposition 11.6 (Adam's law)**.** *For any random variables* X *and* Y*,*

$$\mathbf{E}[\mathbf{E}[Y \mid X]] = \mathbf{E}[Y].$$

Proof. This follows immediately from the conditional expectation property written above. In particular, if we set $h(X) = 1$, then the property reduces to

$$\mathbf{E}[Y - \mathbf{E}[Y \mid X]] = 0,$$

and the rest follows from linearity of expectation. □

Proposition 11.7 (Eve's law)**.** *For all random variables* X*,* Y*,*

$$\mathbf{Var}[Y] = \mathbf{E}[\mathbf{Var}[Y \mid X]] + \mathbf{Var}[\mathbf{E}[Y \mid X]].$$

Proof. Without loss of generality, assume that $\mathbf{E}[Y] = 0$. By Adam's law,

$$\mathbf{Var}[Y] = \mathbf{E}\left[Y^2\right] = \mathbf{E}\left[\mathbf{E}\left[Y^2 \mid X\right]\right].$$

Also, $\mathbf{E}[Y \mid X]$ has mean zero by assumption. Then, observe that

$$\begin{aligned}\mathbf{E}\left[\mathbf{E}\left[Y^2 \mid X\right]\right] &= \mathbf{E}\left[\mathbf{E}\left[Y^2 - \mathbf{E}[Y]^2 \mid X\right] + \mathbf{E}[Y \mid X]^2\right] \\ &= \mathbf{E}[\mathbf{Var}[Y \mid X]] + \mathbf{Var}[\mathbf{E}[Y \mid X]].\end{aligned}$$

□

That concludes a very brief foray into conditional expectation and some of its properties.

12 October 13th, 2020

First, let's talk about the midterm exam. The test-taking window will start on October 22nd, and it will last for 60 hours. As a result, there will be no class next Thursday. The test can be taken in any 3-hour block, although it has been written to be "reasonable" as a 75-80 minute exam to somewhat alleviate time pressure. The exam is open-book, open-note, and open-internet, but you may not ask questions or consult any other students. Submissions are in PDF format and can be handwritten or in LaTeX.

12.1 Conditional Covariance: ECCE

Recall that Eve's law allows us to calculate the variance using conditional distributions, by adding up the inter-group and intra-group variances. We can actually generalize this slightly from variances to *covariances* between two variables.

Definition 12.1 (Covariance)**.** The *covariance* of two r.v.s X and Y, denoted $\text{Cov}(X, Y)$, is defined as $\mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. Note that as a special case $\text{Cov}(X, X) = \mathbf{Var}[X]$.

Proposition 12.2 (ECCE)**.** *For any random variables X, Y, and Z,*

$$\text{Cov}(X, Y) = \mathbf{E}[\text{Cov}(X, Y \mid Z)] + \text{Cov}(\mathbf{E}[X \mid Z], \mathbf{E}[Y \mid Z]).$$

Proof. We will show this by simply applying Adam's law and linearity of expectation:

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] \\
&= \mathbf{E}[\mathbf{E}[XY \mid Z]] - \mathbf{E}[\mathbf{E}[X \mid Z]]\mathbf{E}[\mathbf{E}[Y \mid Z]] \\
&= \mathbf{E}[\mathbf{E}[XY \mid Z] - \mathbf{E}[X \mid Z]\mathbf{E}[Y \mid Z]] \\
&\quad + (\mathbf{E}[\mathbf{E}[X \mid Z]\mathbf{E}[Y \mid Z]] - \mathbf{E}[\mathbf{E}[X \mid Z]]\mathbf{E}[\mathbf{E}[Y \mid Z]]) \\
&= \mathbf{E}[\text{Cov}(X, Y \mid Z)] + \text{Cov}(\mathbf{E}[X \mid Z], \mathbf{E}[Y \mid Z]).
\end{aligned}$$

Alternatively, note that the above argument could be slightly simplified by assuming, without essential loss of generality, that $\mathbf{E}[X] = \mathbf{E}[Y] = 0$. □

As an aside, note that everything stated above has been about conditional expectations. This is because conditional expectations (which are just random variables) are much easier to rigorously talk about than conditional distributions. When we write $X \mid Z \sim \mathcal{N}(Z, 1)$, this is a statement about the conditional distribution of X, not a random variable called "$X \mid Z$" (which does not make sense). Defining conditional distributions rigorously requires some measure theory machinery,[18] which is not the focus of this course.

An interesting generalization of Adam's law, Eve's law, and ECCE is the law of total cumulance. This is not included in the textbook at the moment, but Joe might add it later. Anyway, it's beyond the scope of this course, as the laws written above cover 95% of cases.

Note. Borel's paradox, as mentioned in the book, is an issue when trying to define conditional probability when conditioning on events. It happens with continuous random variables, for example, conditioning on the events $X = X$ based on the equivalent formulations $X - Y = 0$, and $\frac{X}{Y} = 1$. The issue is that we are conditioning on a *event of measure zero* in both cases. Because of the obvious issues, conditioning on events is outside the scope of this course, and we will only condition on r.v.s and σ-algebras, which is well-defined.

[18] For more info and some juicy *pushforwards*, see regular conditional probability.

12.2 Moment Generating Functions

In your typical undergraduate-level probability class, you probably talked about moment generating functions, but perhaps not on generating functions in general.[19] We'll talk about this briefly.

Example 12.3 (Making change)**.** Suppose that you wanted to know how many ways you can make change for 50 cents, given coins of denominations 1, 5, 10, 25, 50 cents. You could do this by writing out the possibilities, but this is tedious. You could also use dynamic programming (Knapsack). One formalism from combinatorics that might help here though, is a *generating function*. Write

$$\begin{aligned} p(t) &= (1+t+t^2+t^3+\cdots)(1+t^5+t^{10}+t^{15}+\cdots)(1+t^{10}+t^{20}+t^{30}+\cdots) \\ &\quad (1+t^{25}+t^{50}+t^{75}+\cdots)(1+t^{50}+t^{100}+\cdots) \\ &= \frac{1}{(1-t)(1-t^5)(1-t^{10})(1-t^{25})(1-t^{50})}. \end{aligned}$$

Then, in the formal power series expansion for $p(t)$, the coefficient of x^k is precisely the number of ways to make k cents, using these types of coins. Mathematically, this is also $\frac{1}{k!}p^{(k)}(0)$. Here t is just a "bookkeeping device" for the sequence of values in the coefficients.

With that informative example, which also shows the two-pronged interpretation of generating functions as either *formal power series* or *convergent functions*, we are now ready to define the notion of a generating function.

Definition 12.4 (Generating function)**.** Given a sequence $(a_0, a_1, a_2, \ldots)$, a *generating function* for the sequence is a power series containing this sequence in its coefficients. There are two kinds:

- The *ordinary generating function* $p(t) = \sum_{n=0}^{\infty} a_n t^n$, and
- The *exponential generating function* $p(t) = \sum_{n=0}^{\infty} \frac{a_n t^n}{n!}$.

Generally we prefer working with the exponential kind in statistics, as it is more likely to converge.

Definition 12.5 (Moment generating function)**.** The *moment generating function* of a random variable X is the exponential generating function of the moments. We denote this by

$$M_X(t) = \mathbf{E}\left[e^{tX}\right] = \mathbf{E}\left[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{\mathbf{E}\left[X^n\right] t^n}{n!}.$$

This function only exists when $M_X(t) < \infty$ for all t in some open neighborhood of 0. Under this assumption, we are allowed to swap the expectation and sum in the last step, due to dominated convergence. This is because

$$\left|\sum_{n=0}^{m} \frac{X^n t^n}{n!}\right| = \sum_{n=0}^{m} \frac{|X|^n |t|^n}{n!} \leq e^{|tX|} \leq e^{tX} + e^{-tX}.$$

The last expression above has finite expectation, by our assumption, so it meets the necessary requirements for dominated convergence.

Now that we've rigorously defined MGFs, let's see some useful properties.

[19] For the definitive text on generating functions, see Herbert Wilf's *generatingfunctionology*.

Proposition 12.6 (MGF of independent sum)**.** *If X, Y are r.v.s and $X \perp\!\!\!\perp Y$, then*

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

Proposition 12.7 (Uniqueness of MGFs)**.** *If X, Y are r.v.s with moment generating functions and $M_X(t) = M_Y(t)$ on some open neighborhood of the origin, then $X \sim Y$.*

Example 12.8 (MGF of the normal distribution)**.** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$M_X(t) = e^{t\mu + t^2\sigma^2/2}.$$

Combined with the previous two propositions, this immediately implies that the sum of independent normals is also normally distributed.

Example 12.9 (No MGF for log-normal distribution)**.** If $Y \sim e^z$ where $z \sim \mathcal{N}(0, 1)$, then all of the moments of Y are defined, as

$$\mathbf{E}\left[Y^n\right] = \mathbf{E}\left[e^{nZ}\right] = M_Z(n) = e^{n^2/2}.$$

Therefore, all of the moments are defined. However, the MGF of Y *does not exist*. We can show that the expectation $\mathbf{E}\left[e^{tY}\right]$ does not converge on any neighborhood of the origin.

Definition 12.10 (Joint moment generating function)**.** Given two random variables X and Y, or in general n variables, the *joint MGF* of X and Y is defined as the function

$$M_{X,Y}(s, t) = \mathbf{E}\left[e^{sX+tY}\right].$$

In addition to the usual moments, the joint MGF also generates joint moments such as the covariance when X and Y are centered. It also fully describes the *joint* distribution of (X, Y), meaning that $X \perp\!\!\!\perp Y$ if and only if their joint MGF factors into the marginal variants.

Finally, one problem with moment generating functions illustrated by an earlier example is *convergence*. To fix this issue somewhat, there is a variant of MGFs based on the *Fourier transform* rather than the *Laplace transform*, which is guaranteed to always exist.

Definition 12.11 (Characteristic function)**.** The *characteristic function* $\varphi_X : \mathbb{R} \to \mathbb{C}$ for a random variable X is defined as

$$\varphi_X(t) = \mathbf{E}\left[e^{itX}\right].$$

This also has uniqueness properties, as we will describe next lecture, but there is also the nice fact that its value always lies within the unit disk (as it's really a convex combination of points on the unit circle).

13 October 15th, 2020

Today, class is starting 15 minutes early due to a last-minute schedule conflict for Joe. Therefore, we will have a brief self-contained topic (cumulants) for the first 15 minutes, before going back to the main topic.

13.1 Cumulants

As a brief aside, recall the definition of the characteristic function. This has the nice property that it always exists, but it is not always smooth. The moment generating function has a redeeming property that it are *infinitely differentiable*, i.e., $\mathcal{C}^\infty$ at 0, and their existence implies that all the moments of the distribution exist. This makes them very useful in many cases.

Proposition 13.1. *If a random variable X has moment generating function $M_X(t)$, then the k-th moment of X exists for all k, and*

$$\mathbf{E}\left[X^k\right] = M^{(k)}(0).$$

Proof. This can be justified by dominated convergence, which can be used to do differentiation under the integral sign.[20] Essentially,

$$M_X'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{E}\left[e^{Xt}\right] = \mathbf{E}\left[Xe^{Xt}\right].$$

Doing this repeatedly lets you illustrate that

$$M_X^{(k)}(t) = \frac{\mathrm{d}^k}{\mathrm{d}t^k}\,\mathbf{E}\left[e^{Xt}\right] = \mathbf{E}\left[X^k e^{Xt}\right],$$

and the result follows. In general, the MGF is always infinitely differentiable, arguing from dominated convergence once again (i.e., MGF existence is stronger than existence of all moments), so this proposition is valid. □

Recall that if two random variables X and Y are independent, then the moment generating function of their sum $X+Y$ is simply the *product* of their individual moment generating functions. In other words, $M_{X+Y}(t) = M_X(t)M_Y(t)$. What if we wanted to turn this product back into a sum?

Definition 13.2 (Cumulant generating function (CGF))**.** The *cumulant generating function* of a random variable X, defined wheneever the MGF exists, is defined by

$$K_X(t) = \log M_X(t) = \sum_{r=1}^{\infty} \frac{\kappa_r}{r!} t^r.$$

The coefficients κ_r of this power series are called *cumulants*.

You can derive formulas for the few cumulants by using the power series expansion for $\log(1+x)$, since the moment generating function satisfies $M_X(0) = 1$. This power series looks like

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \pm \cdots.$$

[20] Joe calls this by the acronym DUThIS.

Example 13.3 (Cumulants)**.** The first four cumulants are:

1. $\kappa_1 = \mathbf{E}[X]$. This is the *expectation*.
2. $\kappa_2 = \mathbf{Var}[X]$. This is the *variance*.
3. $\kappa_3 = \mathbf{E}\left[(X-\mu)^3\right]$. This is the *third central moment*, which is connected to *skewness*.
4. $\kappa_4 = \mathbf{E}\left[(X-\mu)^4 - 3\,\mathbf{Var}[X]^2\right]$. This is the *excess kurtosis* multiplied by $\mathbf{Var}[X]^2$.

We'll come back to cumulants and generating functions later, when we discuss the central limit theorem. However, we can still state some basic facts. One nice property of cumulants is that they are easy to compute, partially due to the following fact.

Proposition 13.4 (Cumulants are additive)**.** *If $X \perp\!\!\!\perp Y$, then $K_{X+Y}(t) = K_X(t) + K_Y(t)$.*

This is a vast generalization of the fact that variances are additive for independent random variables, as the variance is just the second cumulant. Cumulants also give you a good way to find central moments of distributions like the Poisson, as their generating functions are simple.

Example 13.5 (Cumulance of Poisson)**.** The MGF of the Poisson distribution is $e^{\lambda(e^t-1)}$, and therefore the CGF is

$$K(t) = \lambda(e^t - 1) = \lambda\left(x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots\right).$$

This means that all of the cumulants of the Poisson distribution are equal to λ.

Example 13.6 (Cumulance of Normal)**.** The MGF of the normal distribution is $e^{\mu t + \sigma^2 t^2/2}$, and therefore the CGF is

$$K(t) = \mu t + \frac{\sigma^2 t^2}{2}.$$

Therefore, the first two cumulants are nonzero, equal to the mean μ and variance σ^2. The rest of the cumulants are all zero.

Note. As an aside, it turns out that the normal distribution is the *only* nontrivial distribution that has a finite number of nonzero cumulants.

13.2 Characteristic Functions

We want to talk a little bit more about characteristic functions. Recall from earlier today that we mentioned characteristic functions are not always smooth, but you can approximate their values based on how many derivatives are defined, using a power series. We can show characteristic functions are always defined by applying Jensen's inequality,

$$|\mathbf{E}[itX]|^2 \leq \mathbf{E}\left[\left|e^{itX}\right|^2\right] = 1.$$

as the complex magnitude function $|x|^2 = x\overline{x}$ is convex. This should be consistent with your intuitions about the Fourier transform, if you are familiar with that operator. We will not compute many characteristic functions by hand in this course, as the integrals can involve complex analysis machinery like the residue theorem. Still, it can be useful to see some examples.

Example 13.7 (Characteristic of Cauchy)**.** The characteristic function of the Cauchy distribution, with PDF $f(x) \propto 1/(1+x^2)$, is

$$\varphi_X(t) = e^{-|t|}.$$

This should remind you of the PDF of the Laplace distribution. Indeed, the characteristic function of the Laplace distribution is also a scaled version of the Cauchy PDF; this is a consequence of the inversion property of the Fourier transform.

Example 13.8 (Characteristic of Normal)**.** The characteristic function of the normal distribution, where $X \sim \mathcal{N}(\mu, \sigma^2)$, is

$$\varphi_X(t) = M_X(it) = e^{i\mu t - \sigma^2 t^2/2}.$$

Note that the above is a slight abuse of notation, as the moment generating function has real domain, but it works out anyway if we pretend that it's a Laplace transform and extend to $\mathbb{C}$.

13.3 The Multivariate Normal Distribution

We would like to now introduce the multivariate normal (MVN) distribution. This is a distribution with a lot of interesting properties, and you will learn a lot about this in courses like Statistics 230 and CS 229r (Spectral Graph Theory). However, for this course, it suffices to just have a basic understanding.

Definition 13.9 (Random vector)**.** A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of random variables. These components may or may not be independent. Generally, random vectors have some distribution over $\mathbb{R}^n$.

Definition 13.10 (Covariance matrix)**.** Given a random vector $\mathbf{Y}$, we define the *covariance matrix* to be the $n \times n$ matrix of variances and covariances between pairwise components. In other words,

$$\operatorname{Cov}(\mathbf{Y}, \mathbf{Y}) = \mathbf{E}\left[(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])(\mathbf{Y} - \mathbf{E}[\mathbf{Y}])^T\right] = \begin{bmatrix} \mathbf{Var}[Y_1] & \operatorname{Cov}(Y_1, Y_2) & \cdots & \operatorname{Cov}(Y_1, Y_n) \\ \operatorname{Cov}(Y_2, Y_1) & \mathbf{Var}[Y_2] & \cdots & \operatorname{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(Y_n, Y_1) & \operatorname{Cov}(Y_n, Y_2) & \cdots & \mathbf{Var}[Y_n] \end{bmatrix}.$$

Proposition 13.11 (Semidefinite covariance matrix)**.** *The covariance matrix is always positive semidefinite. In other words, it is symmetric, and all eigenvalues are nonnegative.*

We can also define $\operatorname{Cov}(\mathbf{X}, \mathbf{Y})$ similarly. One can show that this is a *bilinear* operator.

Proposition 13.12 (Linearity of vector expectation)**.** *If A is a linear operator (matrix), b is a vector, and $\mathbf{Y}$ is a random vector, then*

$$\mathbf{E}[A\mathbf{Y} + \mathbf{b}] = A\,\mathbf{E}[\mathbf{Y}] + \mathbf{b}.$$

The proof of this is left as an exercise. Generally, you want to stick to vector and matrix notation whenever possible when proving facts about random vectors, as it will make arguments much cleaner (and *natural*). You should avoid explicit sums over bases whenever possible.

Proposition 13.13 (Covariance matrices are conjugate)**.** *If A is a linear transformation and $\mathbf{X}$ is a random vector, then*

$$\operatorname{Cov}(A\mathbf{X}, A\mathbf{X}) = A \operatorname{Cov}(\mathbf{X}, \mathbf{X}) A^T.$$

With all of this machinery for talking about multivariate distributions, it's time to actually create some instances. The nicest multivariate distribution is unambiguously the multivariate normal.[21] It turns out that there are *many ways* you might attempt to construct a multivariate normal, but all of them will end up generating the same distribution.

Definition 13.14 (Matrix square root)**.** If $\Sigma \succeq 0$, then there exists at least one matrix A such that $\Sigma = A^T A = AA^T$. In general, there can exist many A, but they will all be equivalent up to multiplication by an orthogonal matrix.

One can construct matrix square roots explicitly by using the *Cholesky decomposition* algorithm.

Definition 13.15 (Multivariate normal distribution)**.** A *multivariate normal distribution* with mean μ and covariance matrix Σ is defined by representation. If $\mathbf{Z} = (Z_1, \ldots, Z_n)$ is a standard multivariate normal where $Z_i \perp\!\!\!\perp Z_j$ are pairwise independent and $Z_i \sim \mathcal{N}(0, 1)$, then

$$\mathbf{X} = A\mathbf{Z} + \mu$$

is a multivariate normal distributed according to $\mathbf{X} \sim (\mu, \Sigma)$, where $\Sigma = AA^T = A^T A$.

Note. Observe that since the standard multivariate normal is rotationally symmetric, multiplying $\mathbf{Z}$ by any orthogonal rotation matrix does not affect the joint distribution. This means that the above definition is unambiguous with respect to the matrix square root, and multivariate normals are indeed characterized by their covariance matrices.

Note that if $\mathbf{X}$ is multivariate normal, then any projection $\mathbf{b}^T\mathbf{X}$ is a linear combination of independent standard normals plus some mean $\mathbf{b}^T\mu$, so it is itself a univariate normal. This lends itself to an alternative definition.

Proposition 13.16 (Multivariate normal by projections)**.** *A random vector $\mathbf{Y}$ is multivariate normal if and only if every nonzero linear combination of its components, $\mathbf{b}^T\mathbf{Y}$ where $\mathbf{b} \neq 0$, is distributed according to a univariate normal.*

Another interesting fact is that within a multivariate normal distribution, uncorrelated implies independent. This can be most easily shown through representations.

[21] The second-nicest one is the multinomial distribution, but this isn't too much different from the binomial.

14 October 21st, 2020

Today we will finish discussing some useful properties of the multivariate normal. The midterm is on Thursday (so no class then), and it will be timed. Generally the problems will require some tricky thinking, but there should always be a clever solution that does not require tedious calculation.

14.1 More on the Multivariate Normal

Recall that a multivariate normal distribution is defined by $\mathbf{Y} \sim A\mathbf{Z} + \mu$, where $\mathbf{Z}$ is a vector of i.i.d. standard normal random variables. This is denoted $\mathcal{N}(\mu, \Sigma)$, where $\Sigma = A^T A$. If you take the Jacobian, you can compute the probability density function, which in n dimensions is

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}.$$

Joe makes the following optimistic point about reducing high-dimensional random vectors into ordinary random variables.

Proposition 14.1 (Cramér-Wold device)**.** *Given a finite-dimensional random vector $\mathbf{X}$, the joint distribution of $\mathbf{X}$ is uniquely determined by its projections onto 1-dimensional spaces. In other words, knowing the marginal distribution of $\mathbf{t}^T\mathbf{X}$, for every fixed $\mathbf{t}$, is enough.*

The above proposition follows from the fact that joint characteristic functions determine multivariate distributions (a hard but standard fact from analysis). This is because the joint characteristic function is just

$$\varphi_{\mathbf{X}}(\mathbf{t}) = \mathbf{E}\left[e^{i\mathbf{t}^T\mathbf{X}}\right].$$

In particular, $\mathbf{t}^T\mathbf{X}$ occurs in the exponent, so values of the characteristic function are completely determined from marginal distributions of projections of $\mathbf{X}$.

Example 14.2 (MGF of multivariate normal)**.** Recall that the moment generating function of a univariate $\mathcal{N}(\mu, \sigma^2)$ normal distribution is

$$e^{t\mu + \sigma^2 t^2/2}.$$

The joint moment generating function of a multivariate normal $\mathcal{N}(\mu, \Sigma)$ is analogously

$$e^{\mathbf{t}^T(\mu + \frac{1}{2}\Sigma\mathbf{t})}.$$

This is because if we let $\mathbf{W} \sim \mathcal{N}(\mu, \Sigma)$ and consider the projection $\mathbf{t}^T\mathbf{W}$, we get $\mathbf{E}\left[\mathbf{t}^T\mathbf{W}\right] = \mathbf{t}^T\mu$ and $\mathbf{Var}\left[\mathbf{t}^T\mathbf{W}\right] = \mathbf{Var}\left[\mathbf{t}^T(\Sigma^{1/2}\mathbf{W} + \mu)\right] = \mathbf{Var}\left[\mathbf{t}^T\Sigma^{1/2}\mathbf{Z}\right] = \mathbf{t}^T\Sigma\mathbf{t}$. Therefore, the distribution of each projection of a multivariate normal is also normal, so we can compute the joint MGF from the univariate MGF.

Example 14.3 (Closure properties of MVN)**.** The multivariate normal distribution has many nice closure properties, such as:

- If you take a linear combination or shift of multivariate normals, it is also multivariate normal.
- Any vector of projections (i.e., projection matrix) is also multivariate normal.
- The conditional distribution of a multivariate normal is also multivariate normal.[22]

[22] See this page for more info, including the formula involving Schur complements.

It turns out that these closure properties are really useful for applications like Kalman filtering (Branislav's favorite!), where we can exactly compute posteriors due to closure.

Here's a really important fact about multivariate normal distributions.

Proposition 14.4. *Within a multivariate normal distribution, consider any two (possibly vector) projections $\mathbf{Y}_1$ and $\mathbf{Y}_2$. Then, if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated, they are also independent.*

Proof. Consider the multivariate normal random vector $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \mathcal{N}(\mu, V)$. We have

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

as a block matrix, where V_{11} and V_{22} are the covariance matrices of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively. Now we can simply observe that $V_{12} = \operatorname{Cov}(Y_1, Y_2) = 0$, and $V_{21} = \operatorname{Cov}(Y_2, Y_1) = 0$, which is the assumption we made about the vectors being uncorrelated. Then, the matrix is a diagonal block and factorizes into a direct sum of invariant subspaces, as desired. □

Proposition 14.5. *Suppose that $\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$ is multivariate normal with covariance matrix*

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

Also assume that $\mathbf{E}[\mathbf{Y}_1] = \mu_1$ and $\mathbf{E}[\mathbf{Y}_2] = \mu_2$. Then,

$$\mathbf{Y}_2 \mid \mathbf{Y}_1 \sim \mathcal{N}(\mu_2 + V_{21}V_{11}^{-1}(\mathbf{Y}_1 - \mu_1), V_{22} - V_{21}V_{11}^{-1}V_{12}).$$

In particular, the conditional distribution is still normal, its mean is linear with respect to $\mathbf{Y}_1$, and its variance is constant! This is related to formulas from linear regression.[23]

14.2 Example Problem: Socks in a Drawer

For a complete non sequitur, we will now talk about one of Joe's favorite problems. This will serve as a minor review of order statistics for the upcoming midterm.

Exercise 14.1. Suppose that you have n pairs of socks, and each pair is different. You randomly pull socks out of your drawer, one at a time, until you have a matching pair after N socks. Find the expected value of N.

Proof. We choose an embedding of the problem in continuous time. Assume that we continue the task, drawing all of the socks until the drawer is empty (even if we reach a pair after sock N). Each of the $2n$ socks will be assigned a time i.i.d. uniform in $[0, 1]$.

Let the interarrival times be $X_1, X_2, \ldots, X_{2n+1}$, and let T be the time when we draw our first matching pair. Then we can write, using uniform order statistics, that

$$T = \sum_{j=1}^{N} X_j \sim \text{Beta}(N, 2n + 1 - N).$$

Note that by Adam's law and linearity of expectation,

$$\mathbf{E}[T] = \mathbf{E}[\mathbf{E}[T \mid N]] = \mathbf{E}[X_1]\,\mathbf{E}[N] = \frac{\mathbf{E}[N]}{2n+1}.$$

[23] In particular, in a multivariate normal distribution, the best predictor of $\mathbf{Y}_2$ given $\mathbf{Y}_1$ is a linear regression.

We're going to find T in a slightly different way. For each color $i \in \{1, \dots, n\}$, let T_i be the time when we complete the pair of socks with color i, so that $T = \min(T_1, \dots, T_n)$. Note that each individual sock is independent, so all of the T_i are i.i.d. distributed according to the maximum of two independent uniforms. Recall that this is $\sim \text{Beta}(2, 1) \sim \sqrt{\text{Unif}}$.

We can then write that $T = \min(\sqrt{U_1}, \dots, \sqrt{U_n}) = \sqrt{\min(U_1, \dots, U_n)}$, where the U_i are jointly distributed as i.i.d. uniform. Therefore, $T \sim \sqrt{\text{Beta}(1, n)}$. We can verify by LOTUS that

$$\begin{aligned}
\mathbf{E}[T] &= \int_0^1 \frac{x^{1/2}(1-x)^{n-1}}{\text{B}(1,n)}\,\text{d}x \\
&= \frac{\text{B}(3/2, n)}{\text{B}(1,n)} \\
&= \frac{\Gamma(3/2)\Gamma(n+1)}{\Gamma(1)\Gamma(n+3/2)} \\
&= \frac{n!}{(3/2)(5/2)\cdots((2n+1)/2)} \\
&= \frac{4^n(n!)^2}{(2n)!(2n+1)}.
\end{aligned}$$

Therefore, matching up our two expressions for $\mathbf{E}[T]$, we get $\mathbf{E}[T] = \frac{4^n(n!)^2}{(2n)!}$. □

Note that the above problem does not appear to be related to continuous distributions at all (very combinatorial), yet we found a very natural solution by using a continuous embedding!

15 October 27th, 2020

Today marks a little over the halfway point of the course, as we're done with the midterm. Grading will take some time due to logistics. We'll begin Chapter 9 today, on inequalities, and Chapters 10–14 are all posted on Canvas.[24] Overall, the chapters are fairly short, but they're also analysis-heavy and packed with information.

15.1 Intro to Inequalities

In life, numbers are almost never equal. Philosophically, given that the general position of values is to be *unequal*, it can sometimes be lot more important to talk about when $a \leq b$ or $a \geq b$, rather than the specific case when $a = b$. This is why inequalities are important.

The general setup is that you have some probability problem that you'd like to solve, but it's too hard! Yet you have to do something, since you really want to solve it. So you can approach the problem by considering a couple of strategies:

1. **Simulate it by writing code.**

   Simulation is interesting because it can help inform your conjectures and hypotheses. After you write a little bit of code, you will surely get some results or output, so you can at least feel productive when trying to solve a hard problem. One of the most powerful simulation techniques is *Markov Chain Monte Carlo (MCMC)*, which is a specialty of the Harvard Statistics Department.

2. **Make approximations and bound their error.**

   When problems are too hard to solve analytically, you can often approximate some distributions by Poisson, Normal, Negative Binomial, or other special cases. This is awesome, and in particular, the *bounding* part of this technique will be our focus today.

In many problems, approximation can be useful. For example, we can say something like a "linear" or "quadratic" approximation, but these asymptotic bounds don't tell us exactly how close we are to the answer. Convergence in the $n \to \infty$ case isn't immediately applicable to discuss what a distribution looks like when $n = 5$, or even $n = 30$.

We are going to develop a few inequalities, which we can apply to make statements like *p is within ϵ of the true value*. Let's get started with one of the most famous inequalities in math.[25]

Proposition 15.1 (Cauchy-Schwarz inequality)**.** *If X and Y are random variables, then*

$$\mathbf{E}\left[|XY|\right] \leq \sqrt{\mathbf{E}\left[X^2\right]\mathbf{E}\left[Y^2\right]}.$$

Proof. This proof is particularly nice, though slightly algebra-heavy. The key idea is that variances are a sum-of-squares, so for any value of $\beta \in \mathbb{R}$,

$$\mathbf{E}\left[(Y - \beta X)^2\right] \geq 0.$$

This is an infinite family of inequalities. We can take the derivative to find the value of β that gives us the strongest bound. This is a neat problem-solving idea because we added complexity with this

[24] This includes things like exponential families and natural exponential families, convergence theorems, the central limit theorem, and martingales. If we had more time, Joe mentions that he would have also liked to discuss Markov chains. Those are covered in Stat 212 and Stat 171.

[25] For more on this, Joe recommends J. Michael Steele's book *The Cauchy-Schwarz Master Class*.

additional variable, but it actually makes the solution easier. In any case, the optimal value of β is given by the projection of Y onto X in the Hilbert space, which is

$$\beta = \frac{\langle X, Y \rangle}{\langle X, X \rangle} = \frac{\mathbf{E}[XY]}{\mathbf{E}[X^2]}.$$

Substituting this into the inequality and expanding yields

$$\begin{aligned} \mathbf{E}\left[Y^2\right] + \beta^2 \mathbf{E}\left[X^2\right] &\geq 2\beta \mathbf{E}[XY], \\ \frac{\mathbf{E}\left[X^2\right]\mathbf{E}\left[Y^2\right]}{\mathbf{E}[XY]} + \mathbf{E}[XY] &\geq 2\mathbf{E}[XY], \\ \mathbf{E}\left[X^2\right]\mathbf{E}\left[Y^2\right] &\geq \mathbf{E}[XY]^2. \end{aligned}$$

The result follows after assuming, without loss of generality, that X and Y are nonnegative. □

In the probability setting, this means that we can bound the covariance of random variables, which is a 2-variable expectation, by the product of the second moments of their *marginal* distributions. Marginal distributions are often easier to calculate.

Corollary 15.1.1 (Covariance inequality)**.** *For any random variables X and Y, we have*

$$|\mathrm{Corr}(X, Y)| = \left| \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbf{Var}[X]\,\mathbf{Var}[Y]}} \right| \leq 1.$$

Proof. This corollary is almost equivalent to Cauchy-Schwarz, but it admits a particularly elegant direct proof. Assume without loss of generality that X and Y are standardized to have mean 0 and variance 1, and let $\rho = \mathrm{Corr}(X, Y)$. Since variances are nonnegative,

$$\begin{aligned} \mathbf{Var}[X + Y] &= \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\,\mathrm{Cov}(X, Y) = 1 + 1 + 2\rho \geq 0 \implies \rho \geq -1, \\ \mathbf{Var}[X - Y] &= \mathbf{Var}[X] + \mathbf{Var}[Y] - 2\,\mathrm{Cov}(X, Y) = 1 + 1 - 2\rho \geq 0 \implies \rho \leq 1. \end{aligned}$$

□

15.2 Concentration Inequalities

Now let's get into our first inequality for proving tail bounds!

Proposition 15.2 (Markov's inequality)**.** *If $Y \geq 0$, then*

$$P(Y \geq a) \leq \frac{\mathbf{E}[Y]}{a}.$$

Proof. This is an extremely crude bound. Observe that

$$a \cdot I_{Y \geq a} \leq |Y|.$$

This is obviously true. Here, $a \cdot I_{Y \geq a}$ is simply equal to a when $Y \geq a$ and 0 when $Y < a$. Now, we can take expectation on both sides, and the result immediately follows. □

Although Markov's inequality seems really obvious, it's the starting point for pretty much all concentration bounds, as it makes very few assumptions about the random variable Y. If we additionally assume that Y has a second moment, then we can extend our bound slightly.

Proposition 15.3 (Chebyshev's inequality)**.** *If Y is a random variable with finite variance, then*

$$P(|Y - \mathbf{E}[Y]| \geq c) \leq \frac{\mathbf{Var}[Y]}{c^2}.$$

Proof. Apply Markov's inequality to the random variable $X = (Y - \mathbf{E}[Y])^2$. □

Note that even though Chebyshev's inequality is a trivial extension of Markov, you often get much better tail bounds using it, as they are quadratic in the deviation. This idea of applying an increasing function (such as $x \mapsto x^2$) to both sides of Markov's inequality can be used to get even better tail bounds in general, such as the celebrated *Chernoff bound*.[26]

Proposition 15.4 (Chernoff bound)**.** *Let Y be a nonnegative random variable, and let $t > 0$ be a constant. Then,*

$$P(Y \geq a) \leq P(e^{tY} \geq e^{ta}) \leq \frac{\mathbf{E}\left[e^{tY}\right]}{e^{ta}},$$

where the last step follows by Markov's inequality.

Notice the coincidental appearance of the moment generating function $M_Y(t) = \mathbf{E}\left[e^{tY}\right]$ above. This means that for Chernoff bounds to be applied, you essentially need *all of the moments* to be defined. Intuitively, it is the limit case of many concentration inequalities based on moments, as it makes a strong assumption of the MGF existing. The Chernoff bound is also intuitively useful because it lets you optimize for *any* value of t by taking the derivative.

15.3 More Basic Inequalities

We'll continue to cover more concentration inequalities in future weeks, such as Azuma-Hoeffding's inequality for martingales and McDiarmid's inequality (which is my personal favorite). However, let's go back to talking about inequalities in general.

Proposition 15.5 (Jensen's inequality)**.** *If g is a convex function, then*

$$\mathbf{E}[g(X)] \geq g(\mathbf{E}[X]).$$

Proof. Jensen's inequality is interesting because it does not rely on smoothness properties of g, and it also works in any number of dimensions. A proof is given in the book using the supporting hyperplane theorem for convex sets. □

Definition 15.6 (p-norms of random variables)**.** The L^p norm for a random variable X, where we have some fixed $p \geq 1$, is defined by

$$\|X\|_p = \mathbf{E}\left[|X|^p\right]^{1/p}.$$

This is a valid norm for two reasons. First, if $\|X\|_p = 0$, then X is almost surely zero. Second, the norm satisfies the triangle inequality, which is a fact called Minkowski's inequality.

Now let's ask the question of how the r-norm compares to the s-norm, when $1 \leq r < s$. The following result actually holds for any values of r and s, including negative values and zero (in the limit, which is called the geometric mean).

[26] Named after Herman Chernoff, who is faculty emeritus at Harvard.

Proposition 15.7 (Monotonicity of norms)**.** *If $1 \le r < s$, then*

$$\|X\|_r \le \|X\|_s.$$

(This is a continuous version of the so-called inequality of power means.)

Proof. This follows from Jensen's inequality on the convex function $x \mapsto x^{s/r}$. Assume without loss of generality that X is nonnegative. Then X^r is also nonnegative, so

$$\mathbf{E}\left[(X^r)^{s/r}\right] \ge \mathbf{E}\left[X^r\right]^{s/r} \implies \mathbf{E}\left[X^s\right]^{1/s} \ge \mathbf{E}\left[X^r\right]^{1/r}.$$

□

Finally, we write down one of the most famous and classical inequalities, which is a special case of the discrete inequality of power means mentioned above!

Proposition 15.8 (AM-GM inequality)**.** *If $x_1, \dots, x_n \ge 0$ are numbers and $w_1, \dots, w_n \ge 0$ are weights with $w_1 + \cdots + w_n = 1$, then*

$$\sum_{i=1}^{n} w_i x_i \ge \prod_{i=1}^{n} x_i^{w_i}.$$

The left-hand side is called the arithmetic mean, and the right-hand side is called the geometric mean. Equality holds if and only if $x_1 = x_2 = \cdots = x_n$.

Proof. Assume without loss of generality that the x_i are distinct. Let W be a random variable supported on $\{x_1, \dots, x_n\}$, with $P(W = x_i) = w_i$ for each i. By Jensen's inequality on log,

$$\sum_{i=1}^{n} w_i \log x_i = \mathbf{E}\left[\log W\right] \le \log \mathbf{E}\left[W\right] = \log\left(\sum_{i=1}^{n} w_i x_i\right).$$

The result follows after exponentiating both sides. □

Corollary 15.8.1 (Young's inequality)**.** *In the special $n = 2$ case of weighted AM-GM, we have*

$$a^p b^q \le pa + qb,$$

where $a, b, p, q \ge 0$ and $p + q = 1$.

16 October 29th, 2020

Today we continue our discussion of inequalities and norms.

16.1 Hölder's Inequality and Nonnegative Covariance

Recall that we covered p-norms in the last lecture. One nice property of these norms is that if random variables have zero distance under a norm, such as $\mathbf{E}\left[(X-Y)^k\right] = 0$, then X equals Y almost surely. However, to be able to actually show this, we need the k-th moment to exist. A common trend we'll see in future classes is that fact get generalized by assuming less about the existence of higher moments, such as various central limit theorem conditions.

Anyway, with that aside out of the way, we will now cover a classic result in analysis. This is a generalization of Cauchy-Schwarz.

Proposition 16.1 (Hölder's inequality)**.** *Let $r, s \geq 1$ with $\frac{1}{r} + \frac{1}{s} = 1$. Then, for any random variables X, Y where the corresponding norms are defined,*

$$\|XY\|_1 \leq \|X\|_r \|Y\|_s.$$

Here, r and s are sometimes called conjugate norms. The Cauchy-Schwarz inequality is the special case where $r = s = 2$ (see Proposition 15.1).

Proof. Joe notes that you'll often see complex proofs of this, but the proof actually only takes one or two lines once you know how to set it up. We'll start by assuming without loss of generality that $\|X\|_r = \|Y\|_s = 1$, since both sides of the inequality are bilinear. For any $x, y \in \mathbb{R}$, we have by Young's inequality (Corollary 15.8.1) on $|x|^r$ and $|y|^s$ that

$$|xy| \leq \frac{|x|^r}{r} + \frac{|y|^s}{s}.$$

Since the above result holds for real numbers, it also holds for random variables. Therefore,

$$\mathbf{E}\left[|XY|\right] \leq \frac{\|X\|_r^r}{r} + \frac{\|X\|_s^s}{s} = \frac{1}{r} + \frac{1}{s} = 1.$$

□

Hopefully that was an inspiring, short proof of a classic inequality in analysis. Hoping to outdo himself, Joe will now attempt to present an even more inspiring proof of another inequality.

Proposition 16.2 (Nonnegative covariance[27])**.** *If g and h are non-decreasing functions, then*

$$\operatorname{Cov}(g(X), h(X)) \geq 0.$$

Proof. The key idea is to choose i.i.d. $X_1, X_2 \sim X$. Then, observe that

$$(g(X_1) - g(X_2))(h(X_1) - h(X_2)) \geq 0.$$

What happens when we take the covariance of the above expression? Well,

$$\begin{aligned}
&\mathbf{E}\left[(g(X_1) - g(X_2))(h(X_1) - h(X_2))\right] \\
&= \mathbf{E}\left[g(X_1)h(X_1)\right] - \mathbf{E}\left[g(X_1)h(X_2)\right] - \mathbf{E}\left[g(X_2)h(X_1)\right] + \mathbf{E}\left[g(X_2)h(X_2)\right] \\
&= 2\,\mathbf{E}\left[g(X)h(X)\right] - 2\,\mathbf{E}\left[g(X)\right]\mathbf{E}\left[h(X)\right] \\
&= 2\operatorname{Cov}(g(X), h(X)).
\end{aligned}$$

The result follows from the nonnegativity of that expression. □

[27] This is a special case of the FKG inequality in correlation theory. Amusingly enough, it's also a continuous version of Chebyshev's sum inequality from olympiad mathematics — even having essentially the same proof!

16.2 Convergence and the Borel-Cantelli Lemma

Recall in an earlier lecture that we introduced the notions of almost-sure convergence (Definition 11.2) and convergence in probability (Definition 11.1). The first is stronger than the second. Let's rigorously introduce one more useful notion of convergence, the weakest so far.

Definition 16.3 (Convergence in distribution)**.** Consider an infinite sequence of random variables $X_1, X_2, \ldots$, and a random variable X. Let $F_1, F_2, \ldots$ be the CDFs of $X_1, X_2, \ldots$, and let F be the CDF of X. Then we say that $X_1, X_2, \ldots \to X$ *in distribution* if $\lim_{n\to\infty} F_n(x) = F(x)$ for all continuity points x of F.

In some sense, convergence in distribution is **much** weaker than the other two, as it only talks about the marginal distributions of the random variables. Meanwhile, almost sure convergence is only slightly weaker than convergence in probability.

Example 16.4 (Convergence in distribution but not in probability)**.** Consider an infinite sequence of i.i.d. $U, U_1, U_2, \ldots \sim \text{Unif}$. Then, clearly $U_1, U_2, \ldots \sim U$ in distribution, as all of their marginal distributions are the same (uniform). However, they do not converge in probability, as clearly $\Pr(|U_n - U| > \epsilon) \geq 1 - 2\epsilon$.

Example 16.5 (Convergence in probability but not almost surely)**.** Let $X_n \sim \text{Bern}(1/n)$, and assume that all of them are independent. Then $X_1, X_2, \ldots \to 0$ in probability, since

$$\Pr(|X_n - 0| > \epsilon) \leq \Pr(X_n = 1) = 1/n.$$

However, it does not converge almost surely.

How can we show that the above example does **not** converge almost surely? It's true that in the sequence $X_1, X_2, \ldots$, the 1 values get rarer and rarer as $n \to \infty$. If there is a finite number of 1's, then we have almost sure convergence, but if there are an infinite number of 1's, then we do not have convergence.

With this motivation in mind, we will now deliver a two-part lemma that very elegantly describes the above as a dichotomy — useful both for proving and disproving almost sure convergence.

Proposition 16.6 (Borel-Cantelli lemma)**.** *Let $E_1, E_2, \ldots$ be an infinite sequence of pairwise independent events. Let $p = P(\limsup_{n\to\infty} E_n)$ be the probability that infinitely many of the events occur. Then, $p \in \{0, 1\}$, and furthermore:*

1. *(Borel-Cantelli lemma).*[28] *If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $p = 0$.*
2. *(Second Borel-Cantelli lemma). If $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $p = 1$.*

The second lemma immediately shows why Example 16.5 does not have almost sure convergence, as the harmonic series diverges. However, if we had changed it slightly to $X_n \sim \text{Bern}(1/n^{1.001})$ instead, it would converge almost surely by the first Borel-Cantelli lemma.

[28] This first version of the lemma also holds when the E_i are not necessarily independent.

17 November 3rd, 2020

Last time, we were talking about convergence. Let's pick up where we left off.

17.1 More on Convergence

Commonly, we want to get examples and counterexamples of theorems in statistics, to help us intuit about formalisms. One common way of doing this is to take simple random variables, such as a coin flip Bern(p). We can scale these random variables in arbitrary ways and see if any interesting examples come out. In the off chance that we want correlated random variables, one way is to first select p from another distribution (such as uniform) before starting to sample Bernoullis with mean p, which are conditionally independent on p.

Exercise 17.1. Construct an infinite, random sequence of coin tosses such that for any n consecutive coin tosses, the probability of all n tosses coming up heads is $\frac{1}{n+1}$.

Now let's prove the Borel-Cantelli lemma. Joe mentions that this is interesting not just for completeness, but also because it illustrates many useful ideas in analysis — short yet instructive.

Proof of Proposition 16.6. We'll prove each of the two parts separately.

1. Assume that $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then, by the definition of lim sup and a union bound,

$$P\left(\limsup_{n\to\infty} A_n\right) = P\left(\bigcap_{n\geq 1}\bigcup_{m\geq n} A_m\right) \leq P\left(\bigcup_{m\geq n} A_m\right) \leq \sum_{m=n}^{\infty} P(A_m).$$

   This is the tail of the series, but by the definition of convergence of an infinite series, its partial sums must converge. Therefore, the tail of the series $P(A_1) + P(A_2) + \cdots$ must converge to zero, so we conclude.

2. Our strategy in this case will be slightly different. Instead of trying to directly prove that something will happen infinitely often, we're going to show that the complement (event happens finitely often) has zero probability. In other words, we want

$$P\left(\bigcup_{n\geq 1}\bigcap_{m\geq n} A_m^C\right).$$

   A useful fact from measure theory is that the countable union of measure-zero sets also has measure zero. Therefore, it's equivalent to show that the inner intersection has measure zero for any n. Since the A_m are independent, we have

$$P\left(\bigcap_{m\geq n} A_m^C\right) = \prod_{m=n}^{\infty} P(A_m^C) = \prod_{m=n}^{\infty} (1 - P(A_m)) \leq e^{-\sum_{m=n}^{\infty} P(A_m)}.$$

   Since the infinite series in the exponent diverges to ∞, we conclude.

□

The next topic is an example of a *zero-one law* similar to the Borel-Cantelli lemma.

Proposition 17.1 (Kolmogorov zero-one law)**.** *Let $A_1, A_2, \ldots$ be independent events. Recall that we can generate a σ-algebra from a collection of sets (i.e., events) by taking the smallest σ-algebra containing those events. Then, the "tail field" of the A_i is*

$$\mathcal{A} = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, A_{n+2}, \ldots).$$

You can think of the tail field as the set of events that only depends on the limiting tail of the event sequence. Then, for any $A \in \mathcal{A}$, we have $P(A) \in \{0, 1\}$.

Proof. Omitted, but the key idea in this proof is very "cute" — it is to show that $A \perp\!\!\!\perp A$. □

This generalizes part of the Boreli-Cantelli lemma, since $\limsup_{n\to\infty} A_n$ is an example of something that only depends on the limiting values of A_n, so it is in the tail field.

17.2 Building a Hierarchy of Convergence

In this section, we will show that almost sure convergence is stronger than convergence in probability. We've already seen a subtle example in Example 16.5 of how they differ.

Proposition 17.2 (Almost sure convergence implies convergence in probability)**.** *If $X_n \to X$ almost surely, then $X_n \to X$ in probability.*

Proof. First, observe that

$$P(|X_n - X| > \epsilon) \le P\Bigg(\underbrace{\bigcup_{m=n}^{\infty} \{|X_m - X| > \epsilon\}}_{A_n}\Bigg).$$

As notated above, we call the event in parentheses A_n. We have $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. This inequality captures the essence of almost sure convergence (within ϵ for all indices after n), versus convergence in distribution (within ϵ for n). To see why, consider

$$P\left(\bigcap_{n=1}^{\infty} A_n\right) = P(|X_n - X| > \epsilon \text{ infinitely many times}).$$

This is zero precisely when $X_n \to X$ converges almost surely. By our first inequality above, we have proven the proposition. □

Next on our menu is a theorem that Joe calls both "beautiful and useful," which lets you go from convergence in distribution back to convergence in probability. However, there has to be a catch, since convergence in distribution is obviously weaker. We will need to move to a *different probability space*.

Proposition 17.3 (Skorokhod's representation theorem)**.** *Suppose that $X_n \to X$ in distribution. Then, there exists a new probability space $(\Omega^*, \mathcal{F}^*, P^*)$, with random variables $X_n^*, X^* : \Omega^* \to \mathbb{R}$, such that $X_n^* \sim X_n$, $X^* \sim X$, and $X_n^* \to X^*$ almost surely.*

Proof. The proof is omitted because of hard technical details. However, in principle, the key intuition is that you can just take a PIT on all of the X_n variables, which couples them to the same uniform. This fixes the issue where the X_n may be totally independent, in a sequence that converges in distribution. □

Skorokhod's theorem is somewhat of a useful hammer. One neat application is that you can really easily prove the "in distribution" case of the continuous mapping theorem, by reducing it to the "almost sure" case using Skorokhod.

18 November 5th, 2020

Today we will start talking about asymptotics. In other words, how do distributions change in some limit where their parameters go to infinity? Some of these theorems are quite beautiful,[29] but we will specifically focus on facts that have practical applications.

18.1 Major Tools in Asymptotics

Let's review a few facts from analysis, which we'll discuss and use as a stepping stone.

- **Continuous mapping theorem:** If $X_n \to X$ in any kind of convergence, then $g(X_n) \to g(X)$ in the same kind of convergence, assuming that g is a continuous function.
- **Taylor's theorem:** Taylor approximations are also called the Delta method in statistics.
- **Slutsky's theorem:** Convergence of a binary operation on two random variables.
- **LLN, CLT:** What we're going to talk about soon.

For the rest of the semester, we will focus on a few high-level goals. One topic is *natural exponential families*, which unify a lot of distributions that we've seen this semester.[30] This includes the special *NEF-QVF* families. We'll also talk about *martingales*, which are useful for concentration bounds and for modeling financial markets.

Proposition 18.1. *If $X_1, X_2, \ldots \to X$ in distribution and $Y_1, Y_2, \ldots \to Y$ in distribution, and these two sequences are mutually independent, then $X_n + Y_n \to X + Y$ in distribution.*

Proposition 18.2 (Slutsky's theorem)**.** *Assume that we have two sequences of random variables $X_1, X_2 \ldots$ and $Y_1, Y_2, \ldots$, not necessarily independent, such that $X_n \to X$ and $Y_n \to c$ in distribution, where c is a constant. Then,*

- *$X_n + Y_n$ converges in distribution to $X + c$,*
- *$X_n - Y_n$ converges in distribution to $X - c$,*
- *$X_n Y_n$ converges in distribution to cX, and*
- *X_n / Y_n converges in distribution to X/c, as long as we avoid division by zero.*

Proof. This is a somewhat technical fact from analysis, so we omit the proof. □

Proposition 18.3 (Delta method)**.** *Assume that you have a sequence of random variables $T_1, T_2, \ldots$ such that $\sqrt{n}(T_n - \theta_0) \to Z$ in distribution, where θ_0 is some constant. If g is a real function that is $\mathcal{C}^1$ continuous at θ_0, then*

$$\sqrt{n}(g(T_n) - g(\theta_0)) \to g'(\theta_0) Z$$

in probability. In particular, the special case when $Z \sim \mathcal{N}(0, 1)$ is particularly nice because of connections with the central limit theorem.

[29] Joe cites the law of the iterated logarithm as an example.

[30] Around this time, Carl Morris walked into our class and said hello. He is the "originator" of the NEF.

Proof. The proof uses the mean value theorem, which tells us that

$$g(T_n) = g(\theta_0) + g'(\tilde{\theta}_n)(T_n - \theta_0),$$

for some $\tilde{\theta}_n$ between θ_0 and T_n. Then,

$$\sqrt{n}(g(T_n) - g(\theta_0)) = \sqrt{n}g'(\tilde{\theta}_n)(T_n - \theta_0).$$

However, note that $g'(\tilde{\theta}_n) \to g'(\theta_0)$ by continuity. Also, $T_n - \theta_0 \to Z/\sqrt{n}$ in distribution, so after applying Slutsky's theorem to the above equation, we conclude. □

18.2 Natural Exponential Families

Now we'll introduce our next topic, which is NEFs. Natural exponential families are a special case of the more general *exponential family*.

Definition 18.4 (Natural exponential family)**.** A *natural exponential family* with natural parameter η is a family of distributions with CDF F_η, taking the form

$$\mathrm{d}F_\eta(y) = e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y).$$

We give the condition that $F_0(y)$ does not depend on η. In particular, it's just the $\eta = 0$ case.

Here, we can see that $\psi(\eta)$ is a normalizing factor for the rest of the density. The rough idea is that we just shift probabilities by weighting with pointwise multiplication by some exponential of the value. In particular,

$$\int \mathrm{d}F_\eta(y) = \int e^{\eta y - \psi(\eta)}\,\mathrm{d}F_0(y) = 1 \implies e^{\psi(\eta)} = \int e^{\eta y}\,\mathrm{d}F_0(y) = \mathbf{E}_{Y\sim F_0}[e^{\eta Y}].$$

The last step above follows from LOTUS. In particular, this means that $\psi(t)$ is just the cumulant generating function of $Y \sim F_0$. It's also easy to show that the cumulant generating function of F_η for any η is $\psi(t+\eta) - \psi(\eta)$. For this reason, we call ψ the *cumulant function*.

Example 18.5. The binomial distribution $\mathrm{Bin}(n, p)$ is a natural exponential family for any fixed value of n, where we vary p. In this case, the natural parameter is given by the logit function $\mathrm{logit}(p) = \log \frac{p}{1-p}$.

Example 18.6. The normal distribution with unit variance, $\mathcal{N}(\mu, 1)$, is a natural exponential family with natural parameter μ and cumulant function $\mu^2/2$. Notice how this aligns with the cumulant generating function of the standard normal $\mathcal{N}(0, 1)$, which is $t^2/2$.

Another useful fact, which falls out of the cumulant function, is that if we let $\psi'(\eta) = \mu$ and $\psi''(\eta) = \sigma^2$, then $F_\eta \sim [\mu, \sigma^2]$. Note that since variances are positive (except in the degenerate case), this tells us that $\psi'(\eta) = \mu$ is a strictly increasing function, so we can invert it.

Definition 18.7 (Variance function)**.** The *variance function* of a natural exponential family F_η is $V(\mu) = \sigma^2$. In other words, we have for any η that $V(\psi'(\eta)) = \psi''(\eta)$.

Definition 18.8 (NEF-QVF)**.** An *NEF-QVF* is a natural exponential family with variance function of the form $V(\mu) = v_0 + v_1\mu + v_2\mu^2$.

It is a theorem from Carl Morris that there are only **six** NEF-QVF distribution families.

19 November 10th, 2020

Today we will continue talking about NEF-QVFs and asymptotics (delta method), in preparation for the law of large numbers and central limit theorem.

19.1 Example of the Delta Method in Asymptotics

Recall that we proved the general delta method last time, in Proposition 18.3, by using Taylor series approximations. However, although we gave a proof, we didn't actually show how it is often used in practice. Consider the case where we have the sample mean $\overline{X}_n$ of n i.i.d. random variables from $[\mu, \sigma^2]$. The basic statement of the central limit theorem is that as $n \to \infty$,

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} \mathcal{N}(0, 1),$$

assuming some mild regularity conditions on the distribution. This is one of the most celebrated theorems in all of probability and statistics. Note that although CLT is a nice statement strictly about asymptotics, you can actually get a concrete version of the bound using the Berry-Esseen theorem. The Berry-Esseen version is still fairly weak though; it tries to be too general.[31]

Example 19.1 (Central limit theorem for Poissons)**.** Suppose that $X \sim \text{Pois}(\lambda)$, where λ is large. Then, observe that by the properties of the Poisson distribution, X can be represented as the sum of λ i.i.d. copies of a Pois(1), so $X \approx \mathcal{N}(\lambda, \lambda)$ approximately.

(Note that the Poisson distribution has equal mean and variance both tied to the parameter λ, which can lead to some issues in modeling, as many processes exhibit *overdispersion*, where the variance is larger than the mean. The negative binomial is sometimes used as a proxy.)

Suppose that we would like to perform a *variance-stabiling transformation* on $X \sim \text{Pois}(\lambda)$ to approximately disentangle the mean from the variance. Transformations are a big topic in statistics for wrangling data. For example, we might apply a cube-root transformation to minimize skewness and make the CLT approximation better. In this case, we will show that taking the square root is a transformation that helps stabilize variance.

Example 19.2 (Delta method for square root of Poisson)**.** Suppose that λ is large, so $X \sim \text{Pois}(\lambda)$ is approximately $\mathcal{N}(\lambda, \lambda)$. Note that Jensen's inequality tells us that $\mathbf{E}[\sqrt{X}] < \sqrt{\mathbf{E}[X]} = \sqrt{\lambda}$. However, it actually turns out that by applying Proposition 18.3 to the function $g(x) = \sqrt{x}$, we get $\sqrt{X} \xrightarrow{D} \mathcal{N}(\sqrt{\lambda}, \frac{1}{4})$ as $\lambda \to \infty$.

19.2 The Law of Large Numbers

There are several versions of the *Law of Large Numbers (LLN)*, each requiring slightly different assumptions. These can be roughly divided into two types:

- **Strong** laws of large numbers deal with convergence of the sample mean almost surely.
- **Weak** laws of large numbers deal with convergence of the sample mean in probability.

This terminology is unique to the law of large numbers, as the central limit theorem only applies to convergence in distribution. Generally, we will see that weak laws of large numbers have a slightly weaker result, but also require less assumptions.

[31] For practical purposes, Joe suggests using simulation to find the smallest value of n for which the distribution of the mean becomes approximately normal, i.e., $\overline{X}_n \approx \mathcal{N}(\mu, \frac{\sigma^2}{n})$. This is not useful for rigorous proofs though.

Proposition 19.3 (Weak LLN, basic version)**.** *Suppose that $\overline{X}_n$ is the mean of n i.i.d. random variables with mean μ and finite variance σ^2. Then, as $n \to \infty$, $\overline{X}_n \to \mu$ in probability.*

Proof. By Chebyshev's inequality,

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\mathbf{Var}\left[\overline{X}_n\right]}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

This goes to zero as $n \to \infty$, so we're done. □

This was a really simple proof. Let's see how we can relax the assumptions a bit, by being a little more sophisticated in our argument. In particular, what if the random samples were not independent? One strategy to deal with this is to consider

$$\mathbf{Var}\left[\overline{X}_n\right] = \frac{\sum_{i,j} \mathrm{Cov}(X_i, X_j)}{n^2\epsilon^2}.$$

If the above value goes to zero as $n \to \infty$, then we have an equivalent to the weak law of large numbers. However, sometimes this strategy does not work, and we need to try something else. For example, consider the characteristic function, which always exists and can be approximated by derivatives. It turns out that with a linear approximation (first term) of the characteristic function, we get LLN, and with a quadratic approximation, we get CLT.

In the case when random variables are guaranteed to be independent, we can prove incredibly strong LLNs and CLTs due to the multiplicativity of the characteristic function. However, the dependent case is harder, and we might give a couple of examples, later on, where we relax the independence assumptions.

Proposition 19.4 (Strong LLN)**.** *Assume that $X_1, X_2, \ldots$ are i.i.d., with $\mathbf{E}[X_j] = \mu$ and the average absolute deviation is bounded, i.e., $\mathbf{E}[|X_j|] < \infty$. Then, $\overline{X}_n \to \mu$ almost surely.*

The above version of the strong LLN is hard to prove and fairly technical. This is because it only assumes first moments. For now, we will prove an easier version with a different set of assumptions — more moments, but also not necessarily i.i.d. this time.

Proposition 19.5 (Strong LLN, fourth moments)**.** *Assume that X_j are independent and have mean zero, and also that $\mathbf{E}[X_j^4] \leq b < \infty$ for some bound b. Then $\overline{X}_n \to 0$ almost surely.*

Proof. By the Borel-Cantelli lemma (Proposition 16.6), it suffices to check that for any $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P(|\overline{X}_n| > \epsilon) < \infty.$$

However, we have $P(|\overline{X}_n| > \epsilon) = P(\overline{X}_n^4 > \epsilon^4)$, applying Markov's inequality tells us that

$$\sum_{n=1}^{\infty} P(|\overline{X}_n| > \epsilon) \leq \frac{1}{\epsilon^4} \sum_{n=1}^{\infty} \mathbf{E}\left[\overline{X}_n^4\right].$$

This is fair enough, but how do we get rid of the fourth moment of the mean? One way to deal with this is by brute forcing through the multinomial theorem on $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. However, a nicer approach is to use cumulants, which are additive. Note that

$$\mathbf{E}\left[\overline{X}_n^4\right] = \kappa_4(\overline{X}_n) + 3\,\mathbf{Var}\left[\overline{X}_n^2\right]^2 = \frac{1}{n^4}(\kappa_4(X_1) + \cdots + \kappa_4(X_n) + 3(\mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n])^2).$$

Now we just need to bound $\kappa_4(X_j)$ and $\mathbf{Var}\left[X_j\right]$, for each j. This turns out to be very simple. First, the fourth cumulant is strictly less than the first moment (smaller by three times the variance), so $\kappa_4(X_j) \leq \mathbf{E}[X_j^4] \leq b$. Also, by Jensen's inequality, $\mathbf{Var}[X_j] \leq b$ as well. Therefore, the last summation above is bounded by

$$\frac{1}{\epsilon^4} \sum_{n=1}^{\infty} \mathbf{E}\left[\overline{X}_n^4\right] \leq \frac{1}{\epsilon^4} \sum_{n=1}^{\infty} \left(\frac{b}{n^3} + \frac{3b}{n^2} \right).$$

This summation converges because sums of $1/n^s$ are finite for $s > 1$, so we are done. □

Note. Even when b is not bounded by a constant, we can still use the argument above as long as the summation converges, i.e., when $b = o(n)$. For example, $b = n^{0.999}$ would work just as well.

It's instructive to ask why we need finite *fourth* moments in the law of large numbers above, versus two or six or some other number. This is because only having finite variances gives you linear falloff similar to the above, and the harmonic series diverges, so we end up on the wrong side of Borel-Cantelli for almost sure convergence.

20 November 12th, 2020

Today we will discuss the central limit theorem.

20.1 The Central Limit Theorem

Let's first provide a bit of intuition to motivate the central limit theorem. There's a couple of different ways to think about this:

1. **Cumulants:** Suppose that $X_1, X_2, \ldots$ are i.i.d. with mean 0 and variance 1. Then, the r-th cumulant of the sum of these variables, divided by $\sqrt{n}$, is

$$\kappa_r\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) = \frac{n}{n^{r/2}} K_r(X_1).$$

   This just follows from the additivity of cumulants. Notice that because of the exponent, this fraction approaches 0 as $n \to \infty$ for any $r > 1$. Therefore, we should expect the limiting distribution to only have finite first two cumulants — which makes it a normal distribution!

2. **Entropy:** The normal distribution maximizes entropy for a given mean and variance. When you add independent random variables together, their entropy increases, which is a statistical analogue of the second law of thermodynamics. Andrew Barron has a paper in *The Annals of Probability* where he proves CLT using an entropy-type argument.

3. **Stability:** Let $S_n = X_1 + \cdots + X_n$, and suppose that $S_n/\sqrt{n}$ converges in in distribution to some distribution Z. Why must Z be normal? Well, note that in convergence, we can replace n by $2n$, so

$$\frac{S_{2n}}{\sqrt{2n}} = \frac{X_1 + \cdots + X_n}{\sqrt{2n}} + \frac{X_{n+1} + \cdots + X_{2n}}{\sqrt{2n}} \xrightarrow{D} Z.$$

   However, we can also write the second expression above as converging in distribution to $Z_1/\sqrt{2} + Z_2/\sqrt{2}$, where Z_1, Z_2 are i.i.d. $\sim Z$. Since a sequence can't converge to two different distributions, these must be the same, so

$$Z \sim \frac{1}{\sqrt{2}}(Z_1 + Z_2).$$

   This is a stable law, and the only stable law with finite variance is the normal distribution.

Actually, although we only promised to give some intuition above, we can also formalize the third point to produce a rigorous proof as well. First, a quick lemma.

Lemma 20.1 (Taylor approximation for characteristic function)**.** *If X is a random variable with finite m-th moment $\mathbf{E}[|X|^m] < \infty$, and X has characteristic function φ, then*

$$\varphi(t) = \sum_{k=0}^{m} \frac{(it)^k \mathbf{E}\left[X^k\right]}{k!} + o(|t|^m),$$

for small t approaching 0.

Proof. This follows almost immediately from the Peano form of the Taylor series remainder for e^x. The only slight hiccup is that we need to apply dominated convergence, due to the expected value. This is also why we need to assume finite m-th moment. □

Proposition 20.2 (Stable law with finite variance)**.** *Let Z_1, Z_2 be i.i.d. random variables with mean 0 and variance 1. If $Z_1 + Z_2 \sim \sqrt{2}Z_1$, then $Z_1 \sim \mathcal{N}(0,1)$.*

Proof. We'll turn the condition into a functional equation of the characteristic function. Let φ be the characteristic function of Z_1. Then, the characteristic function of $\frac{Z_1+Z_2}{\sqrt{2}} \sim Z_1$ is

$$\mathbf{E}\left[e^{\frac{it}{\sqrt{2}}(Z_1+Z_2)}\right] = \varphi\left(\frac{t}{\sqrt{2}}\right)^2 = \varphi(t).$$

By iterating this functional equation, we get

$$\varphi(t) = \varphi\left(\frac{t}{\sqrt{2}}\right)^2 = \varphi\left(\frac{t}{\sqrt{2^2}}\right)^{2^2} = \cdots = \varphi\left(\frac{t}{2^{n/2}}\right)^{2^n}.$$

Since $\kappa_1(Z_1) = 0$ and $\kappa_2(Z_1) = 1$, we have by Lemma 20.1 that

$$\lim_{n\to\infty} \varphi\left(\frac{t}{2^{n/2}}\right)^{2^n} = \lim_{n\to\infty}\left(1 - \frac{t^2}{2!\cdot 2^n} + o\left(\frac{1}{2^n}\right)\right)^{2^n} = e^{-t^2/2}.$$

Therefore, by uniqueness of characteristic functions (Fourier transform), we conclude. □

Therefore, from the intuition at the start of this section, this stable law immediately implies the basic result of the central limit theorem itself.

Proposition 20.3 (Classical CLT (Lindeberg–Lévy))**.** *If $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$, then as $n \to \infty$, we have in distribution that*

$$\frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

20.2 More Central Limit Theorems

What are the limitations of Proposition 20.3? Even though it's an extraordinarily powerful result, we still want to handle cases when our random variables are independent but not identically distributed. Sometimes we even want CLTs for *weakly dependent* random variables.

We will now introduce four different central limit theorems. The common setting is that we have a sequence $X_1, X_2, \ldots$ of independent random variables, where X_j has mean 0 and variance σ_j^2. Let $S_n = X_1 + \cdots + X_n$, let $Y_j = X_j/\sigma_j \sim [0,1]$ be the standardized version of X_j. Let

$$Z_n = S_n/s_n = \sum_{j=1}^{n} \sigma_j Y_j / s_n \sim [0,1]$$

be the standardized version of S_n, where

$$s_n^2 = \mathbf{Var}[S_n] = \sum_{j=1}^{n} \sigma_j^2.$$

Definition 20.4 (UAN)**.** The *uniform asymptotic negligibility* condition is that none of that none of the n terms in S_n has a large asymptotic variance, in comparison to the total variance s_n^2. In general, UAN holds when

$$u_n := \frac{\max_{1\le j\le n} \sigma_j}{s_n} \to 0.$$

We can interpret u_n^2 as the largest fraction of the variance contributed by any single term X_j of the entire sum S_n.

It turns out that the UAN is *almost* a sufficient condition to prove a central limit theorem.[32] As a consequence, we will primarily focus on the setting in which UAN holds, which allows us to prove two central limit theorems that turn out to be equivalent. The first is due to Morris and Blitzstein, while the second is very famous.

Proposition 20.5 (Fundamental bound)**.** *Define the "fundamental bound"* FB_n *by*

$$\text{FB}_n = \sum_{j=1}^{n} \mathbf{E}\left[\left(\frac{X_j}{s_n}\right)^2 \min\left(1, \frac{|X_j|}{s_n}\right)\right].$$

If $\text{FB}_n \to 0$*, then* Z_n *converges in distribution to* $\mathcal{N}(0,1)$*.*

Proposition 20.6 (Lindeberg CLT)**.** *Define Lindeberg's condition, for any* $\epsilon > 0$*, to be*

$$\text{Lind}_{\epsilon,n} = \sum_{j=1}^{n} \mathbf{E}\left[\left(\frac{X_j}{s_n}\right)^2 I_{|X_j|/s_n > \epsilon}\right] \to 0.$$

If Lindeberg's condition holds for each ϵ*, then* $Z_n \to \mathcal{N}(0,1)$*.*

In some sense, the fundamental bound and Lindeberg CLT are both equivalently strong, as well as the "strongest" central limit theorem. In the presence of the UAN, both of these central limit theorems are necessary and sufficient for convergence to take place. They also both immediately imply the UAN condition, i.e., they are supersets of the UAN.

However, both of these conditions (fundamental bound and Lindeberg) are fairly technical, so we oftentimes prefer to apply simpler variants of the central limit theorem. These are not as general, but they can still be useful in many practical applications.

Proposition 20.7 (Lyapunov CLT)**.** *Define Lyapunov's condition, for any* $r > 2$*, to be*

$$\text{Lyap}_{r,n} = \sum_{j=1}^{n} \mathbf{E}\left[\left|\frac{X_j}{s_n}\right|^r\right] \to 0.$$

If Lyapunov's condition holds for some r*, then* $Z_n \to \mathcal{N}(0,1)$*.*

Proof. It's easy to show that this implies Lindeberg's condition. For any $\epsilon > 0$, note that

$$\left(\frac{X_j}{s_n}\right)^2 I_{|X_j|/s_n > \epsilon} \le \frac{1}{\epsilon^{r-2}}\left|\frac{X_j}{s_n}\right|^r I_{|X_j|/s_n > \epsilon} \le \frac{1}{\epsilon^{r-2}}\left|\frac{X_j}{s_n}\right|^r.$$

This works for any $r > 2$, and $\epsilon > 0$ is just a constant here, so we can ignore it. Therefore, if Lyapunov's condition holds, so does Lindeberg's condition, and we are done. □

Proposition 20.8 (Fourth cumulant CLT)**.** *If the UAN condition holds and* $|\kappa_4(Z_n)| \to 0$*, then* $Z_n \to \mathcal{N}(0,1)$*.*

Proof. This turns out to be equivalent to the $r = 4$ case of Lyapunov's CLT. Observe that

$$\text{Lyap}_{4,n} = \sum_{j=1}^{n} \kappa_4(X_j/s_n) + 3\,\mathbf{Var}\left[X_j/s_n\right]^2 = \kappa_4(Z_n) + 3\sum_{j=1}^{n}(\sigma_j/s_n)^4.$$

If we assume the UAN condition, then the latter term definitely tends to zero, so we are done. □

[32] In fact, the UAN condition is also *almost* necessary, in the sense that if any term contributes an asymptotically nontrivial portion to the variance, then CLT only holds when that term is already normally distributed itself.

The proofs of all of these CLTs are given in the textbook. They are all pretty technical arguments involving analysis on the characteristic function. Anyway, we can now do an illustrative example.

Example 20.9. Assume that $Y_1, \dots, Y_n$ are i.i.d. $\sim [0,1]$, and let $S_n = c_1 Y_1 + \cdots + c_n Y_n$. Then, the UAN can be written as

$$\frac{\max_{1 \le j \le n} c_j^2}{\sum_{j=1}^n c_j^2} < \infty,$$

and it turns out that this condition alone is enough to prove that $\frac{S_n}{\sqrt{c_1^2 + \cdots + c_n^2}}$ converges in distribution to $\mathcal{N}(0,1)$. Let's see how to do this with the κ_4 method. Observe that

$$|\kappa_4(Z_n)| = \frac{\left(\sum_{j=1}^n c_j^4\right) |\kappa_4(Y_1)|}{\left(\sum_{j=1}^n c_j^2\right)^2}.$$

To bound this, we use a really neat trick. Note that $c_j^4 = c_j^2 c_j^2$. So, we can write

$$\frac{\sum_{j=1}^n c_j^4}{\left(\sum_{j=1}^n c_j^2\right)^2} \le \frac{\left(\max_{1 \le j \le n} c_j^2\right) \sum_{j=1}^n c_j^2}{\left(\sum_{j=1}^n c_j^2\right)^2} = \frac{\max_{1 \le j \le n} c_j^2}{\sum_{j=1}^n c_j^2} \to 0,$$

which is simply the UAN condition. Therefore, assuming that the fourth cumulants of Y_i exist, we are done by Proposition 20.8.

21 November 17th, 2020

First, some administrative information. The midterm grades will be posted tonight, and information about the final project will be added shortly. We're getting close to the end of the course, and today we'll continue discussing the central limit theorem. The last topic after this will be martingales. Future courses to consider include Stat 171 and Stat 212.

21.1 Examples of the Central Limit Theorem

We discussed CLT last lecture, but we didn't go into many examples of each of the cases. These central limit theorems are most useful in problems where we sum many independent, but not identically distributed random variables. Here's an example involving Lyapunov CLT.

Example 21.1 (Record values)**.** Suppose that we have an infinite sequence of i.i.d. random variables $X_1, X_2, \ldots$, which are *measurements*. We say that X_j sets a *record* if $X_j > \max(X_1, \ldots, X_{j-1})$. The probability of setting a record is strictly decreasing as time passes.[33] Let R_n be the number of records up to time n. Show that $R_n \sim \mathcal{N}(\log n, \log n)$. More formally,

$$\frac{R_n - \log n}{\sqrt{\log n}} \xrightarrow{D} \mathcal{N}(0,1).$$

For each j, let I_j be the indicator variable of X_j setting a record. Then, by a symmetry or exchange argument, $I_j \sim \text{Bern}(\frac{1}{j})$. Furthermore, each of these I_j are *independent*, which is a result due to Rényi. The intuition for them being independent is that given all values $I_1, \ldots, I_{j-1}$, we can rearrange all of the variables $X_1, \ldots, X_{j-1}$ in an arbitrary order to change the prior records. It doesn't matter which order these previous measurements were in; either way, X_j has to be greater than all of them to set a record.

Assuming that I_j are independent, we can apply Lyapunov CLT with $r = 3$ to justify the convergence in distribution. In this case, the Lyapunov condition is

$$\frac{\sum_{j=1}^n \mathbf{E}\left[\left|I_j - \frac{1}{j}\right|^3\right]}{s_n^3} \to 0,$$

where $s_n^2 = \mathbf{Var}[I_1 + \cdots + I_n]$. The top expression looks hard, but in this case the I_j are simply indicators, so it's easy to compute that

$$\mathbf{E}\left[\left|I_j - \frac{1}{j}\right|^3\right] = \left(1 - \frac{1}{j}\right)^3 \frac{1}{j} + \frac{1}{j^3}\left(1 - \frac{1}{j}\right) \le \frac{1}{j} + \frac{1}{j^3},$$

where we're choosing to use a crude bound for simplicity. Also,

$$s_n^2 = \sum_{j=1}^n \mathbf{Var}[I_j] = \sum_{j=1}^n \frac{1}{j} - \sum_{j=1}^n \frac{1}{j^2}.$$

Plugging all of this in, our Lyapunov condition becomes

$$\frac{\sum_{j=1}^n \frac{1}{j} + \frac{1}{j^3}}{\left(\sum_{j=1}^n \frac{1}{j} - \frac{1}{j^2}\right)^{3/2}} = \frac{\log n + O(1)}{(\log n + O(1))^{3/2}} \to 0.$$

Therefore, we've shown that CLT holds. The only remaining things to check are that $\mathbf{E}[R_n] = \log n + O(1)$ and $\mathbf{Var}[R_n] = s_n^2 = \log n + O(1)$, so we conclude by Slutsky's theorem.

[33] This leads to some interesting behavior. For example, what's the expected number of variables before the first value greater than X_1, i.e., the first record? This actually turns out to be ∞!

21.2 The Replacement Method

We're going to now discuss an interesting method used by Lindeberg and Lyapunov in the past. We will use this strategy to prove an i.i.d. version of the central limit theorem, but it can also be used more generally to prove Lindeberg's CLT.

Suppose that we have $X_1, X_2, \dots$ be i.i.d. random variables with mean 0 and variance 1. Let $S_n = X_1 + \cdots + X_n$. Also, suppose that we have i.i.d. standard normals $Z_1, Z_2, \dots \sim \mathcal{N}(0,1)$. The idea of the *replacement method* is to simply "install" Z_j by replacing X_j in the sum, swapping in the normals one-by-one.[34] It turns out that each of these steps is negligible both individually and as a whole on the final distribution, which implies $X_1 + \cdots + X_n \sim Z_1 + \cdots + Z_n \sim \mathcal{N}(0, n)$.

Let's go over each step of this argument in detail. We initially start by letting $T_0 = S_n$, and define $T_j = Z_1 + \cdots + Z_j + X_{j+1} + \cdots + X_n$ for each $1 \le j \le n$. To show convergence in distribution, we will use an equivalent definition of convergence in terms of expectation.

Proposition 21.2 (Convergence of expectations in distribution)**.** *If $Y_1, Y_2, \dots$ is a sequence of random variables and Y is a random variable, then $Y_n \xrightarrow{D} Y$ if and only if $\mathbf{E}[g(Y_n)] \to \mathbf{E}[g(Y)]$ for all suitable test functions g. For example, one class of test functions is $g_t(y) = e^{ity}$, which can be shown through uniqueness of characteristic functions. Another is the class of all bounded, $\mathcal{C}^k$ functions g for fixed $k \ge 0$, which includes "ramp" functions approximating an indicator.*

In the above proposition, using indicators gives us the original definition of convergence in distribution, but this creates big issues at discontinuities. Therefore, we prefer to apply the other kinds of test functions, which are continuous and simpler to use in an argument. Motivated by this, we'll show for all $\mathcal{C}^3$ functions g such that g, g', g'', and g''' are bounded that

$$\mathbf{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right) - g\left(\frac{T_n}{\sqrt{n}}\right)\right] \to 0.$$

The replacement idea manifests itself in a *telescoping series*. Observe that

$$g\left(\frac{T_n}{\sqrt{n}}\right) - g\left(\frac{T_0}{\sqrt{n}}\right) = \sum_{j=1}^{n}\left[g\left(\frac{T_j}{\sqrt{n}}\right) - g\left(\frac{T_{j-1}}{\sqrt{n}}\right)\right].$$

Each of the differences in this telescoping sum can be thought of as replacing X_j by Z_j. Anyway, using a third-order Taylor series expansion with error term, we can show that each of these differences is locally bounded by $O(n^{-3/2})$, and therefore the entire sum vanishes as $n \to 0$.

[34] Joe compares this argument to the ship of Theseus thought experiment.

22 November 19th, 2020

Today we cover some central limit theorems on sums of dependent random variables, and we begin our discussion of martingales.

22.1 Dependent Central Limit Theorems

First, we will introduce a famous dependent central limit theorem.

Definition 22.1 (Stationary sequence)**.** We call a sequence of random variables $X_1, X_2, \ldots$ *stationary* if for all k, the joint distribution of all length-k windows $(X_n, X_{n+1}, \ldots, X_{n+k-1})$ is the same for all starting indices n.

Definition 22.2 (m-dependence)**.** We saw that a sequence $X_1, X_2, \ldots$ of random variables is m-dependent if each element can only have dependence with other variables that are at most m apart. In other words, for all n,

$$(X_1, \ldots, X_n) \perp\!\!\!\perp (X_{n+m+1}, X_{n+m+2}, \ldots).$$

The motivation for the m-dependence definition is that it can be really useful for *time series* data. For example, when $m = 0$, we get ordinary independence. For larger values of m, we can imagine a "horizon" of observations in the past, which can influence our future observations.

Proposition 22.3 (m-dependent CLT)**.** *Let $(X_n)_n$ be a stationary, m-dependent sequence of random variables, such that $\mathbf{E}[X_j] = \mu$ and $\mathbf{Var}[X_j] = \sigma^2 < \infty$ for all j. Then,*

$$\sqrt{n}(\overline{X}_n - \mu) \quad \xrightarrow{D} \quad \mathcal{N}(0, \nu),$$

where the variance is

$$\nu = \sigma^2 + 2\sum_{j=2}^{m+1} \operatorname{Cov}(X_1, X_j).$$

Proof. The full proof of this theorem is very technical. However, we will provide an outline of the key idea, which involves an argument where we split up the series into two parts. This is called a "big block-little block" strategy.

Choose some value $k > 2m$, and divide up the sequence of variables $X_1, \ldots, X_n$ into alternating blocks of length $k - m$ (big block) and m (little block). The idea is that because of m-dependence, the sum of each of the big blocks constitutes an i.i.d. random variable. Meanwhile, as k grows larger, we can show that the little blocks contribute a negligible amount to the total sum of the series. Therefore, after many technical details, one can show convergence by piggybacking off of the standard CLT (Proposition 20.3) for big blocks, and concentration bounding the little blocks. □

Another common case of dependent random variables is when we sample n elements *without replacement* from a finite population of size N. Since we don't have replacement, the samples $Y_1, \ldots, Y_n$ must be dependent. We can prove finite population central limit theorems in this case, showing that $\overline{Y}$ is approximately normal, as $N, n \to \infty$ while maintaining that $n \ll N$.

One interesting duality in the finite population case is between $\overline{Y}_n$ and $\overline{Y}_{N-n}$. The distributions of these two means for samples of size n and $N - n$ have the exact same shape (just reversed). This leads to the interesting observation that as you increase the sample size in a finite population, the normal approximation gets more accurate, but once you increase it *too* far, the normal approximation once again decreases in accuracy, until finally you get a full census when $n = N$.

Other CLTs of interest are Markov chain CLTs, which are useful for proving facts about MCMC algorithms, and martingale CLTs, which we may cover at the end of the course if time permits. Anyway, that concludes our unit on central limit theorems!

22.2 Martingales

Now we discuss *discrete-time martingales*, which are a useful model of stochastic processes that maintain a certain "fairness" property.

Definition 22.4 (Discrete-time martingale)**.** We say that $X_1, X_2, X_3, \dots$ is a *martingale* with respect to another sequence $Y_1, Y_2, Y_3, \dots$ if for all n,

1. (Regularity) $\mathbf{E}[|X_n|] < \infty$,
2. (Measurability) $X_n \in \sigma(Y_1, \dots, Y_n)$,
3. (Fairness) $\mathbf{E}[X_{n+1} \mid Y_1, \dots, Y_n] = X_n$.

You can more generally think of $(X_n)_n$ as a martingale with respect to the *filtration* $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$, where $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$. This is a more general definition, but it's also more abstract.

Note. The etymology of the word "martingale" is complicated. In the context of probability theory, the *martingale* was a risky betting strategy where you double your bet after each loss. This would theoretically lead to a +\$1 payoff if you had infinite money, but in practice, you will eventually run out of money after enough consecutive losses.

Note. There are other interesting models of stochastic processes like Brownian motion, which we won't cover in this course. Brownian motion is a special kind of continuous-time martingale where the deviations are Markov and multivariate normal. This gives it particularly nice properties, but it also has some weird properties like being continuous everywhere yet differentiable nowhere.

Many stochastic processes will be both Markov chains and martingales. However, in general the Markov property (memorylessness) and martingale property are different, as martingales are allowed to depend on *all* previous events.

Definition 22.5 (Submartingale and supermartingale)**.** We call a sequence of random variables a *submartingale* if the third property above is replaced by $\mathbf{E}[X_{n+1} \mid Y_1, \dots, Y_n] \geq X_n$. On the other hand, it is a *supermartingale* if $\mathbf{E}[X_{n+1} \mid Y_1, \dots, Y_n] \leq X_n$.

Oftentimes we will just write "X_n is a martingale" without specifying the sequence Y_n. The following proposition justifies why this is unambiguous.

Proposition 22.6. *If $(X_n)_n$ is a martingale with respect to $(Y_n)_n$, then $(X_n)_n$ is also a martingale with respect to $(X_n)_n$.*

Proof. We can simply check the properties. The first two properties are trivial, while the third property can be verified by using Adam's law, since $\sigma(X_1, \dots, X_n) \subseteq \sigma(Y_1, \dots, Y_n)$. □

This is a relatively simple definition, and we'll see how it can be used to prove really nice facts about various processes, with machinery like Doob's optional stopping theorem, Azuma's inequality, Kolmogorov's inequality, and others.

Example 22.7 (Random walks are martingales)**.** If $X_1, X_2, \dots$ is a sequence of random variables that have mean 0, then $S_n = X_1 + \cdots + X_n$ is a martingale. Similarly, if $\mathbf{E}[X_j] \geq 0$, then S_n is a submartingale, and if $\mathbf{E}[X_j] \leq 0$, then S_n is a supermartingale.

23 November 24th, 2020

Today we continue to discuss martingales and their applications. The two key theorems in this area are the martingale convergence theorem and the optional stopping theorem.

23.1 Examples of Martingales

Recall that a martingale is any process that doesn't tend to either increase or decrease, so it won't drift in either direction on expectation. We will consider the following example now.

Definition 23.1 (SSRW)**.** A *simple symmetric random walk* is the sequence S_n for $n \geq 0$, where $S_n = X_1 + \cdots + X_n$ and the X_j are i.i.d. random signs.

Clearly, S_n is a martingale because it does not tend to drift in either direction. Also, S_n^2 is a submartingale because it tends to increase over time (and in general, you could replace x^2 with any other convex function), but we can "correct" the drift by taking $S_n^2 - n$, which is a martingale!

As an aside, there is a continuous analogy to the simple symmetric random walk called *Brownian motion*. If B_t is Brownian motion with $B_t \sim \mathcal{N}(0, t)$, where $B_t(\omega)$ for each $\omega \in \Omega$ is a continuous sample path, then B_t is a martingale, and furthermore $B_t^2 - t$ is also a martingale. Although stochastic processes are not the focus of this course, here is an interesting fact in this vein, which is a partial converse that characterizes the normal distribution.

Proposition 23.2 (Lévy)**.** *If X_t is a process with continuous sample paths, $X_0 = 0$, X_t is a martingale, and $X_t^2 - t$ is a martingale, then X_t is Brownian motion.*

Here's another cool fact about Brownian motion, since Joe likes it so much.

Proposition 23.3. *Brownian motion in two dimensions will write your name. In other words, for any time $\epsilon > 0$, Brownian motion B_ϵ will almost surely draw any continuous path with some error tolerance, after sufficient amounts of zooming in.*

This is quite hard to visualize because Brownian motion is "infinitely wiggly" with detail similar to a fractal. But you can generalize this fact to say some incredible things, such as, Brownian motion will write all of the complete works of Shakespeare in ϵ time. See infinite monkey theorem.

Let's return to the discrete setting now.

Example 23.4 (Pólya urn)**.** Suppose that you have an urn with balls of two different colors. We start with $a \geq 1$ orange balls and $b \geq 1$ blue balls. On each step of the process, we draw a single ball from the urn and replace it with *two* balls of that color. Let M_n be the proportion of orange balls at time n. Then, M_n forms a martingale, and in the limit distribution is

$$M_\infty \sim \text{Beta}(a, b).$$

This has applications to reinforcement learning, such as in multi-armed bandits.

Example 23.5 (Likelihood ratio)**.** There is a famous test in statistics known as the *likelihood ratio test*. Suppose that we have two possible density functions f and g, and we draw i.i.d. data samples $Y_1, \ldots, Y_n$. Consider the values

$$M_n = \frac{f(Y_1)}{g(Y_1)} \cdot \frac{f(Y_2)}{g(Y_2)} \cdots \frac{f(Y_n)}{g(Y_n)}.$$

Suppose that either f or g is the true density for Y_j. Then, if g is the true density of Y_j, then M_n is a martingale. When $Y_j \sim f$, we get that M_n is a submartingale.

Example 23.6 (Branching process)**.** Suppose that you have a process that spreads to many individuals through a tree structure, such as a viral disease or a family tree. We can write down the number of members in the process at time t as a stochastic process. Although this is not a martingale (it's increasing), it becomes a martingale after an appropriate rescaling.

Example 23.7 (Doob martingale)**.** Suppose that we have a random variable Y with $\mathbf{E}[|Y|] < \infty$. Then, $Z_t = \mathbf{E}[Y \mid \mathcal{F}_t]$ is a martingale with respect to filtration $\{\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \ldots\}$.

23.2 Martingale Convergence and Optional Stopping

A martingale is just a stochastic sequence, so sometimes we might ask if they converge almost surely. Unfortunately this is not always the case, as some martingales intuitively just "keep moving" like the SSRW example (Definition 23.1). The following theorem provides some conditions under which convergence happens.

Proposition 23.8 (Martingale convergence theorem)**.** *Let M_n be a submartingale such that*

$$\sup_n \mathbf{E}[|M_n|] \leq c,$$

for some constant $c < \infty$. Then, there exists a random variable M_∞ such that $M_n \to M_\infty$ almost surely, and also, $\mathbf{E}[|M_\infty|] < \infty$.

Proof. This is technical but interesting, and we'll try to cover it in the next lecture. □

Corollary 23.8.1. *A nonnegative supermartingale converges almost surely.*

The intuition behind this theorem is that submartingales are similar to a monotone increasing sequence in their convergence properties. Bounded monotone sequences must converge. Although submartingales are not strictly increasing because they have bumpiness, these bumps are not enough to significantly impact the convergence properties.

The next theorem is very useful for generalizing some intuitions about martingales not drifting in expectation, to the case where our stopping time may be unbounded. If M_n be a martingale, then it's easy to show that $\mathbf{E}[M_t] = \mathbf{E}[M_0]$ for any positive integer t. However, what if our stopping time t is a random variable, instead?

Proposition 23.9 (Optional stopping theorem)**.** *A random variable T supported on $\mathbb{N} \cup \{\infty\}$ is called a stopping rule with respect to $\{X_0, X_1, \ldots\}$ if for all n, the event $\{T \leq n\}$ lies in the sigma-algebra $\sigma(X_0, \ldots, X_n)$.*[35] *If M_n is a martingale and T is a stopping time, then $\mathbf{E}[M_T] = \mathbf{E}[M_0]$ if any of the following conditions holds:*

1. *(Bounded time). T is almost surely bounded by a constant $c \in \mathbb{N}$, i.e., $P(T \leq c) = 1$.*
2. *(Bounded space). M_n is almost surely bounded, i.e., $P(|M_n| \leq c) = 1$.*
3. *(Bounded increments). $|M_n - M_{n-1}| \leq c$ and $\mathbf{E}[T] < \infty$.*

Example 23.10. To show that at least one of these conditions is necessary, consider the SSRW S_n with stopping time $T = \inf\{n \in \mathbb{N} \mid S_n = 1\}$. This stopping time is almost surely finite by the martingale convergence theorem, so we have $\mathbf{E}[S_T] = 1$, but $S_0 = 0$. An interesting conclusion is that by the contrapositive of the optional stopping theorem, $\mathbf{E}[T] = \infty$.

[35] In other words, you can't use "psychic powers" to see into the future when deciding whether to stop at time t.

24 December 1st, 2020

This is the last week of classes. Today, we continue discussing martingales, and we'll prove the optional stopping theorem. This will give us some reusable tools that we can use more generally in martingale problems.

24.1 The Optional Stopping Theorem

Recall that we covered Proposition 23.9 last week, known as the *optional stopping theorem*, which provides sufficient conditions for $\mathbf{E}[M_T] = \mathbf{E}[M_0]$ where T is a stopping time. We gave three sufficient conditions, but in general, there are many variants on the same theorem.

Proof of Proposition 23.9. First consider the bounded time condition. If $T \le n$ almost surely, then we can write M_T in terms of a telescoping series with indicators,

$$M_T = M_0 + \sum_{j=1}^{T}(M_j - M_{j-1}) = M_0 + \sum_{j=1}^{n}(M_j - M_{j-1})I_{T\ge j},$$

where the second equality holds almost surely. If we show that the summation above has expectation zero, then we're done. To do this, we use Adam's law to get

$$\begin{aligned}\mathbf{E}[(M_j - M_{j-1})I_{T\ge j}] &= \mathbf{E}[\mathbf{E}[(M_j - M_{j-1})I_{T\ge j} \mid Y_0, \dots, Y_{j-1}]] \\ &= \mathbf{E}[I_{T\ge j}\,\mathbf{E}[M_j - M_{j-1} \mid Y_0, \dots, Y_{j-1}]] \\ &= 0.\end{aligned}$$

We can factor $I_{T\ge j}$ out of the conditional expectation because it is a stopping time, therefore in the sigma-algebra generated by events up to time $j-1$, and the last equality is just the definition of a martingale. This finishes the proof for the bounded-time case.

What about the second condition, where $|M_n| \le c$ almost surely for all n? For this, we will use a technique called *truncation*. Let $T_n = \min(T, n)$ for all n, so T_n is clearly a bounded stopping time. Furthermore, as $n \to \infty$, we have $T_n \to T$ almost surely. The idea of truncation is that we have the result $\mathbf{E}[M_{T_n}] = \mathbf{E}[M_0]$ in the *truncated* case, and we use a convergence theorem to deduce the same result in the *general* case. In this case, the bounded convergence theorem yields

$$\mathbf{E}[M_{T_0}], \mathbf{E}[M_{T_1}], \mathbf{E}[M_{T_2}], \dots \to \mathbf{E}[M_T].$$

Since we have that $\mathbf{E}\left[M_{T_j}\right] = \mathbf{E}[M_0]$ for all j by the telescoping argument above, we conclude that $\mathbf{E}[M_T] = \mathbf{E}[M_0]$. The third condition with bounded increments can be proven in a similar manner by using the dominated convergence theorem. □

Note that with minor modifications to the above proof, we can get variants of the optional stopping theorem for submartingales and supermartingales, where $\mathbf{E}[M_T] \ge \mathbf{E}[M_0]$ and $\mathbf{E}[M_T] \le \mathbf{E}[M_0]$ respectively.

Example 24.1 (Gambler's ruin)**.** Consider a simple symmetric random walk on $\mathbb{Z}$, starting at 0, with absorbing barriers at $-a$ and b. The position S_t at time t is a bounded martingale. Therefore, by the optional stopping theorem, the probability of being absorbed at a is $b/(a+b)$, while the probability of being absorbed at b is $a/(a+b)$.

Example 24.2 (Asymmetric random walk)**.** Consider the same problem as the previous example, but we instead have an unfair game where we win with probability $p \neq 1/2$ and lose with probability $q = 1 - p$. Then, $(q/p)^{S_t}$ is a martingale.

Example 24.3 ("Say red")**.** Consider a deck of cards in random order, with 26 red cards and 26 black cards. A dealer is flipping over cards, one at a time, and after each step they give you the option to stop. When you stop, the next card in the deck is revealed. You win that card is red. It turns out that no strategy for this game achieves a success probability different from 50%. This is because the fraction of red cards M_n left in the deck after n draws is a martingale, and it is also your success probability when stopping.

Joe mentions that the above example has shown up in many job interviews. Indeed, I remember Paul Christiano giving us this exercise as a brain teaser at SPARC. Another slick argument is to use the *interchangeability* of the cards, which says that choosing the top card of the deck is completely interchangeable with choosing the last card of the deck, and therefore none of your actions matter!

24.2 Doob's Martingale Inequality

Now we'll quickly see an application of martingales to concentration bounds. Let $X_0, X_1, X_2, \dots$ be a nonnegative sequence of random variables. Markov's inequality tells us that

$$P(X_n \geq a) \leq \mathbf{E}[X_n]/a.$$

However, if we additionally assume that X_n is a submartingale, we get a much stronger inequality.

Proposition 24.4 (Maximal inequality)**.** *If $X_0, X_1, \dots$ is a nonnegative submartingale, then*

$$P\left(\max_{0 \leq j \leq n} X_n \geq a\right) \leq \frac{\mathbf{E}[X_n]}{a}.$$

Proof. We will define a bounded stopping time T by

$$T = \min(\inf\{j \leq n : X_j \geq a\}, n).$$

Therefore, by the optional stopping theorem,

$$P\left(\max_{0 \leq j \leq n} X_n \geq a\right) \leq P(X_T \geq a) \leq \frac{\mathbf{E}[X_T]}{a} \leq \frac{\mathbf{E}[X_n]}{a}.$$

□

One application of this inequality is in proving the martingale convergence theorem.

25 December 3rd, 2020

Today is the last lecture of the course, before reading period! We will talk about a selection of topics, as requested by the students.

25.1 Completeness of Natural Exponential Families

We will prove a completeness property of certain natural exponential distributions.

Proposition 25.1. *Given a natural exponential family F_η, if there is a function h such that $\mathbf{E}[h(Y_\eta)] = 0$ for all η, where $Y_\eta \sim F_\eta$, then $h = 0$ almost everywhere on the support of Y_η.*

Proof. We can split h into positive and negative components, $h = h^+ - h^-$. Then we have

$$\int_{-\infty}^{\infty} e^{\eta y} h^+(y) f_0(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} e^{\eta y} h^-(y) f_0(y)\,\mathrm{d}y.$$

When $\eta = 0$, this means that $\int h^+(y) f_0(y)\,\mathrm{d}y = \int h^-(y) f_0(y)\,\mathrm{d}y = c$, so we can divide both sides of the above equation by c to turn this into a probability distribution. Then, the above integrals are precisely the moment generating functions of two distributions (according to LOTUS), so

$$h^+(y) f_0(y) = h^-(y) f_0(y)$$

almost everywhere, by the uniqueness of moment generating functions. Therefore, $h^+(y) = h^-(y)$ almost everywhere on the support of y, where $f_0(y) \neq 0$, so therefore $h(y) = 0$. □

25.2 Bounded Central Limit Theorem from Lindeberg

As an example of Lindeberg's CLT, we will prove a nice bounded central limit theorem.

Proposition 25.2. *Suppose that X_j are independent and have zero mean, such that $|X_j| \leq c$ almost surely for all j, where c is a constant. Then, assuming that the variance of the partial sum*

$$s_n^2 = \mathbf{Var}[X_1 + \cdots + X_n] = \sigma_1^2 + \cdots + \sigma_n^2 \to \infty,$$

we have $S_n/s_n \to \mathcal{N}(0,1)$ in distribution.

Proof. Let's verify Lindeberg's condition for the CLT. This is written as

$$\mathrm{Lind}_{n,\epsilon} = \sum_{j=1}^{n} \mathbf{E}\left[\left(\frac{X_j}{s_n}\right)^2 I_{|X_j|/s_n > \epsilon}\right] \to 0,$$

as $n \to \infty$, for any fixed value of ϵ. However, note that since $|X_j| \leq c$, the indicator random variable must be zero for all sufficiently large values of n. Therefore, the Lindeberg condition converges to zero for any choice of value for ϵ, as desired. □

25.3 Poisson Embedding for the Coupon Collector's Problem

Consider the following problem, which appeared on a past midterm.

Exercise 25.1 (b-coupon collector)**.** There are n empty boxes. Balls are put into the boxes one at a time, independently, with each ball equally likely to go into any of the boxes. Find the expected number of balls needed to make it so that all of the boxes have at least b balls each. Your answer can be left as a single integral. Hint: Imagine putting balls in boxes according to a Poisson process of rate 1.

This is a generalization of the so-called *coupon collector's problem* where $b = 1$, and the expected amount of time is simply equal to the n-th harmonic number $H_n = \sum_{j=1}^{n} \frac{1}{j}$. However, while the basic coupon collector's problem is easy, this problem is significantly more difficult. The really interesting thing is that although the problem is ostensibly discrete, adding the continuous-time Poisson process greatly simplifies the solution, as this distribution has nice properties.

25.4 Final Thoughts

That concludes the last lecture of Stat 210. Here's a quote from Joe, about the final exam:

> My belief is that probability is inexhaustibly rich. There will never be a shortage of interesting problems; it just might take a while to come up with them.

Probability is a vast subject. There's plenty of courses like Stat 212 that go further. Joe mentions that his book is quite different from the standard courses on the material, emphasizing probabilistic thinking and not just measure theory. That's it for the semester!

References

[BM20] J.K. Blitzstein and C. Morris. *Probability for Statistical Science*. Unpublished draft, 2020.