MAT 235A / 235B: Probability

Instructor: Prof. Roman Vershynin

2007-2008

Contents

OCTOBER 1, 2007

This Friday, we will have a quiz. It will be on undergraduate probability, and it will not count toward your grade. It will probably be on discrete probability, things like coins and balls in urns, and the like. So just typical problems. I will look over this quiz and next Monday, I will outline the solutions a little, perhaps place them online. You're also welcome to come to my OHs to discuss.

On the webpage, I'll post notes (just one page) on the interpretation of set theory in probability theory. Friday's quiz will take the whole time.

Now, we'll start on Lecture 2. So, we're trying to build a mathematical foundation into probabalistic thinking. How do we put mathematical rigor into the mathematical uncertainty? That's our program for the time. As you know, we will do it using *measure theory*.

For the measure theory, we'll need a big set Ω, called the *sample space*. The set Ω will be finite or infinite. It will list all possible outcomes of the experiment. When we talk about the *events* $A, B, \ldots \subseteq \Omega$, the events are subsets of Ω. The event that "you have two tails" will be one of the events. Not all possible collections of outcomes can be considered events! (More colloquially, not all possible events are allowed.) Thus $\mathcal{F}$ will be the collection of all permissible events.

It may be unclear what should be the permissible events, but it is going to become clear that the collection must satisfy certain rules. For example, if you can talk about A, you can also talk about the complement of A. There will be some rules on $\mathcal{F}$, and the simplest rules will lead to the concept of *algebras*.

Definition 0.1. *A collection $\mathcal{F}$ of events is called an algebra if*

1. *It is closed under complementation:*

$$A \in \mathcal{F} \;\Rightarrow\; A^c \in \mathcal{F}$$

2. *It is closed under taking finite union:*

$$A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}.$$

It is immediate from these two conditions that algebras are also closed under finite intersections:

$$A, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}.$$

Proof. De Morgan's Law states $(A \cap B)^c = A^c \cup B^c$. Thus, $A \cap B = (A^c \cup B^c)^c$.[1] □

Example 0.2.

- *Let $\mathcal{F} = \{\emptyset, \Omega\}$. This is the trivial algebra. In some sense, it is the "poorest" algebra.*
- *At the opposite extreme let $\mathcal{F} = \mathcal{P}(\Omega)$ the power set of Ω. This is the "richest" algebra.*

[1] So far so good?

The true life is between these extremes.

One more example: I want a natural algebra.

Example 0.3. *Take $\Omega = \mathbb{R}$. I want all possible intervals to be in the algebra. Because of the properties of Definition 0.1, we will have the smallest working $\mathcal{F}$ to be*

$$\mathcal{F} = \{\textit{finite unions of intervals in } \mathbb{R}\}.$$

Note that $\mathcal{F}$ will not contain sets that can not be expressed as the finite union of intervals in $\mathbb{R}$. The unfortunate result is that the concept of algebra is too weak for probability theory. It does not allow us to talk about these sets. It does not allow us to talk about limits, which is what we want. So, what is a remedy to this problem?

You can not put a probability measure on $\mathbb{R}$ such that all subsets will be measurable. Thus, arbitrary infinite unions will be too strong for probability theory.

What's the compromise? Only include **countable** unions (and intersections)! This is the dogma of probability theory, to include only countable unions. Then, the concept of algebra will then be replaced by the concept of *σ-algebras*.

Definition 0.4. *A collection $\mathcal{F}$ of events is called an σ-algebra if*

1. *It is closed under complementation:*

$$A \in \mathcal{F} \;\Rightarrow\; A^c \in \mathcal{F}$$

2. *It is closed under taking countable unions:*

$$A_1, A_2, \ldots \in \mathcal{F} \Longrightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}.$$

If we go back to our examples, which are σ-algebras? The first two are, but the third is not.

You should not worry too much about σ-algebras when working in discrete probability. Then, you can just take the power set for your algebra.

Let me illustrate this with a very concrete concept, which will only make sense in the settings of σ-algebras. This is the concept of the *limits of sets*. In both the cases of numbers and vectors, we have some topology. The meaning of closer is quantified by the topology. So this is some kind of limit without topology. So, here's a definition:

Definition 0.5. *Suppose $\mathcal{F}$ is a σ-algebra*[2]*. Let $A_1, A_2, \ldots \subseteq \mathbb{R}$.*

[2] on Ω. At this point, I'll stop specifying the universal set.

- *Then the*

$$\begin{aligned}\limsup A_n &:= \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k \\ &= \{\omega \in \Omega : \forall\ n\ \exists\ k \geq n \text{ such that } \omega \in A_k\} \\ &= \{\omega \in \Omega : \omega \text{ belongs to infinitely many } A_n\text{'s}\} \\ &= \text{"Infinitely many events occur".}\end{aligned}$$

- *Then the*

$$\begin{aligned}\liminf A_n &:= \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k \\ &= \{\omega \in \Omega : \exists\ n \text{ such that } \forall\ k \geq n,\ \omega \in A_k\} \\ &= \{\omega \in \Omega : \omega \text{ belongs to all except finitely many } A_n\text{'s}\} \\ &= \text{"All except finitely many events } A_n \text{ occur".}\end{aligned}$$

The only reason for A_n not to occur is some finite isolation. So these are two sets. Finally,

Definition 0.6. *If* $\limsup A_n = \liminf A_n$*, then we say*

$$A = \lim A_n = \limsup A_n = \liminf A_n$$

and we say that A_n *converges to* A *and write* $A_n \longrightarrow A$.

In order to talk about these sets, we need the unions and intersections in Definition 0.5 to exist. So, this is where we need the σ-algebra axioms.

Example 0.7. *Let's actually go from what we know about the real numbers* $\mathbb{R}$*. What's an example where we have the* $\limsup$ *and* $\liminf$ *are different? How about the sequence:*

$$a_n = \frac{1}{2} + \frac{(-1)^{n+1}}{2}$$

Then the $\limsup a_n = 1$ *and the* $\liminf a_n = 0$.

How can we use this to make a simple example for the case of sets?

Example 0.8. *Let* $A_1 = A_3 = A_5 = \cdots = A$ *and* $A_2 = A_4 = A_6 = \cdots = B$ *where* B *is a proper subset of* A.
Then $\limsup A_n = A$ *and* $\liminf A_n = B$.

OCTOBER 3, 2007

There will be a quiz this Friday. The quiz will be on undergraduate probability. It will not count toward your final grade. The TA has posted office

hours: My webpage always contains the most current information. The TA's office hours are on Tuesday.

One more announcement: We have a Research Focus Group (RFG) this year, led by Prof. Janko Gravner. The organizational meeting is today at 5:10 pm.

A little note about the text book: I will not be following exactly as you probably saw already. We will mostly follow the text book. The HW problems, which I'll post today right after lecture, will be from the book. Any more questions?

So, what are we doing so far? We're looking at foundations. We had Ω, an abstract set, called the *sample space*. It lists all possible elementary outcomes of your experiment. Elementary outcomes may be very well be small. For example, it's highly unlikely that your plane leaves in **exactly** 10 minutes. What we will do is look at special subsets of Ω: We are looking at some σ-algebra $\mathcal{F}$ of subsets of Ω. We will speak of the probability of the subsets $A \in \mathcal{F}$.

So why do we need the strange countable restriction? This is so that we can now talk about *limits* of events/sets. Recall Definition 0.5. Usually, in this hard definition, we are only considering sets that are increasing or decreasing families of sets.

Definition 0.9. *The sequence of events $A_1, A_2, \ldots$ is nonincreasing if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. In this case, we write that $A_n \searrow$.*

Similarly, the sequence of events $A_1, A_2, \ldots$ is nondecreasing if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$. In this case, we write that $A_n \nearrow$.

Proposition 0.10. *If $A_n \nearrow$, then A_n converges, and*

$$\lim A_n = \bigcup_{n=1}^{\infty} A_n = A.$$

If $A_n \searrow$, then A_n converges, and

$$\lim A_n = \bigcap_{n=1}^{\infty} A_n = B.$$

Proof. Exercise. □

Question: *(ambiguously stated)* Is A in the σ-algebra?

- **Answer:** From now on, whenever I say that A is an event, I implicitly mean that $A \in \mathcal{F}$.

Question: Are we assuming $\Omega \in \mathcal{F}$?

- **Answer:** This turns out to always be, as long as $\Omega \neq \emptyset$. Indeed, take $\{a\} \cap \{a\}^c = \emptyset \in \mathcal{F}$.

0.1 Generating σ-algebras

We'll have the following approach to generate a σ-algebra containing a certain collection $\mathcal{A}$ of subsets of Ω that we (initially) deem interesting:

- Start with an arbitrary collection $\mathcal{A}$ of subsets of Ω.

  Example 0.11. *If $\Omega = \mathbb{R}$, then $\mathcal{A}$ might consist of all intervals.*

- We want the smallest σ-algebra containing $\mathcal{A}$. That's what we actually do: We look at all σ-algebras that contain $\mathcal{A}$, and we pick the smallest[3].
- Choose the smallest one. Call it $\sigma(\mathcal{A})$, the *σ-algebra generated by $\mathcal{A}$*.

What is the smallest one? Why does the smallest one exist? Note, the arbitrary intersection of σ-algebras is again a σ-algebra. You should check this. Thus, we can actually define

$$\sigma(\mathcal{A}) = \bigcap_{\Sigma \text{ is a } \sigma\text{-algebra containing } \mathcal{A}} \Sigma.$$

This is a very non-constructive idea. You can't actually observe all σ-algebras in this intersection. But, heuristically, how do you **view** $\sigma(\mathcal{A})$? The σ-algebra $\sigma(\mathcal{A})$ is obtained by taking countably many unions, intersections and complements.

As an application, let's do this on the real line, continuing our previous example. This is an important standard example (in probability and analysis) called the *Borel σ-algebra on $\mathbb{R}$*.

Example 0.12. *Again, we had $\Omega = \mathbb{R}$ and $\mathcal{A} = \{(a, b) : -\infty < a < b < \infty\}$, the collection of open intervals*[4]*. Then $\sigma(\mathcal{A})$ is called the Borel σ-algebra, denoted by $\mathcal{R}$, at least in this text.*

Elements $A \in \mathcal{R}$ are called Borel sets. The pair $(\mathbb{R}, \mathcal{R})$ is called Borel space.

Many fractals are non-Borel sets. Even fractals are. The set $\mathcal{R}$ contains all "interesting"[5] sets on $\mathbb{R}$, except some fractals. In particular, it includes all open sets, all closed sets, and their "mixtures." This is an exercise in topology.

Remark 0.13. *Everything can be carried over into higher dimensions. In fact, it is true for all metric spaces. In particular, it is true for $\mathbb{R}^n$. Here, $\mathcal{R}$ is the σ-algebra generated by open sets.*

Also, it easily generalizes to intervals $\Omega = [0, 1]$ (or any other interval). Why? By restriction.

It is very natural in probability, analysis, and topology to just cut out a little window, and look at a subobject in your category.

[3] What do you do in practice? Keep taking elements in $\mathcal{A}$ and apply σ-algebra operations. Heuristically, this will work.

[4] Including the closed intervals would have been redudnant.

[5] Yes, this is an opinion.

Proposition 0.14. *Let $\mathcal{F}$ be a σ-algebra on Ω. Let $A \in \mathcal{F}$. Then $\mathcal{F}|_A = \{F \cap A : F \in \mathcal{F}\}$ is a σ-algebra on A.*

Applying Proposition 0.14 to $\Omega = \mathbb{R}$ and $A = [0,1] \in \mathcal{F}$, you obtain the Borel σ-algebra $\mathbb{R}|_{[0,1]}$ on $[0,1]$. You can (alternatively) do this by thinking of $[0,1]$ as a metric space and doing a direct construction.

I promise not to talk about σ-algebras at such length anymore, because soon we move into probability.

1 Probability Measures

Again, we have a sample space Ω and a σ-algebra $\mathcal{F}$ on Ω. Here's the bright idea: each probability should just be an assignment of numbers.

Definition 1.1. *A measure on $(\Omega, \mathcal{F})$ is a map $\mu : \mathcal{F} \to [0, +\infty]$ such that:*

- *For every disjoint events $A_1, A_2, \ldots \in \mathcal{F}$, one has $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$.*

This property is also is phrased by saying that μ is countably additive.

Definition 1.2. *If $\mu(\Omega) = 1$, then μ is called a probability measure. The triple $(\Omega, \mathcal{F}, \mu)$ is called a probability space.*

OCTOBER 5, 2007

Today, we had a quiz.

OCTOBER 8, 2007

I haven't look at the quiz yet, but I will and then I'll tell you something. I don't know what yet. By Friday. If you were trying to decide whether or not to take the course or not, we can look at your quiz together and decide. On Friday, I'll go over the quiz with you on the board, and we'll also post the solutions.

About the homeworks, some announcements. First of all, this I is the *indicator function*.

$$I(A) = \mathbb{I}_A$$

where

$$I(A)(x) = \begin{cases} 1, & x \in A. \\ 0, & x \notin A. \end{cases}$$

Now, about my office hours. Please come to my office hours. There are no other classes I'm teaching. If there's anything you'd like to discuss, please come. Today, I feel sick, so I don't know if I can survive office hours. Those of you who want to discuss anything today, will you be able to come right after this class?[6] Well, who can come from 2 to 2:30? Who can not come?

Question: What kind of convergence do we have on functions?

[6] Awww...

- **Answer:** As in for Exercise 7, it's point-wise.

Recall we are discussing measure. Let Ω be the sample space, and let $\mathcal{F}$ be a σ-algebra of subsets of Ω. Recall we had a definition: A *measure* is a function $\mu : \mathcal{F} \to [0, +\infty]$. So, for every event, we define a number. But, there's a condition: For every collection of disjoint events $A_1, A_2, \ldots \in \mathcal{F}$, one has

$$\mu\left(\bigcup_n A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

As an example of measure, we can take the cardinality. We can also take, in the case of $\mathbb{R}$, the length.

In the case of probability, we should only consider values of the measure μ to be between 0 and 1. Thus, if $\mu(\Omega) = 1$, then μ is called a *probability measure*. Usually, denoted by $\mathbb{P}$, to emphasize the probability. And, in this case, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a called a *probability space*.

You should think of the Ω, $\mathcal{F}$ and $\mathbb{P}$ as sort of different creatures: If you have fully-described this triple, then you know everything. That's why the discussion of probability theory is based on the discussion of probability space.

So, an example: a simple example is "You roll a dice." It has six facets, since it's a cube. Then the Ω in this case is the set

$$\Omega = \{1, 2, \ldots, 6\}.$$

So, the events are the things we want to ask of. In the discrete setting, we take everything, so this is just the power set. Our σ-algebra $\mathcal{F}$ is $P(\Omega)$. What is the probability that we'll have? Say our dice is fair. Then, we have

$$\mathbb{P}(A) = \frac{|A|}{6}.$$

Question: Are we just rolling the dice once? So, how come $\mathcal{F}$ is so big?

- **Answer:** For clarification, we can have other interesting events, like A is the event that "we get any even number". Of course, this is just in words. Mathematically, we say $A = \{2, 4, 6\}$. What is the probability $\mathbb{P}(A)$ here? It's $\frac{3}{6} = \frac{1}{2}$.

This is why we consider all of the power set. We are thus able to discuss any combination (such as $A = \{1, 4, 5\}$). So, this is the discrete probability model.

So, now we go into some elementary properties of probability measures, which are helpful when you get down to computing probabilities of different combinations of events:

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

   Proof. $\Omega = A \cup A^c$, a disjoint union. Hence, $\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$. Since $\mathbb{P}(\Omega) = 1$, we are done. □

2. $A \subseteq B$ implies $\mathbb{P}(A) \leq \mathbb{P}(B)$.

   Proof. $B = A \cup (B \setminus A)$, a disjoint union. By the additivity axiom, we have $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$, and $\mathbb{P}$ only takes on non-negative values. □

   Let's call this property is *monotonicity*.

3. Then, we have a similar property, called *sub-additivity*. I'll do it for two sets, just for simplicity:

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

   for arbitrary events A and B.[7]

   Proof. Note that $A \cup B$ can be written as a disjoint union $A \cup (B \setminus A)$. Hence, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \leq \mathbb{P}(A) + \mathbb{P}(B)$, with the inequality resulting from monotonicity. □

4. (Countable sub-additivity) Whenever you have sub-additivity of two sets, you can always add (countably many) more sets. Each time, you just use the additivity property for two sets. A little harder will be to do it for countable number of sets. So, the property is

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

   Proof. Exercise[8]. □

1.1 Continuity of Probability Measures

We have a concept of continuity of probability measures. There are two results on this. The first is:

Theorem 1.3. *(Continuity) Let $A_1, A_2, \ldots$ be a sequence of events. Then*

1. *If $A_n \nearrow A$, then $\mathbb{P}(A_n) \nearrow \mathbb{P}(A)$ as $n \to \infty$.*
2. *If $A_n \searrow A$, then $\mathbb{P}(A_n) \searrow \mathbb{P}(A)$ as $n \to \infty$.*

Proof. Let $A_1, \ldots$ be a sequence of sets. The general idea is to decompose the union $\bigcup_n A_n$ into "shells."

[7] Again, a Venn diagram may be helpful.

[8] I won't check if you've done this, but you really should do this problem, especially if your future work is in probability!

1. That is, let $B_1 = A_1$. And for $k \geq 2$, define $B_k = A_k \setminus A_{k-1}$. We have A is the disjoint union
$$A = \bigcup_n B_k$$
Hence, $\mathbb{P}(A_n) = \mathbb{P}(\bigcup_{k=1}^n)B_k)$, because the B_k's are disjoint. Then, we have $\mathbb{P}(\bigcup_{k=1}^n)B_k) = \sum_{k=1}^n \mathbb{P}(B_k)$, by additivity. All of these partial sums are bounded above by 1. Further, the sequence of partial sums are non-decreasing. Therefore, it converges (to their least upper bound). In particular, it converges from below to
$$\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k\right) = \mathbb{P}(A).$$

2. Exercise, done similarly.

Thus, both properties are proved for probability measures $\mathbb{P}$. □

There are uses for negative measures, in things like functional analysis. There, this kind of argument breaks down. The second property is:

Corollary 1.4. *(Continuity II) Let $A_1, A_2, \ldots$ be a sequence of events. Then*

1. $\mathbb{P}(\liminf A_n) \leq \liminf \mathbb{P}(A_n) \leq \limsup \mathbb{P}(A_n) \leq \mathbb{P}(\limsup A_n)$.
2. *As an immediate consequence, $A_n \to A$ implies $\mathbb{P}(A_n) \to \mathbb{P}(A)$ as $n \to \infty$.*

OCTOBER 10, 2007

1.2 Lebesgue measure

Let $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{R} = \{$ all Borel sets in $\mathbb{R}\}$. Let μ be the Lebesgue measure. Then, the theorem is

Theorem 1.5. *There exists a unique measure μ on $(\mathbb{R}, \mathcal{R})$ such that $\mu([a,b]) = b - a$ for all $a \leq b$.*

This measure is then called the *Lebesgue measure* on $\mathbb{R}$. It makes sense on intervals, but it's already mildly problematic if you want to talk about unions of intervals, due to convergence issues in infinite series. The construction is actually explicit, and this is half a course in measure theory.

Similarly, a result of this kind holds for the interval $\Omega = [0,1]$ instead of $\mathbb{R}$. Similar results hold also in higher dimension. Here, $\Omega = \mathbb{R}^n$. This measure is natural. In $\mathbb{R}^3$, what is it? It's volume. Here, we have
$$\mu\big([a_1,b_1] \times \cdots [a_n,b_n]\big) = \prod_{i=1}^n |b_i - a_i|.$$
Also similarly, instead of the whole space, we can look at $\Omega = [0,1]^n$, an n-cube.

1.2.1 Recurring Example: A delayed flight

Now, we can look at the example of the flight, and discuss the probability space there. This example was from Lecture 1. Suppose that a flight is delayed by at most one hour. So, we let $\Omega = [0, 1]$. The σ-algebra we consider is the collection $\mathcal{F} = \{\text{Borel subsets of } [0, 1]\}$. And then the measure[9] that we take is uniform (because we said that every possible moment of delay was equally likely). So, let $\mathbb{P}$ be the Lebesgue measure. Thus, the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is our probability space.

2 Conditional Probability

Now, we depart a little from measure theory. We'll discuss conditional probability. I'll assume that you have all seen this a little bit before. The occurrence of other events, and the knowledge of them, may affect the probability of our event occurring.

Definition 2.1. *Let A and B be two events, and suppose that $\mathbb{P}(B) > 0$. The conditional probability of A given B is defined to be*

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

The idea is that we zoom down from the whole space Ω to the subspace B. If we condition on the event B, that "Delta cancels its flights," then we see that A and B, so to speak, "heavily intersect."

2.0.2 Example: Longevity in Developing Countries

Let's do an example. Only half of the population of the developing countries lives longer than 40 years. The problem is that the child mortality rate is high, say 40%. We ask, "What is the probability that an adult (in these countries) lives longer than 40 years?"

Because the question asked about adults, we condition on the event that the people in question pass the childhood stage. Let B be the event that "the person does not die as a child" and A be the event that "the person lives longer than 40 years."

So, we wish to compute $\mathbb{P}(A|B)$, the probability that the average adult lives longer than 40 years, given that this adult survives childhood. Note that $\mathbb{P}(A) = 0.5$. The probability of B is $\mathbb{P}(B) = 0.6$. We need to compute the probability of $A \cap B$. What is this intersection? This is the event that both A and B happen, that the person lives longer than 40 years and does not die as a child. So, this is just A. In other words, $A \subseteq B$. So, the probability of $A \cap B$ is the probability of A, the smaller set. So, now it's easy to compute by Definition 2.1. We have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{0.5}{0.6} = \frac{5}{6} \approx 0.83.$$

[9] That is, the probability

Recall Bayes formula. This will give you a good reminder of how this whole conditional probability thing works.

3 Independence

Now for the major concept of probability theory. It's independence. Perhaps the most significant concept in probability theory is independence. It's sort of the analogue of continuity for analysis.

You probably know this concept. It's, in some sense, a negative statement.

Definition 3.1. *Events A and B are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$.*

Equivalently, people will sometimes give (based on the formula in Definition 2.1) an alternate definition

Definition 3.2. *Events A and B are independent if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Of course these are the same, but the first one seems more natural. The second is an easy by-product of the first. The second definition is nice because we can more easily add on more events. More generally,

Definition 3.3. *Events $A_1, A_2, \ldots, A_n$ are independent if for every collection of indices $1 \leq k_1 < k_2 < \cdots k_m \leq n$*

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m}) = \mathbb{P}(A_{k_1}) \cdot \mathbb{P}(A_{k_2}) \cdot \cdots \cdot \mathbb{P}(A_{k_m}).$$

We can extend this even to the countable intersections:

Definition 3.4. *Events $A_1, A_2, \ldots$ are independent if every finite sub-collection is independent (according to Defintion 3.3).*

3.0.3 Example: Tossing a coin

This is the canonical example of independence. So, suppose you toss a coin twice. Let A be the event "heads comes up in the first toss" and B be the event that "heads comes up in the second toss."

We compute: $\mathbb{P}(A) = \frac{1}{2}$ and $\mathbb{P}(B) = \frac{1}{2}$. To count the probability when both occur, we list out the situations and note that $\mathbb{P}(A \cap B) = \frac{1}{4}$. Thus A and B are independent.

Really, I don't know which way you're thinking of the logic: Either the probabilities compute the way they do and we conclude that A and B are independent, or we say that the events are independent and then compute these probabilities.

3.0.4 Example: Sampling

Suppose an urn contains 5 white balls and 5 black balls. Now we pick 2 balls without replacement. (That is, after picking up a ball, we do not put it back in the urn: we set it aside.) Let's look at similar events as before

$$A = \text{first ball is white}$$
$$B = \text{second ball is white.}$$

Are events A and B independent? No. We can say they are **not** independent. Why? After picking up a white ball, we can say that the urn will contain only 4 white balls and 5 white balls. So, $\mathbb{P}(B|A) = \frac{4}{4+5}$, and the probability of B itself (when it is not know what happens with the first ball) is $\mathbb{P}(B) = \frac{1}{2}$ (because $\mathbb{P}(B) = \mathbb{P}(B^c)$, and thus the roles of the colors are symmetric.).

But, on the other hand, if we pick **with** replacement, then the situation is the same as the coin tosses. Then events A and B will be independent. (The exercise is to check this!)

OCTOBER 12, 2007

There is a correllation on the weather with the number of people. The homework solutions will be posted online once the remaining people have submitted the assignment. Let's discus the HW policy: There's a 10% penalty every day past due. Because I also want to post the solutions online, I'll have to cut off the solution submission at some time. Let's say the absolute cut off is Friday, the end of the week. The reason for all of this is because there is just too much strain on the TA. Homework solutions will be posted online[10]. Quiz solutions are online. The quiz results were good: They were better than I expected. On the problem with the 10 keys, we should be careful about conditional probability: To open the door (within the first five keys) and to not open the door are not completely symmetric events. Note that the keys are not "replaced into the urn." Also, today's office hours are from 4 o'clock to 5 o'clock p.m.

For a pair of events A and B, we recall Definition 3.2 for the definition of independence. We can look at all of the combination of events, and we can ask questions about their independence too. So, we often think of *families* of events.

Definition 3.5. *Two families of $\mathcal{A}$ and $\mathcal{B}$ events (think σ-algebra) are independent if every two events $A \in \mathcal{A}$ and $B \in \mathcal{B}$ are independent.*

The definition requires that we check all possible pairs of events. Similarly, if you have more than two families, for any number of families (possibly even infinitely-many), one has the definition:

[10] I posted them already, but the permissions are not set.

Definition 3.6. *A collection of families $\mathcal{A}_1, \mathcal{A}_2, \ldots$ of events are called independent if every $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2, \ldots$ are independent.*

Proposition 3.7. *Let A and B be events. Let A_1 and A_2 be events.*

1. *If A is independent[11] of B, then A^c is independent of B.*
2. *Suppose A_1 and A_2 are disjoint[12]. If A_1 is independent of B and A_2 is independent of B, then $A_1 \cup A_2$ is independent of B.*

Proof. We'll check the first property.

1. We want to check that $\mathbb{P}(A^c \cap B) = \mathbb{P}(A^c) \cdot \mathbb{P}(B)$. We can compute

$$\begin{aligned}\mathbb{P}(A^c \cap B) &= \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) \text{ by the independence} \\ &= \mathbb{P}(B)\,(1 - \mathbb{P}(A)) \\ &= \mathbb{P}(B) \cdot \mathbb{P}(A^c)\end{aligned}$$

The second is left to the reader. □

As an exercise (following a student question to part 1 above), can you do this for more than one A?

The need for disjointness in part is pivotal. Pairwise independence (A independent of C, B independent of C) does not imply the full independence (A, B, C independent). See an example in the textbook. This will also work as a counter-example to the exercise of the footnote in part 2 of the proposition above.

You might think that this notion of independence is now too subtle to know how to deal with in combinations, but we do have the following:

Theorem 3.8. *Suppose families of events $\mathcal{A}$ and $\mathcal{B}$ are closed under intersections (that is, $A_1, A_2 \in \mathcal{A} \Rightarrow A_1 \cap A_2 \in \mathcal{A}$). If $\mathcal{A}$ and $\mathcal{B}$ are independent, then[13] $\sigma(\mathcal{A})$ and $\sigma(\mathcal{B})$ are independent.*

This is a very strong result, and it's a non-trivial result. So, we'll prove it. This is the first hard result of the course.

Proof. 1. First, we have a reduction. It suffices to prove that $\sigma(\mathcal{A})$ and $\mathcal{B}$ are independent (because afterwards, we reapply this result, to $\mathcal{B}$ and $\sigma(\mathcal{A})$, in that order).

[11] Here, I mean the pair A and B are independent. The relation is symmetric, so sometimes we use this kind of language that makes it sound one-sided, when really it's not.

[12] If A_1 and A_2 are **not** disjoint, then this property may fail! This will be the first non-trivial exercise: Construct an example to show that just individual independence will not be enough. Similarly for the intersection, the property will fail. This is very interesting: A_1 is independent of B, A_2 is independent of B, yet their union will be correlated! The example will not be exotic: you can do it with (around four) balls in an urn, for example.

[13] any combination of events form independent events, namely,

2. We will go around this problem, and consider the form of this set $\sigma(\mathcal{A})$. We will consider a family, a subset of $\sigma(\mathcal{A})$ for which we will know some property. To slow down, we want to show:

   For all $A \in \sigma(\mathcal{A})$ and for all $B \in \mathcal{B}$, events A and B are independent.

   So, fix $B \in \mathcal{B}$. Consider

   $$\mathcal{F}_{\mathcal{B}} := \{A \in \sigma(\mathcal{A}) \text{ such that } A, B \text{ are independent}\}.$$

   We know that $\mathcal{A} \subseteq \mathcal{F}_{\mathcal{B}} \subseteq \sigma(\mathcal{A})$.

3. It suffices to show that

   $$\mathcal{F}_{\mathcal{B}} \text{ is a } \sigma\text{-algebra} \tag{1}$$

   since $\sigma(\mathcal{A})$ is the smallest σ-algebra containing $\mathcal{A}$).

   This is nice: We will not even touch $\sigma(\mathcal{A})$ itself, other than to use this property. We just have some simple problem[14].

   Now, we prove property 1.

   (a) Let $A \in \mathcal{F}_{\mathcal{B}}$. Then $\Rightarrow A^c \in \mathcal{F}_{\mathcal{B}}$, by Property 1.

   (b) $A_1, A_2, \ldots \in \mathcal{F}_{\mathcal{B}} \Longrightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}_{\mathcal{B}}$? Well, it is enough to show $A_1, A_2, \ldots$ independent of $B \Longrightarrow \bigcup_{n=1}^{\infty} A_n$ is independent of B.

   - Can we do this for finite unions first? We define a disjoint collection

     $$\begin{aligned} \overline{A_1} &= A_1 \\ \overline{A_2} &= A_2 \setminus A_1 \\ \overline{A_3} &= A_3 \setminus (A_1 \cup A_2) \\ &\vdots \end{aligned}$$

     First, note that

     $$\bigcup_n A_n = \bigcup_n \overline{A_n}.$$

     [PROOF POSTPONED.]

□

OCTOBER 15, 2007

The HW solutions are posted online. Also, the quiz solutions are there.

We go back to our problem, which was what we can do with independence. We go back to our theorem from last time, which was about the power of combining events.

[14] We avoid some ϵ-δ-like annoyingness brought to us courtesy of analysis.

Theorem 3.9. *Suppose that families of events $\mathcal{A}$ and $\mathcal{B}$ are independent, and both $\mathcal{A}$ and $\mathcal{B}$ are closed under finite intersections. Then, $\sigma(\mathcal{A})$ and $\sigma(\mathcal{B})$ are independent.*

It's a pretty general theorem. As a specific example, you can think of $\mathcal{A}$ and $\mathcal{B}$ being some intervals in $\mathbb{R}$. Not necessarily all intervals. A note on the Borel σ-algebra: As soon as you know some property on intervals, you can extend it to all measurable sets.

So, recall the wrong proof. The first reduction was that we don't have to do it all at once. In other words, it's enough to prove that $\sigma(\mathcal{A})$ and $\mathcal{B}$ are independent. It means that every independent in $\sigma(\mathcal{A})$ is independent of any event $B \in \mathcal{B}$. Fix an event $B \in \mathcal{B}$. So, we want to prove that every event in $\sigma(\mathcal{A})$ is independent of B.

We do not even know the form of a general set in the σ-algebra! How do we go about this? We just use the minimality property of the σ-algebra, and we don't rely on any form of the set. So, we look at all of the events that are independent: That is, consider

$$\mathcal{F}_{\mathcal{B}} := \{A \in \sigma(\mathcal{A}) \text{ independent of } B\}. \tag{2}$$

We want to show that $\mathcal{F}_{\mathcal{B}} = \sigma(\mathcal{A})$. It is enough to prove that $\mathcal{F}_{\mathcal{B}}$ is a σ-algebra, because $\mathcal{F}_{\mathcal{B}} \supseteq \mathcal{A}$ and $\sigma(\mathcal{A})$ is the minimal σ-algebra with this property implies that $\sigma(\mathcal{A}) \subseteq \mathcal{F}_{\mathcal{B}}$.

So, how do we prove (2)?

Proof. We just check that the properties of a σ-algebra hold:

1. If $A \in \mathcal{F}_{\mathcal{B}}$, then $A^c \in \mathcal{F}_{\mathcal{B}}$, by "Property 1" of the last lecture.
2. Suppose that $A_1, A_2, \ldots \in \mathcal{F}_{\mathcal{B}}$. Then, we want

$$\bigcup_{n=1}^{\infty} A_n \ \in \ \mathcal{F}_{\mathcal{B}}.$$

   This is where things got out of control. Why is this hard? If the A_n's are disjoint, then this is true. Why? There was "Property 2" of the last lecture. If two events are disjoint and independent of B, then their union is as well. You'd also have to check the limiting of this procedure, but the continuity would end up telling you that this is okay.

   But they may not be disjoint. We only know that A_n are elements of $\sigma(\mathcal{A})$. Our intersection property was true only for the elements of $\mathcal{A}$. So, we can not just take intersections. So, how do we get around this obstacle? One possibility is to give up. Another possibility is to change the definition of σ-algebra: Let us require that the sets that we union must be **disjoint**. Then, we'll prove that this new definition of σ-algebra is the same as the old, provided that you have the property of "closed under intersection."

   So, we change the definition of σ-algebra, adding disjointedness.

Definition 3.10. *(Dynkin Systems) A collection $\mathcal{D}$ of sets is a Dynkin system if:*

(a) $A \in \mathcal{D} \Rightarrow A^c \in \mathcal{D}$.

(b) $A_1, A_2, \ldots \in \mathcal{D}$ *are disjoint* $\Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{D}$.

Theorem 3.11. *(Monotone Class Theorem) If $\mathcal{A}$ is closed under intersections, then $\sigma(\mathcal{A}) = \mathcal{D}(\mathcal{A})$, where $\mathcal{D}$ is the smallest Dynkin system containing $\mathcal{A}$.*

So, for free, we can replace σ-algebras with Dynkin systems. So, in order prove Theorem 3.9, we can replace $\sigma(\mathcal{A})$ and $(\mathcal{B})$ by $\mathcal{D}(\mathcal{A})$ and $\mathcal{D}(\mathcal{B})$ by Theorem 3.11. Then, the wrong proof will now hold (because of disjointedness).

So, what we've done in the proof is to verify the axioms of the Dynkin system. □

This Dynkin system notion is a bright idea from measure theory that finishes off the proof. As a simple corollary of Theorem 3.9, we have

Corollary 3.12. *Let $A_1, A_2, \ldots$ be independent events. If $I \cap J = \emptyset$, then $\sigma(A_k : k \in I)$ and $\sigma(A_k : k \in J)$ are independent.*

That is, no matter how you split of the two groups into a partition, then their sigma algebras will be independent.

This corollary is not immediate, right? We don't have the closed under intersection property. So, what do you do there? You may not have all intersections. So, you simply add them in:

Proof. Define $\mathcal{A} = \{$ all finite intersections of $A_k, k \in I\}$, and define similarly: $\mathcal{B} = \{$ all finite intersections of $A_k, k \in J\}$ But now, how do we know that $\mathcal{A}$ and $\mathcal{B}$ are independent? Let's leave this as an exercise[15]. □

That may all seem very abstract, but now we go to a significant application of this in probability theory, which is called Kolmogorov's 0-1 Law.

4 Kolmogorov's 0-1 Law

In physics, people usually call this *phase transitions*. Think of water and ice. This is trying to codify that you really don't have observed middle ground.

Let $A_1, A_2, \ldots$ be arbitrary events. Every other event that is correlated with these events should have probability either 0 or 1. Our event A will depend on all of these $A_1, A_2, \ldots$. But, how do we say that? We try to say something happens differently if we were to, say, leave off the first 9 events. So, we introduce a new notion:

[15] This is a very nice exercise. You should use the independence of all of the sets $A_1, A_2, \ldots$.

Definition 4.1. *The tail σ-algebra is*

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

The elements of $\mathcal{T}$ are called tail events.

So, the tail events are the on

Example 4.2. *In a series of coin tosses, the event that "A sequence of 100 consecutive heads occurs infinitely often" is a tail event.*
A_n is "Heads in n^{th} toss."

Now, the famous theorem:

Theorem 4.3. *(Kolmogorov's 0-1 Law) Let $A_1, A_2, \ldots$ be independent events. Then every tail event A has probability either 0 or 1.*

So, in that example, the probability that you'd have 100 consecutive heads infinitely-often is either 0 or 1. Who thinks it's 0? How about 1? Yes, it's 1. Try to prove it.

Proof. Let $A_1, A_2, A_3, \ldots, A_{n-1}, A_n, A_{n+1}, \ldots, \ldots$ be events (and our event A is concerned with $A_n, A_{n+1}, \ldots$. Then, by Corollary 3.12, $\sigma(A_1, \ldots, A_{n-1})$ and $\sigma(A_n, \ldots A_{n+1}, \ldots)$ are independent. Note, $A \in \sigma(A_n, \ldots A_{n+1}, \ldots)$. Thus, the events $A_1, A_2, \ldots, A_{n-1}, A$ are independent.

Thus, the sequence $A_1, A_2, \ldots, A$ are independent. Then, by Corollary 3.12, $\sigma(A_1, A_2, \ldots)$ and $\sigma(A)$ are independent. But $A \in \sigma(A_1, A_2, \ldots)$ and $A \in \sigma(A)$, which are independent collections.

Thus, A is independent of itself. Thus

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2.$$

and so $\mathbb{P}(A)$ is either 0 or 1. □

4.1 Applications: Percolation Theory

Suppose, given porous material, some liquid is poured on top. Will it reach the bottom? How fast?

In mathematics, we think of porous material as a lattice, say $\mathbb{Z}^2$. We connect neighboring vertices of this lattice independently with some probability p. The result is a graph. The problem about this graph is that: does it contain an infinite cluster? That is, does that complement (the "empty part") contain an infinite cluster? Here is an example (this looks like a maze).

The elementary events will be the connections between the neighbors. If you know the events, then you know exactly the maze. If you know about some local neighborhood, it does not containing the infinite cluster. Thus, this (containing an infinite cluster) is a tail event. The A_n's are whether or not the n^{th} edge is connected (after choosing some enumeration of the edges). By the 0-1 Law, the

probability of an infinite cluster is either 0 or 1. So there is a critical probability p_c below which the probability is 0 and above which the probability is 1.

So, this justifies the phase transition property in physics. It is very hard to prove that $p_c = \frac{1}{2}$. It was proved in the 1980's. This is a difficult problem of percolation theory. What happens in $\mathbb{Z}^3$, for example, is not known.

People in percolation theory are lucky because they can borrow experimental results (i.e. intuition for the value p_c) from physics.

OCTOBER 17, 2007

Did you take your old homework? Your graded homework is here. Otherwise, you can just get it in my office. New homework will be online today. On the first homework, the solutions are posted online. Solutions for the homework turned in today will be online on Friday. Today, we begin a new big topic on probability theory (in chapter 2 of the textbook).

5 Random Variables

We often are not interested in the specifics of the probability space. We do not want to be bothered with what's heads and what's tails. For example, you ask yourself, "How many heads occur on average in n tosses?" How do you solve it? Our sample space $\Omega = \{H, T\}$ is the outcome of every (individual) experiment. In every experiment, we do not look at "heads" and "tails" as items by themselves, but we associate numbers to heads and tails. We assign:

$$X : \left\{ \begin{array}{l} H \to 1 \\ T \to 0 \end{array} \right\}$$

This is called a *random variable*.

Why do we do this? Then, instead of counting heads, we can sum up numbers. So, let $X_1, X_2, \ldots$ be independent copies of the random variable X. Then, the number of heads in n tosses is conveniently represented by:

$$\#(heads) = \sum_{k=1}^{n} X_k.$$

Now, we can take expectation on both sides

$$\mathbb{E}\#(heads) = \mathbb{E}\sum_{k=1}^{n} X_k.$$

and now, we can swap sum and expectation (a dream!) and get

$$\sum_{k=1}^{n} \mathbb{E}X_k = \sum_{k=1}^{n} \frac{1}{2} = \frac{n}{2}.$$

So, we'll do this in much bigger generality, and first study the notion of *measurable functions*, functions from one probability space to another probability space.

5.1 Measurable Functions

Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be probability spaces. We often call elements of $\mathcal{F}_1$ and $\mathcal{F}_2$ *"measurable sets"*. In the case of probability spaces, these are the events.

Now, we consider a different function. Consider a function

$$f : \Omega_1 \to \Omega_2$$

In general, the trouble is that the function f better respect measurability. Why so? Because Ω_1 and Ω_2 are not abstract sets: they come with a structure (from the σ-algebra). For example, if you have a topological space Ω_1 and Ω_2, the concept of goodness is continuity. How do we do this in analysis? We have pullbacks of open sets. We do the same thing here, but we replace the open sets by measurability.

Definition 5.1. *A function $f : \Omega_1 \to \Omega_2$ between measure spaces is called measurable if the preimage of every measurable set is measurable.*

In mathematical terms, we summarize:

$$A \in \mathcal{F}_2 \implies f^{-1}(A) \in \mathcal{F}_1.$$

Recall that the preimage $f^{-1}(A) := \{\omega \in \Omega_1 : f(\omega) \in A\}$.

So, if you're familiar with a little bit of topology or analysis, then this is the same idea, but just on a different structure.

Suppose we have two functions f and g, and they are very similar, except on a finite number of points. We'll try to ignore what they do on this small set of measure zero. So, if you have two functions that differ on a small set, we'll think of them as being the same. This is formalized in the notion of *equivalence*.

Definition 5.2. *Measureable functions f and g that differ on a null set (that is, a set of measure zero) are called equivalent. That is,*

$$\mu_1\big(\{\omega \in \Omega_1 : f(\omega) \neq g(\omega)\}\big) = 0.$$

Such f and g are also called equal almost surely. Or, in analysis, people sometimes write

$$f \overset{a.e.}{=} g,$$

where "a.e." stands for almost everywhere.

Proposition 5.3. *A composition of two measurable functions is measurable.*

Proof. Exercise. □

Just check it like for continuous functions.

This is all too abstract. So, now, we want to do this for random variables. What is a random variable? It is a measurable map from a probability space into $\mathbb{R}$. It's a special case of Definition 5.1.

Definition 5.4. *A random variable is a measurable function on a probability space.*

Usually, random variables are denoted by capital Roman letters. For example,

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{R}, \mu),$$

where μ is the Lesbesgue measure, but it doesn't really matter. It's usually Lesbesgue measure. What's really important is the probability $\mathbb{P}$. For $\mathbb{R}^n$, X is called a *random vector*.

5.1.1 Examples

1. The first example is a coin toss. So, $\Omega = \{H, T\}$ and $\mathcal{F} = \mathcal{P}(\Omega)$, the power set. And, the measure $\mathbb{P}$ is the uniform measure on Ω, which means it gives the same weight to each outcome. Thus $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}$.

   Then, our random variable

   $$X : \Omega \to \mathbb{R}$$

   is defined by the following values:

   $$X(H) = 1,$$
   $$X(T) = 0.$$

2. The next example is the indicator of an event A. This is usually denoted by $X = 1_A$, the book denotes it by $I\{A\}$. Thus, this is the function

   $$X(\omega) = \begin{cases} 1 & , \quad \omega \in A \\ 0 & , \quad \omega \notin A. \end{cases}$$

   So, why is this measurable? The Borel set does not contain 0 or 1, then the preimage is empty. If it contains 0 but not 1, then the preimage is A^c. If it contains 1 but not 0, then the preimage is A. If it contains both 0 and 1, then the preimage is Ω. This verifies that X is measurable.

3. A little more advanced example are the *simple random variables*. Instead of having the image of X be a two-element set (like $\{0, 1\}$), we allow some other finite set.

   For example, let $X = \sum_{k=1}^{n} a_k 1_{A_k}$, where a_k are some weights and A_k are disjoint events that partition Ω:

   $$\bigcup_{k=1}^{n} A_n = \Omega.$$

   For every A_k, we assign some value a_k. Again, you can check that simple random variables are random variables indeed.

So you have a map, and how do you check it's measurable. Well, you should check for each Borel set, but that's too much work. What if you just checked on intervals? Is this enough? Yes. In fact, it's true in general. So, it's enough to check on the generators of a σ-algebra.

It suffices to check the measurability property of a map on generators of $\mathcal{F}_2$.

Theorem 5.5. *Consider a function $f : (\Omega_1, \mathcal{F}_1, \mu_1) \to (\Omega_2, \mathcal{F}_2, \mu_2)$, where $\mathcal{F}_2 = \sigma(\mathcal{A})$. Suppose $f^{-1}(A)$ is measurable for every $A \in \mathcal{A}$. That is,*

$$A \in \mathcal{A} \Rightarrow f^{-1}(A) \in \mathcal{F}_1. \tag{3}$$

Then f is measurable:

$$A \in \sigma(\mathcal{A}) \Rightarrow f^{-1}(A) \in \mathcal{F}_1. \tag{4}$$

This simplifies our life a lot, in the future. If we can not use this trick, then life will be complicated. Why? It's because we don't know the form of everything. This is analogous to the concept of bases in algebra. So, how can we do that? We just use the minimality in the definition of σ-algebra.

Proof. Consider $\mathcal{F} := \{A \in \sigma(\mathcal{A}) : f^{-1}(A) \in \mathcal{F}_1\}$. We wish to show that $\mathcal{F} = \sigma(\mathcal{A})$.

We know that $\mathcal{A} \subseteq \mathcal{F} \subseteq \sigma(\mathcal{A})$. Recall that $\sigma(\mathcal{A})$ is the minimal σ-algebra containing $\mathcal{A}$. So, if we show that $\mathcal{F}$ is a σ-algebra, then all three of these are actually equal (and there is no "floating around").

Now, we just use the definition of σ-algebra. We do not even mention generators: they were just part of this trick. We check:

1. We check $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$. Note that $f^{-1}(A^c) = (f^{-1}(A))^c$, which you have to check. And this, proves the first point.
2. Secondly, we check $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$. So, we need something like
$$f^{-1}\left(\bigcup_n A_n\right) = \bigcup_{n=1}^{\infty} f^{-1}(A_n)$$
Again, you'll just need to check, by definition of the preimage. This proves the second point.

Thus, $\mathcal{F}$ is a σ-algebra and $\mathcal{A} \subseteq \mathcal{F} \subseteq \sigma(\mathcal{A})$ actually implies $\mathcal{A} \subseteq \mathcal{F} = \sigma(\mathcal{A})$. □

Corollary 5.6. *Consider a function $X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$. Then X is a random variable if and only if*[16] *the set $\{\omega : X(\omega) \leq a\}$ is an event for every $a \in \mathbb{R}$.*

Proof. Use Theorem 5.5 with

$$(\Omega_1, \mathcal{F}_1, \mu_1) = (\Omega, \mathcal{F}, \mathbb{P})$$
$$(\Omega_2, \mathcal{F}_2, \mu_2) = (\mathbb{R}, \mathcal{R}, \mu)$$

and recall that $\mathcal{R} = \sigma\big((-\infty, a], a \in \mathbb{R}\big)$. Then $X^{-1}((-\infty, a]) = \{\omega : X(\omega) \leq a\}$. □

[16] also abbreviated as "iff".

So, this is a very convenient way to check that X is a random variable.

OCTOBER 19, 2007

People who run this lab have asked us to turn off cell phones every time we enter the room. Please power them off. My office hours are posted: please come and ask us questions. That said, I will not be able to answer e-mailed questions. On the web page, you'll find notes for the class in PDF format.

See http://www.math.ucdavis.edu/~ekim/classes/235/

5.2 Functions of Random Variables

Recall that a function $X : \Omega \to \mathbb{R}$ is a *random variable* if for every Borel set $A \subseteq \mathbb{R}$,

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} = \{X \in A\} \tag{5}$$

is an event (is measurable). Typically in probability theory, we drop mention of ω, so this explains the notation of the middle and right sets in (5).

Proposition 5.7. *If $X_1, \ldots, X_N$ are random variables, then $(X_1, \ldots, X_n)$ is a random vector.*

What do we have to do? We have to take a Borel set in the higher-dimensional space, and pull back. Again, that's hard: we don't know what the Borel sets in higher-dimensional spaces look like. But we can't just use our knowledge about $\mathbb{R}^1$. So, we can use the theorem from last time. We can check just on generators, as was given in the corollary from last lecture.

Proof. In view of Corollary 5.6 from the last lecture, it suffices to check the measurability of the map

$$\omega \to (X_1(\omega), \ldots, X_n(\omega))$$

on some generator of Borel σ-algebra in $\mathbb{R}$ such as

$$\{(-\infty, a_1] \times (-\infty, a_2] \times \cdots \times (-\infty, a_n] : a_1, a_2, \ldots, a_n \in \mathbb{R}\}.$$

Consider the preimage:

$$\begin{aligned}
&\{(X_1, \ldots, X_n) \in (-\infty, a_1] \times (-\infty, a_2] \times \cdots \times (-\infty, a_n]\} \\
&= \{X_1 \in (-\infty, a_1], \ldots, X_n \in (-\infty, a_n]\} \\
&= \bigcap_{k=1}^{n} \{X_k \in (-\infty, a_k]\} = \bigcap_{k=1}^{n} \{X_k \leq a_k\}.
\end{aligned}$$

So, each set $\{X_k \leq a\}$ is measurable (since X_k is a random variable), so their intersection must be measurable too, by the σ-algebra properties.

□

It's a nice illustration of how generators help us.

Theorem 5.8. *(Functions of Random Varibles) Let $X_1, \ldots, X_n$ be random variables and $f : \mathbb{R}^n \to \mathbb{R}$ be a (Borel) measurable function. Then, $f(X_1, \ldots, X_n)$ is a random variable.*

So, as an example of using this, if X is a random variable, then X^2 and $\sin(x)$ are random variables. The function

$$X_1 + \cdots + X_n$$

is a random variable if $X_1, \ldots, X_n$ are random variables.

So, how do we prove Theorem 5.8? We have a composition of measurable functions.

Proof. $f(X_1, \ldots, X_n)$ is a composition of two measurable functions:

1. $$\omega \mapsto (X_1(\omega), \ldots, X_n(\omega))$$
   is measurable by Proposition 5.7.

2. The function
   $$(x_1, \ldots, x_n) \mapsto f(x_1, \ldots, x_n)$$
   is measurable by the assumption.

The composition of two measurable maps is a measurable map (cf. Problem 6). □

Question: So a measurable map is the same thing as a measurable function?

- **Answer:** Yes, we usually say measurable function, however. We use term *random vector* when the codomain is more-than one-dimensional.

We'll certainly need this when we talk about random series:

Theorem 5.9. *(Limits of random variables) Let $X_1, X_2, \ldots$ be random variables. Then, $\sup X_n$, $\inf X_n$, $\limsup X_n$, $\liminf X_n$, and $\lim X_n$ are random variables when they exist.*

Proof. 1. Proof for $\sup X_n$: By the Corollary in the last lecture, it suffices to prove that $\{\sup X_n \leq a\}$ is a random variable $\forall\ a \in \mathbb{R}$. But, this is equal to
   $$\bigcap_{n=1}^{\infty} \{X_n \leq a\},$$
   which is indeed measurable (as in the Proposition).

2. Proof for $\inf X_n$: Exercise[17].

[17] It's not so trivial: You can't just mimick the proof for part 1

3. Proof for $\limsup X_n$: Note that this set is just

$$\inf_n \sup_{k \geq n} X_k$$

and then we use parts 1 and 2.

□

More examples (that follow now from this Corollary) include:

- If $X_1, X_2, \ldots$ are random variables, then so is their infinite sum

$$\sum_{n=1}^{\infty} X_n.$$

Why? The value of an infinite series is the limit of partial sums, and the partial sums are each random variables by the corollary.

Question: How, starting from random variables, can we recover the σ-algebra they came from? Originally, we have Ω and $\mathcal{F}$, and we had a good map (the measurable function X). Now, how do we go backwards?

In our "observables" in the physical sense, we can measure the value of random variables. We don't really have a full description of Ω and $\mathcal{F}$. How do random variables generate σ-algebras?

5.3 Random Variables generate σ-algebras

Let X be a random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 5.10. *The σ-algebra on Ω generated by X, denoted by $\sigma(X)$ is*

$$\begin{aligned}\sigma(X) &:= \{X^{-1}(A) : A \text{ Borel in } \mathbb{R}\} \\ &= \{\{X \in A\} : A \text{ Borel in } \mathbb{R}\}\end{aligned}$$

- $\sigma(X)$ is a σ-algebra indeed. (CHECK)
- $\sigma(X)$ is the smallest σ-algebra on Ω that makes X measurable (that is, a random variable)[18].

This may seem a bit too abstract, so let's do some examples.

1. Let X be the indicator function for some event A. (There is a picture that shows $A = [a, b]$ for some $0 < a < b < 1$.) So, what sets belong in $\sigma(X)$? We can see that $\emptyset$ is there (by taking an appropriate pre-image of a set not containing 0 or 1). Similarly, Ω is there, A is there, and A^c is there. Thus,

$$\sigma(X) = \{\emptyset, \Omega, A, A^c\}$$

Look at how small our σ-algebra is! This is the smallest possible. It's minimal.

[18] Is it clear why? Our definition just lists the sets that need to be measurable!

2. A little bit more complicated example. Suppose that X is a simple random variable (that is, it is a finite combination of indicator functions for events that partition Ω.) That is,

$$X = \sum_{k=1}^{n} a_k I\{A_n\}$$

Let's say that there are $n = 4$ sets in our example. If the a_k are all different[19], then we see that

$$\sigma(X) = \sigma(A_1, \ldots, A_n).$$

3. How about X = delay of the flight (flight is delayed by at most 1 hour, uniformly). Then $\sigma(X) = \mathbb{R}$, all Borel sets in $[0, 1]$.

So this is how random variables generate σ-algebras. Then, we can sort of forget the ambient σ-algebra and outcome space Ω. The next thing is to forget probabilities.

5.4 Random variables induce probability measures on $\mathbb{R}$

Let X be a random variable. For every Borel set $A \subseteq R$, we define the probability $P(A)$ of A by

$$P(A) := \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$$

The first P is different from the second $\mathbb{P}$. Then P is a probability measure on $\mathbb{R}$ ("induced"). Then, P is called the *distribution* of X. If you know the probability of the value of every X, then you know the distribution.

So, we can forget about $\mathcal{F}$, Ω, and $\mathbb{P}$ as well.

Sometime ago, I had this collaborator. I asked him why we distinguished probability and measure theory. In measure theory, we study measure spaces $(\Omega, \mathcal{F}, \mu)$. In probability theory, we try to quickly[20] forget about measure spaces and discuss distributions.

OCTOBER 22, 2007

What defines a random variable? We start with the introduction of the distribution function.

6 Distribution Function

What is the domain where the random variable is defined is not as important as the values. The specifics of the nature of the sample space Ω is not as important

[19] This is a nice idea. It says that if you have a random variable, you don't need the whole real line $\Omega = \mathbb{R}$. You just need the appropriate "chunks" of the real line.

[20] I guess by the ninth lecture

as the likelihood of the values on $\omega \in \Omega$. It may be that Ω's definition is an artifact of something, or we may not be able to have a handle on it.

So, we already know how to get rid of the (original) σ-algebra. Basically, we take the preimages (in our random variable X) of all Borel sets.

What we will do now will all be done by the induced distribution on $\mathbb{R}$. Let's repeat some of the last lecture. If X is a random variable, then X induces a probability measure P on $\mathbb{R}$ defined by taking the preimage for every Borel set A:

$$P(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A). \tag{6}$$

Then, the problem (it's on the homework as problem 7) to prove that it is indeed a probability measure on $\mathbb{R}$. That is, prove that $(\mathbb{R}, \mathcal{R}, P)$ is a probability space.

So here in (6), there is no reference to Ω whatsoever. We're defining a new probability on the real line.

This probability measure P is called *the distribution of X*. In other words (being a little bit more informal), the distribution of X is the list of all values: it lists the probabilities of each Borel (output) set:

$$\text{distribution of } X = \big\{\mathbb{P}(X \in A), \quad A \text{ Borel in } \mathbb{R}\big\}.$$

The distribution is good to know, but it does not define the random variable uniquely. If we have two random variables, they may be pretty different yet have the same distribution. Again, different random variables (even very different random variables) may have the same distribution.

Let's look at some examples:

a). You could have $X(H) = 1, X(T) = 0$, say, for the coin toss. And, you have $Y(H) = 0, Y(T) = 1$. These are very different, but X and Y have the same distribution.

b). You have two coin tosses. Say $X = 1$ if H appears in first toss, and $X = 0$ otherwise. $Y = 1$ if H appears in the second toss, and $Y = 0$ otherwise. Now these variables are not opposite anymore. But the experiments are independent. But X and Y have the same distribution, even though say are so different.

So, to straighten out this gap, in probability theory, we introduce two notions.

6.1 Two notions of Equivalence

The stronger notion will be "almost sure equality", and the other will be "equality in distributions", such as the above examples.

Definition 6.1. • *Random variables X and Y are equal almost surely, and we will write*

$$X \overset{a.s.}{=} Y,$$

if $\mathbb{P}(X \neq Y) = 0$.[21]

[21] In analysis, we say "*almost everywhere*," and write *a.e.*.

- *Random variables X and Y are equal in distribution, and we write*

$$X \overset{d.}{=} Y,$$

if they have the same distribution. Equivalently,

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A) \text{ for all } A \text{ Borel.}$$

If $X \overset{\text{a.s.}}{=} Y$, then $X \overset{\text{d.}}{=} Y$, but not vice versa (as these examples show).

Now we'll work with the notion of distribution for a while. In the second notion, we have nothing else to work with. We just identify them. So, we'll work with this notion $X \overset{\text{d.}}{=} Y$. What is the distribution? It's the list of all probabilities for Borel sets. But as usual, this is too big. The distribution will be uniquely-determined by generators. If you know the probabilities for generators $(-\infty, a]$, then you'll know everything.

The distribution of X is deteremined by its values of the form $\mathbb{P}(X \in A), A = (-\infty, a], a \in \mathbb{R}$, because such intervals form a generating collection for $\mathcal{R}$, and a measure is uniquely determined by its values on a collection of generators (Uniqueness Theorem of Measure Theory[22]).

So, $\mathbb{P}(X \in A)$ is just the probability $\mathbb{P}(X \leq a), a \in \mathbb{R}$, and probabilities of this form $\mathbb{P}(X \leq a), a \in \mathbb{R}$ determine the distribution of X. Now, these are parametrized by a single number a. So now, we can slide our X along the real line $\mathbb{R}$. So now we can define an increasing function from $\mathbb{R}$ to $\mathbb{R}$. This function is called the *distribution function*. This is an important notion which we have just discovered.

Definition 6.2. *The distribution function $F : \mathbb{R} \to [0, 1]$ of a random variable X is defined by the rule*

$$F(x) := \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}.$$

It determines uniquely the distribution of X.

This is a powerful extraction of the experiment. We don't talk about Ω and so on. Everything is summarized be the graph of an $\mathbb{R} \to \mathbb{R}$ function.

Example 6.3. *Let $X = \#$ of H in two independent tosses of a fair coin. So, we have*

$$X = \begin{cases} 0 & , \ prob \ = \frac{1}{4} \\ 1 & , \ prob \ = \frac{1}{2} \\ 2 & , \ prob \ = \frac{1}{4} \end{cases}$$

So then, we can convert this to a function. There will be jumps on this function (at 0, 1, and 2).

$$F(X) = \mathbb{P}(X \leq x) = \begin{cases} 0 & , \quad x < 0 \\ \frac{1}{4} & , \quad x \in [0, 1) \\ \frac{1}{4} + \frac{1}{2} & , \quad x \in [1, 2) \\ \frac{1}{4} + \frac{1}{2} + \frac{1}{4} & , \quad x \in [2, +\infty) \end{cases}$$

[22] We are not proving this: We're believing this.

The distribution function of any simple random variable will look like this, that is, it will be piecewise-constant (with finitely-many jumps, of course).

Let's look at a second example: the poor flight that's always delayed from the first lecture.

Example 6.4. *X = delay of the flight. Recall that the delay $\in [0,1]$ is uniform. Then,*

$$F(x) = \mathbb{P}(X \le x) = \begin{cases} 0, & x < 0 \\ x, & x \in [0,1] \\ 1, & x \ge 1. \end{cases}$$

In this example, the function F is continuous everywhere.

These two examples are very typical. They are on the opposite ends of the spectrum of what happens. Every distribution function has a certain look:

- Its value is 0 at $-\infty$.
- Its value is 1 at $+\infty$.
- It's an increasing (rather, non-decreasing) function in between. It can have jumps in between.

These two examples are representative: they suggest a general form for arbitrary distribution functions. We'll put those properties up as a theorem. Most of these are pretty intuitive from the examples.

Theorem 6.5. *(Properties of a distribution function) The distribution function $F(x)$ of a random variable X has the following properties:*

(i) *F is nondecreasing and $0 \le F(x) \le 1$.*

(ii) *$F(x) \to 0$ as $x \to -\infty$, and $F(x) \to 1$ as $x \to +\infty$.*

(iii) *F is right-continuous, i.e.*

$$\lim_{y \to x_+} F(y) = F(x).$$

(iv) *F has left-hand limits at all points*[23]*. Moreover,*

$$F(x_-) := \lim_{y \to x_-} F(y) = \mathbb{P}(X < x).$$

In particular,

$$\mathbb{P}(X = x) = F(x) - F(x_-).$$

(v) *F has at most countable number of discontinuities.*

[23] And these jumps on the distribution function will mean something. What are these jumps? These jumps mean the probability that X takes this specific value.

All of this is pretty intuitive, except for maybe the last part. Let's prove these one-by-one.

Proof. (i) trivial.

(ii) This is clear, but not trivial. This $F(x)$ is the probability that $X \leq x$. If we can prove that these events have limit the empty set, then we are done. So, let's write this down. The events $\{X \leq x_n\} \searrow \emptyset$ as $x_n \to -\infty$. By the continuity, $\mathbb{P}(X \leq x_n) \to \mathbb{P}(\emptyset) = 0$.

Similarly, $\{X \leq x_n\} \nearrow \Omega$ as $x_n \to +\infty$. So, $\mathbb{P}(X \leq x_n) \to \mathbb{P}(\Omega) = 1$. This completes the proof of *(ii)*.

□

OCTOBER 24, 2007

If X is a random variable, the distribution function $F(x)$ is defined as

$$F(x) = \mathbb{P}(X \leq x),$$

and we know that the distribution uniquely defines your probability. Last time, we listed some of the properties and started to prove them. They were:

(i) F is non-decreasing, $0 \leq F(x) \leq 1$.

(ii) $F(x) \to 0$ as $x \to -\infty$, and $F(x) \to 1$ as $x \to \infty$.

(iii) F is right-continuous: As we approach x from the right, we obtain the limiting value:

$$\lim_{y \to x_+} F(y) = F(x).$$

Proof. Suppase we have a sequence converging to x from the right. We want to show that this arbitrary sequence $y_n = F(x_n)$ converges to x (even $y_n \searrow x$). We have

$$\mathbb{P}(X \leq y_n) \to \mathbb{P}(X \leq x), \qquad n \to \infty.$$

The conclusion follows from the convergence of the sets

$$\{X \leq y_n\} \searrow \{X \leq x\}$$

and the continuity. □

(iv) F has left-hand limits[24] at all points, and

$$F(x_-) := \lim_{y \to x_-} F(y) = \mathbb{P}(X < x).$$

[24] Here, we address what happens from the left. There is no continuity, but there is still a limit.

In particular, we can give meaning to the jumps[25]:

$$\mathbb{P}(X = x) = F(x) - F(x_-).$$

Proof. Similar to (ii), we just take $y_n \nearrow x$, and we want to show

$$\mathbb{P}(X \leq y_n) \to \mathbb{P}(X < x), \qquad n \to \infty.$$

This follows from the convergence of the sets

$$\{X \leq y_n\} \nearrow \{X < x\}$$

and the set on the right is correct (as a strict inequality) when you think about what is the union. This follows from the continuity of probability. □

(v) F has at most countable number of discontinuities.

Proof. The proof doesn't use much anything, other than continuity and monotonicity of the function F. First, we look at the jumps of at least $\frac{1}{2}$. There are at most two of these kinds of jumps. Similiarly, there are at most three jumps of size $\frac{1}{3}$. This continues. Let's write this down.

Since $0 \leq F(x) \leq 1$,

- there can be at most 2 jumps (i.e. discontinuities) of height $\geq \frac{1}{2}$. Here, "jumps" means points x where $F(x) - F(x_-) \geq \frac{1}{2}$.
- there can be at most 3 jumps of size $\in [\frac{1}{3}, \frac{1}{2})$.
- there can be at most 4 jumps of size $\in [\frac{1}{4}, \frac{1}{3})$.
- $\cdots$

The total number of jumps is countable, and every jump is in $[\frac{1}{n}, \frac{1}{n-1})$ for some n, so this list contains all possible jumps. □

Why is this useful? Let's make a corollary:

Corollary 6.6. $\mathbb{P}(X \in (a, b]) = F(b) - F(a)$.

So, we can compute the probability of every interval, just by looking at the distribution function. This intuitively shows that you can compute the probability of any event, because each Borel set is the combining of these kinds of intervals. So the distribution function becomes even more important because we can make the procedure rigorous.

We can start with a function satisfying properties and get a random variable:

Theorem 6.7. *If a function F satisfies (i), (ii), and (iii), then F is the distribution function of some random variable X.*

[25] Jumps mean, so to speak the atoms. They are exactly the places where the probability becomes discrete.

The proof is constructive. You can actually write out the random variable X for which F is a distribution function.

Proof. Consider the probability space[26] $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = [0,1]$, $\mathcal{F}$ is the collection of Borel sets, and $\mathbb{P}$ is the Lesbesgue measure.

Define the values of the random variable X as follows:

Well, how would we do this if F were continuous and strictly monotone? We think of the interval $\Omega = [0,1]$ as the codomain of F. Then we define $X(\omega) = F^{-1}(\omega)$, if F is one-to-one.

Why is this correct? If you ask the question that $X \leq a$ (that is, what is the probability of $(-\infty, a]$, then we'll have exactly $[0, \omega = F(a)]$.

You can rectify the situation in general by assigning

$$X(\omega) = \sup\{y : F(y) < \omega\}$$

The exercise (in the homework) is to complete this proof. □

This is a common way do generate random variables on a computer. You start with a distribution function F, and then you can compute a random variable X. First, you generate the probability (uniform) space $\Omega = [0,1]$. Then you just apply the "inverse" of your distribution function.

6.2 Types of Distributions

There are three types of distributions:

1. Discrete
2. Continuous
3. Something wild (that's not just a combination of the previous two).

6.2.1 Discrete Distributions

Let X be a random variable with the distribution function[27] F (with the induced probability $\mathbb{P}$). The distribution of X (and X itself) is *discrete* if there exists a countable set S such that

$$\mathbb{P}(S^c) = 0.$$

So, the random variable can only take on a countable number of values with certain probabilities. So, in other words, X takes a countable number of values (S).

We can enumerate, by writing $S = \{x_1, x_2, \ldots\}$, and we can look at $p_k := \mathbb{P}(X = x_k)$. Our random variable will take these values S with certain probabilities, and that's it. Then, we have

$$F(x) = \sum_{k:\ x_k \leq x} p_k,$$

[26] We'll make this very complete

[27] We'll study a random variable, from now on, through its distribution function F.

and this is a sum over a countable set. These p_k are sometimes called *point masses*.
Here are some examples:

1. One point mass at zero:
$$\mathbb{P}(X = 0) = 1.$$
So, the distribution function F is:
$$F(x) = \begin{cases} 0 & , x < 0 \\ 1 & , x \geq 0 \end{cases}$$
This is sometimes called *Dirac measure (at 0)*.

2. Any simple random variable is discrete. A simple random variable is a random variable that takes finitely-many values. For example, the coin toss example of the last lecture (where we have the four values) is a discrete random variable.

3. Finally, there are "wilder" examples than this. For example, where the set S of values is are all rationals: $S = \mathbb{Q}$, where you have arbitrary point masses summing up to 1, e.g. $p_k = \frac{1}{2^k}$. Then, the function F is discontinuous at every(?) point. (It is continuous at every irrational point and discontinuous at rational points.)

The discrete random variables are the simplest examples of random variables.

6.2.2 Absolutely continuous distributions

This is on the complete other side of the scale. So, the distribution of X (and X itself, sometimes people will say) is *absolutely continuous* if there exists f, called the *density of X*, such that
$$F(x) = \int_{-\infty}^{x} f(y)\, dy \qquad \text{for every } x.$$
So clearly, $f \geq 0$ is necessary, and the total mass of the density is one, that is:
$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

- In particular, if f is continuous, then
$$F'(x) = f(x),$$
that is, it is the derivative of a distribution function.

- Then, we can say
$$\mathbb{P}(X \in (a, b]) = F(b) - F(a) = \int_a^b f(x)\, dx.$$
The meaning of the density is that it indicates the likelihood of X to be "near" x.

- Alternatively, we can start with a function f such that

$$f \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} f(x)\ dx = 1,$$

and define the distribution function by the formula

$$F(x) := \int_{-\infty}^{x} f(y)\ dy.$$

(This is a distribution function indeed.)

Question: Little "f" (f) is the density?

- **Answer:** Yes, once the conditions listed are satisfied.

The basic example of the absolutely continuous random variable is the example about the flight. This is the *uniform distribution on* $[0, 1]$. Its density is a constant:

$$f(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} F(x) &= \int_{-\infty}^{\infty} f(y)\ dy. \\ &= \begin{cases} x & , \text{ if } x \in [0,1] \\ 0 & , \text{ if } x \leq 0 \\ 1 & , \text{ if } x \geq 1. \end{cases} \end{aligned}$$

The example of uniform distribution is the delayed flight.

OCTOBER 26, 2007

There are three types of random variables (and distributions): Discrete, absolutely continuous (where the random variable has a density), and something that's wild, which is an exceptional case. So, let's finish the absolutely continuous distributions. In this case,

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)\ dy,$$

and we can compute the probability as

$$P(X \in (a, b]) = \int_{a}^{b} f(x)\ dx.$$

So, we see that f will be larger to show a concentration of probability for appropriate points. So, our example is

Example 6.8. *Standard normal distribution on $\mathbb{R}$. This is more convenient to define using the density.*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

This is the familiar "bell-shaped curve." There is no closed-form expression for the distribution function $F(x)$, other than saying that it is the value of the integral. However, there are good estimates on $F(x)$.

So, what is $F(x)$? You take an $x \in \mathbb{R}$, and you have

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy$$

The biggest concentration for x will be at x. So, the complement to the density $1 - F(x)$ should be roughly proportional to $e^{-x^2/2}$ for large x.

So, then we can prove the following. The bound from above is used much more than the bound from below:

Proposition 6.9. *For all x,*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) e^{-x^2/2} \leq 1 - F(x) \leq \frac{1}{x} e^{-x^2/2}$$

This inequality is way sharper (on the left) than I ever use.

6.2.3 Cantor Distribution, and other wild distributions

So here will be an example that doesn't fall into the category of being mixtures of the previous two types. These are usually considered on fractals.

So, let's remember what is the Cantor set. Consider $[0, 1]$. Remove the interval $[\frac{1}{3}, \frac{2}{3}]$. Then, from what remains, remove the middle set(s) again (on the intervals that remain). You iterate this. What remains is called the *Cantor set*. Call it C. It has lots of wild properties:

- The (Lesbesgue) measure of C is zero (i.e. C is a null set) and C is uncountable[28].

Now, that we've defined C, we can define the *Cantor distribution*, defined using the distribution function. $F(x)$ will be defined on $[0, 1]$.

The values are given: $F(0) = 0$, $F(1) = 1$, and in the middle, F is $= frac12$. That is, $F(x) = \frac{1}{2}$ on $x \in [\frac{1}{3}, \frac{2}{3}]$. Then, we iterate. So $F(x) = \frac{1}{4}$ on the interval $[\frac{1}{4}, \frac{2}{9}]$ and $F(x) = \frac{3}{4}$ on $[\frac{7}{9}, \frac{8}{9}]$. And so on. You get sort of a "stair" shape in the graph. Then, extend F on the interval $[0, 1]$ continuously.

F is a self-similar function, and it is continuous. F is extended on to $\mathbb{R}$ be setting $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$. Now, some properties of F:

[28] So, we know that a countable set has measure zero, but this shows that the other way does not hold. It's easy to prove that C is measure zero. Wikipedia contains a pretty nice proof that C is uncountable!

1. F is continuous[29]

2. F is constant on "middle thirds," so the derivative is constant there. But, these are the intervals we excluded. So F is constant (rather $F' = 0$) a. e. (on C^c).

Therefore, F is not discrete (by 1, and the intermediate value theorem[30]). Further, F is not absolutely continuous (by 2).

Hopefully, in probability, we rarely come across such examples. In practice, our distributions will be absolutely continuous or discrete.

So, this completes our discussion of types of distributions and examples.

7 Integration

Why do we need to talk about integration? It's because of the unfortunate fact that the expectation of a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ needs to be defined by an integral.

Heuristically, the expectation is the average of the values of X on Ω. The average is usually defined as some integral over Ω. In calculus 1[31], you defined an average value, and it was an integral over some interval. So,

$$\mathbb{E}X = \frac{1}{\mathbb{P}(\Omega)} \int_\Omega X(w)\, d\mathbb{P}(w).$$

The problem, is that we do not have a definition of integral over arbitrary spaces Ω. What is the integral $\int_\Omega$ over Ω for an abstract sample space Ω.

So, we'll quickly review the type of integration we already know, the Riemann integral. This does not solve our problem, because it's for $\mathbb{R}$, but we'll quickly jump into other integrals (Lebesgue, etc.)

7.1 Riemann Integral

It is defined for a fnuction $f : [a, b] \to \mathbb{R}$. Partition the domain $[a, b]$ into n consecutive intervals $\Delta_1, \ldots, \Delta_n$. Then,

$$\Delta_k = (x_{k-1}, x_k].$$

We consider the Riemann sums

$$R(\Delta) = \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}), \tag{7}$$

where $t_k \in \Delta_k$.

[29] CHECK

[30] Because it does not take all intermediate values.

[31] or maybe 2

Then, this approximates the Riemann integral well, as you refine your partition. How fine your partition is is quantified by the *mesh*, which is defined to be

$$\|\Delta\| = \max_k |\Delta_k|.$$

Definition 7.1. *The Riemann integral of f exists if there exists A such that*

$$\lim_{\|\Delta\|\to 0} |R(\Delta) - A| = 0$$

for arbitrary portitions Δ and t_k. This limit A is called the value of the integral, and we write

$$A = \int_a^b f(x)\ dx.$$

Theorem 7.2. *If f is bounded and continuous a. e , then f is Riemann integrable.*

You've probably seen a stronger version of this theorem.

How many of you have heard about the Riemann-Stieltjes integral? It's a generalization of the Riemann integral, and it's useful in probability theory. Instead of defining Riemann sums as in (7) we assume that your partition is an equi-partition. Then, we sum by the values of the function, multiplied by the same number. But sometimes, it's nice to introduce some weights (and debias the function). So, instead of measuring the values uniformly, you weight the x values.

Let $F : [a, b] \to \mathbb{R}$ be a monotone function. Define the *Riemann-Stieltjes* sums as

$$RS(\Delta) = \sum_{k=1}^{n} f(t_k)(F(x_k) - F(x_{k-1})$$

The result (for an analogous definition of integral) is called the *Riemann-Stieltjes integral of f* and is written

$$\int_a^b f(x)\ dF(x).$$

This notation is pretty justified, because if F is continuously differentiable, then $dF(x)$ will indeed equal $F'(x)\ dx$:

$$\int_a^b f(x)\ dF(x) = \int_a^b f(x) \cdot F'(x)\ dx.$$

Thus, with $F = 1$ constantly, this is the usual Riemann integral. For nice functions f, adding this F doesn't do anything. These increments (differences) will pick up the jumps.

So, this is basically the Riemann integral.

7.2 Lebesgue integral

So, let me give you a heuristic idea. Once you know Lebesgue integral, you do not need to know any other integral. Whenever you write this sign

$$\int,$$

you can just think about the Lebesgue integral. It's the most general thing you need.

The Lebesgue integral is defined on an arbitrary measure space $(\Omega, \mathcal{F}, \mu)$. For simplicity, we'll assume that μ is a *finite measure*, i.e. $\mu(\Omega) < \infty$. This assumption is not actually needed: you can get rid of it. It's not a problem in probability measures.

We can not just partition the domain (because it is abstract). So, we avoid partitioning Ω. Instead, we partition $\mathbb{R}$, the range. A function $f : \Omega \to \mathbb{R}$ is much easier to partition in the range. The result is called the *Lebesgue integral*. We have to know how to partition the range, but this is pretty much the same thing. It will be harder to prove properties about the Lebesgue integral. So, this is possible, but we will not define the Lebesgue integral this way.

Instead, we define Lebesgue integral by steps (four or so).

1. First, for indicator functions (where it's simple)[32]
2. Then, we extend it for simple functions. These are linear combinations of indicator functions.[33]
3. Third, for general functions (which, by the homework, every function can be approximated), as the limit of integrals of simple functions.

OCTOBER 29, 2007

Let f be a measurable function on a measure space $(\Omega, \mathcal{F}, \mu)$ with finite measure μ (i.e. $\mu(\Omega) < \infty$). We have a function $f : \Omega \to \mathbb{R}$. Then, what is

$$\int f \, d\mu = ?$$

So, we'll do this in steps. Step zero is to define it on indicator functions. So, if $f = \mathbf{1}_A$, define

$$\int f \, d\mu = \mu(A).$$

Here are the remaining steps:

[32] What should be the integral of such a set? Its measure.
[33] We extend, thus, by linearity.

1. For simple functions
$$f = \sum_{k=1}^{n} a_k \mathbf{1}_{A_k}$$
for some decomposition $\Omega = \bigcup_{k=1}^{n} A_k$, define
$$\int f \, d\mu = \sum_{k=1}^{n} a_k \mu(A_k).$$

2. For bounded functions $|f| \leq M$. (See Problem 2 of HW4). Here,w
   (a) There is some simple functon $f_n \to f$ pointwise.
   (b) Define
   $$\int f \, d\mu = \sup_{\phi \leq f} \int \phi \, d\mu = \inf_{\psi \geq f} \int \psi \, d\mu.$$
   where ϕ, ψ are simple functions[34].

3. Then, for non-negative functions $f \geq 0$. Define
$$\int f \, d\mu = \sup_{0 \leq \phi \leq f} \int \phi \, d\mu = \lim_{n \to \infty} \int \min(f, n) \, d\mu.$$
   where ϕ is bounded. The second equality[35] is an alternative definition.
   If this integral is finite, we'll say that the non-negative function f is *integrable*.

4. For general functions f, this is straightforward given step 3. We just look at the positive and negative parts separately. So, we decompose f into its positive part f^+ and negative part f^-
$$f = f^+ - f^-,$$
   where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$.

Question: Doesn't this make a function like $\sin x$ have integral unbounded?

- **Answer:** Yes, even more quickly so. But, this integral does not exist, even in the Riemannian sense.

So, let's make that point clear with a definition.

[34] We really want to use a limit of bounded variables, because then we can later say "truncate."

[35] This is a point-wise minimum of f and the constant function with output value n.

Definition 7.3. *1. We call f integrable if $|f|$ is integrable. Equivalently, if both f^+ and f^- are integrable, and in this case,*
footnoteThat is the the integral. Sometimes, we'll denote this

$$\int_\Omega f \, d\mu$$

to indicate the abstract space Ω., define

$$\int f \, d\mu = \int f^+ \, d\mu - \int f^- \, d\mu.$$

2. The integral on a measurable set $A \subseteq \Omega$ is

$$\int_A f \, d\mu = \int_\Omega f \cdot \mathbf{1}_A \, d\mu.$$

Then, with this definition, the properties of the Lebesgue integral are easy to prove.

7.2.1 Properties of Lebesgue Integral

Proposition 7.4. *1. (Monotonicity). If $f \leq g$ almost everywhere, then $\int f \, d\mu \leq \int g \, d\mu$.*
If $f = g$ almost everywhere, then the integrals are equal[36].

2. (Linearity) $\int (af + bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu$, where $a, b \in \mathbb{R}$.

Proof. We'll skip the proof, but it's not difficult to prove this (as an exercise) for simple functions. □

Corollary 7.5. $\left|\int f \, d\mu\right| leq \int |f| \, d\mu$.

Why is this true? We have to check two inequalities.

Proof. Since $f \leq |f|$, this implies $\int f \leq \int |f^+| \, d\mu$. Similarly, $-f \leq |f|$, so $-\int f \, d\mu \leq \int |f^-| \, d\mu$. □

This does not happen for Riemann integrals, right? Highly-oscillatory functions are still non-integrable. So, with that, let's compare integrals.

7.2.2 Comapring Lebesgue and Riemann integrals

So, what's the punchline? Lebesgue integral is better. Whenever possible, you should use this.

Theorem 7.6. *If f is Riemann integrable, then f is Lebesgue integrable, and the two integrals agree.*

[36] Bascially, the integral over any set of mesasure zero will be zero.

This is true at least on finite domains.

So, the Lebesgue integral is at least as good as the Riemann integral. However, the converse does not hold. For example,

Example 7.7. *$f = \mathbf{1}_{\mathbb{Q}}$. This is a simple function*[37]*. Since $\mathbb{Q}$ is countable, $\mu(\mathbb{Q}) = 0$. Therefore $\int f \, d\mu = 0$. So, f is Lebesgue integrable.*

But, f is <u>not</u> Riemann integrable. We'll leave this as an exercise. It shouldn't be hard though. No matter where you look, independent of how fine your mesh is, an upper sum will use 1s and a lower sum will use 0s.

One remark is 'How to view step 2' above. After the homework, you see that you take simple function approximations by decomposing functions by range.

So you can rethink of step 2 as partitioning by range, instead of by domain.

So, the next step is to prove properties that hold in the limit.

7.2.3 Integration to the Limit

So, the main lemma that we'll look at is as follows (Fatou's Lemma).

Lemma 7.8 (Fatou's Lemma)**.** *If $f_n \geq 0$, then $\int \liminf f_n \, d\mu \leq \liminf \int f_n \, d\mu$.*

As an exercise, you should show that the converse does not hold. Actually, proving the converse is an easy way to remember Fatou's lemma. Let's try to sketch the proof of this, at least for indicator functions.

Proof. The proof is by steps: for indicator functions, then simple functions, and so on.

So for indicator functions. Let $f_n = \mathbf{1}_{A_n}$. Then,

$$\int \liminf \mathbf{1}_{A_k} \, d\mu = \int \mathbf{1}_{\liminf A_n} \, d\mu.$$

This was used in one solution of # 7 of HW 1. But then, we know that this is equal to

$$= \mu(\liminf A_n).$$

Our goal is to "pull the lim inf out, from the measure." And we have

$$\mu(\liminf A_n) \leq \liminf \mu(A_n),$$

proved in class (and Theorem 3.2(i) in the text).

Then, we recall the definition. This is equal to

$$= \liminf \int \mathbf{1}_{A_n} \, d\mu.$$

So, then it's an exercise to show it's true for simple functions. Use linearity. □

[37] It was asked if this is simple. Indeed, we can use A_1 is the rationals and A_2 the irrationals.

From this, you can deduce certain properties. Consider the following problem:

Suppose $f_n \to f$. Is it true that

$$\int f_n \, d\mu \to \int f \, d\mu?$$

It's not true in general (exercise).

Theorem 7.9 (Monotone Convergence Theorem)**.** *If $f_n \geq 0$ and $f_n \nearrow f$, then $\int f_n \, d\mu \to \int f \, d\mu$.*

A similar statement is true for $\searrow$. This is almost immediate from Fatou's lemma.

Proof. By Fatou's Lemma,

$$\int f \, d\mu \leq \liminf \int f_n \, d\mu$$

On the other side, each single f_n,

$$\limsup \int f_n \, d\mu \leq \int f \, d\mu$$

by the monotonicity and the property that each f_n is smaller than f ($f_n \leq f$). This completes the proof. □

So, another time when a statement of this kind is true is when every function is bounded by another function. This is the Dominated Convergence Theorem.

Definition 7.10. *$f_n \to f$ almost everywhere (a. e.) if $\mu\{\omega : f_n(\omega) \not\to f(\omega)\} = 0$.*

Theorem 7.11 (Dominated Convergence Theorem)**.** *If $f_n \to f$ a. e. and $|f_n| \leq g$ for all n, where g is integrable, then*

$$\int f_n \, d\mu \to \int f \, d\mu.$$

Proof. We apply a "shift" so that we can apply Fatou's lemma. $f_n + g \geq 0$ for all n. By Fatou's Lemma,

$$\int (f + g) \, d\mu \leq \liminf \int (f_n + g) \, d\mu.$$

Thus, $\int f \, d\mu \leq \liminf \int f_n \, d\mu$. Then, repeat the argument for $-f_n$. This completes the proof. □

OCTOBER 31, 2007

The Midterm will be posted today online, and due next Wednesday. No late submissions. No collaboration. It is weighted higher. Unlike the HWs, I will not be able to consult on it to give hints. This is similar to an actual in-class exam. My office hours will be on Friday as usual, and I have to cancel office hours on Monday. So, be sure the statements are clear by then. The TA still has office hours on Tuesday. So, use any text that you want, but not a human. The TA is not grading every single problem in the HW. I will post which problems have been graded on each HW, so you will know what you can feel confident about. Any questions?

Now, we're finished with integration, and we can define expectation of random variables.

8 Expectation

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. So $X : \Omega \to \mathbb{R}$ is a measurable function.

Definition 8.1. *The expectation (mean) of X is*

$$\mathbb{E}X = \int X \, d\mathbb{P} = \int_\Omega X(\omega) \, d\mathbb{P}(\omega),$$

the usual Lebesgue integral.

So, you should think about expectation as an average value. But, this is "weighted" according to the probabilities. There are many properties of the expectation. The properties of the expectation follow from the properties of the Lebesgue integral:

For example, the expectation is linear: $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$, and the expectation of a constant random variable is a constant: $\mathbb{E}(\text{const}) = \text{const}$. Also, if $X \geq Y$ almost everywhere, then this implies $\mathbb{E}X \geq \mathbb{E}Y$. That's the monotonicity of the integral. You can rethink other properties of the integral as problems, to see what properties you'd obtain.

For simple random variables, the expectation reduces to a simple sum. What is this? Let

$$X = x_k \text{ with probability } p_k, \qquad k = 1, \ldots, n.$$

(Recall, X is a linear combination of indicators.) Then, the definition of Lebesgue integral for simple functions gives

$$\mathbb{E}X = \sum_{k=1}^{n} x_k p_k,$$

or more generally, we have

$$\sum_{k=1}^{n} x_k \mathbb{P}(x = x_k).$$

So, a special case of the integral is the sum. As an exercise,

Exercise. Show that for discrete random variables,

$$\mathbb{E}X = \sum_{k=1}^{\infty} x_k \mathbb{P}(x = x_k).$$

Question: Does the series converge?

- **Answer:** The series converges if it has expectation. For positive random variables, you either have the expectation equal to ∞, or it's finite. But it can not oscillate. Maybe we'll modify the exercise to:

Exercise. Show that for discrete random variables,

$$\mathbb{E}X = \sum_{k=1}^{\infty} x_k \mathbb{P}(x = x_k), \tag{8}$$

if the series is absolutely convergent.

As an example, let's consider:

Example 8.2. *Let X be the number of heads in 2 tosses of a fair coin. Recall*

$$X = \begin{cases} 2, & \textit{probability } \frac{1}{4} \\ 2, & \textit{probability } \frac{1}{2} \\ 0, & \textit{probability } \frac{1}{4} \end{cases}$$

We had this somwhere. So the expectation is to sum up these values with these weights:

$$\mathbb{E}X = 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{4} = 1.$$

This is what you should expect. In two tosses, you should get one head.

This example is great, because it shows: You only need values and the probabilities of these values, but you do not need to know which points go to which values. So for example, the random variable for the tails will have the same distribution. It will be the compeletly "opposite" random variables, but the expectation will be the same. It doesn't matter what the values of X signify. So, the somewhat-nontrivial theorem is:

Theorem 8.3. *Distribution determines expectation. That is, $X \stackrel{d.}{=} Y$ implies $\mathbb{E}X = \mathbb{E}Y$.*

This is good for us. The distribution does not deal with the sample space. It's good to know that the expectation can be proved without knowing the sample space. We'll see how this is useful for us later. The proof proceeds in steps, much like some of the other recent proofs.

Proof. First, we prove this (Step 1) for simple random variables X and Y. In (8), there is no mention of the sample space. So, this follows from the form of the expectation of X and the expectation of Y, as seen in (8).

Step 2, for bounded random variables: Let $X \geq 0, Y \geq 0$, with $X \leq R, Y \leq R$. By Problem 2 of HW4, there exist $X_n \geq 0, Y_n \geq 0$ simple random variable that converge $X_n \nearrow X$, $Y_n \nearrow Y$ point-wise. (Remember how to do this? So, we partition the range into equal intervals.) Then $X_n \overset{\text{d.}}{=} Y_n$. So, by Step 1, $\mathbb{E}X_n = \mathbb{E}Y_n$, the expectations agree. It's only left to say that they actually approximate the $\mathbb{E}X$. So, why does $\mathbb{E}X_n \to \mathbb{E}X$? They converge point-wise, right? You have to use some limit theorem. It's not always true that point-wise convergence implies convergence of the integral. We can use dominated convergence theorem (or monotone convergence theorem). Let's say "CHECK", ... it's a nice exercise. By the Monotone Convergence THeorem,

$$\mathbb{E}X_n \to \mathbb{E}X, \qquad \mathbb{E}Y_n \to \mathbb{E}Y,$$

because $X_n \geq 0, X_n \nearrow X$.

Step 3 is for non-negative (but not necessarily bounded) random variables $X \geq 0, Y \geq 0$. But this is easy. We can say

$$\mathbb{E}X = \sup\{\mathbb{E}X' : X' \leq X \text{ and } X' \text{ is bounded}\},$$

or we can write some truncation instead of X'. (The supremum exists by hypothesis.) So, we can use step 2, and similarly for Y. By Step 2, these suprema are equal.

In step 4, we have general X and Y, and we decompose

$$X = X^+ - X^-, \qquad Y = Y^+ - Y^-.$$

We use a little bit of thinking to show that

$$X \overset{\text{d.}}{=} Y \text{ implies } X^+ \overset{\text{d.}}{=} Y^+ \text{ and } X^- \overset{\text{d.}}{=} Y^-.$$

Check this property: It makes sense that X and Y which are equal up to distribution share positive parts and negative parts (up to a null set). Then, we use Step 3 on the positive and negative parts, respectively. □

But this is not good enough. What would be good enough is if we could discuss expectation with an integral, NOT over Ω, but over $\mathbb{R}$. So, how do we do this? We think about induced $\mathbb{R}$ probability spaces. This change of variables changes our integral to an integral over $\mathbb{R}$.

8.1 Change of Variables

We want to compute expectation by integrating over $\mathbb{R}$, and not Ω. To do this, we use the induced probability measure P on $\mathbb{R}$. The measure is defined as follows:

$$P(A) = \mathbb{P}(X \in A), \qquad A \subset \mathbb{R} \text{ Borel}$$

defined on every Borel set A. What this means, in particular, is that we can cook up another random variable which will be a map (not from Ω to $\mathbb{R}$, but) from $\mathbb{R}$ to $\mathbb{R}$. In particular, the "identical" random variable assigning $x \mapsto x$ on $\mathbb{R}$ equals X is distribution. It is clear that this is a random variable: the identity is measurable. But this random variable is equal to the random variable X (in distribution).

Therefore, by Theorem 8.3,

$$\mathbb{E}X = \int_{\mathbb{R}} x \; dP.$$

And more generally, the expectation of a function $g(X)$ is

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x) \; dP.$$

Check this: it's the same concept. The random variable is defined as you think it should be.

Now things will become easier in a very great speed. Because, the dP is the rate in which our measure changes. And the rate in which our measure changes is the speed of these "jumps" in our probability. So, the proposition we will not prove is

Proposition 8.4 (Lebesgue to Riemann-Stieltjes)**.**

$$\int_{\mathbb{R}} x \; dP = \int_{\mathbb{R}} x \; dF(x),$$

where F is the distribution function of X. More generally,

$$\int g(x) \; dP = \int_{\mathbb{R}} g(x) \; dF(x).$$

For simple random variables, this should be easy. For F a piecewise-constant function dF is only non-zero at a finite collection of points. As an exercise, prove this proposition for simple random variables. Once you prove it for simple random variables, it holds in general due to some limit theorems.

The outcome of this is this theorem, very very useful.

Theorem 8.5 (Change of Variables)**.**

$$\mathbb{E}X = \int_{\mathbb{R}} x \; dF(x)$$

and more generally,

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x) \; dF(x).$$

This gives you a hands-on formula for computing distributions/expectations. If X is discrete, then the theorem is not very useful. It doesn't make sense to take integration in this case, when we already have everything in terms of sums. The formula is very useful if X is absolutely continuous. Then, the theorem is useful, and it gives the following formula:

$$\mathbb{E}X = \int_R f(x)\ dx,$$

where $f = dF$ is the density function of X. And more generally,

$$\mathbb{E}g(X) = \int_R g(x)f(x)\ dx. \tag{9}$$

Heuristically, you should think of this formula (9) is you average over all values of $g(x)$ with this weight $f(x)$. If we just had dx, the $f(x)$ shows that it must prefer some values over another.

The interesting thing to do here in (9) is to integrate by parts. So, the corollary to this theorem is

Corollary 8.6. *If $X \geq 0$, then*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x)\ dx,$$

and more generally,

$$\mathbb{E}g(X) = g(0) + \int_0^\infty \mathbb{P}(X > x)\ dg(x).$$

We'll prove it next time (using integration by parts). If you know something about the distribution, we talk about tails $\mathbb{P}(X > x)$. If the tails decay fast, then there's no weight around the "ends." For example, we can use this on the Gaussian random variable. Then, we can ask what is the likelihood of the random variable outside of some range.

NOVEMBER 2, 2007

We'll work expectation a little bit more. For a random variable X with a distribution function F,

$$\mathbb{E}X = \int_{\mathbb{R}} x\ dF(x).$$

Note, $dF(x)$ is a function, so this is Riemann-Stieltjes integral. And if X has density f (that is, X is absolutely continuous), then

$$\mathbb{E}X = \int_{\mathbb{R}} x\ f(x)\ dx$$

can be computed by this standard (Riemann) integral. The moment that you see

$$\mathbb{E}X = \int_{\mathbb{R}} x \; dF(x),$$

you should think about integration by parts. So, this wil be the corollary that we just stated:

Corollary 8.7. *Let $X \geq 0$ be a random variable*[38]*. Then,*

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x) \; dx,$$

and more generally, you can do the same for a function of a random variable:

$$\mathbb{E}g(X) = \int_0^\infty \mathbb{P}(X > x) \; dg(x) + g(0).$$

Proof. For the proof, we'll use integration by parts for Riemann-Stieltjes integral (it has the same form as for Riemann integrals). So, the formula is

$$\int_a^b G(x) \; dF(x) = G(x)(Fx)|_a^b - \int_a^b F(x) \; dG(x).$$

The problem is we have an infinite range integral. In what range is the integral going to be zero? For negative values of x, certainly. If you have any arbitrary bounded random variable, then the distribution is zero outside of a range: Note, for bounded random variables, say $X \in [a, b]$,

$$\mathbb{E}X = \int_a^b x \; dF(x).$$

This is because $F(x) =$ constant outside the range $[a, b]$. So, this is how we reduce this integral. But, we know that an arbitrary random variable is approximated by bounded random variables.

1. Reduction to Bounded random variables (by truncation). We consider the random variable $X_R = \min(X, R)$. So X_R is a non-negative random variable (when $R \geq 0$), and $X_R \nearrow X$ pointwise. This is the setting of one of the theorems (Monotone Convergence Theorem). The MCT yields $\mathbb{E}X_R \to \mathbb{E}X$.

   The distribution function $F_R(x)$ of X_R is

$$F_R(x) = \begin{cases} F(x), & x < R \\ 1, & x \geq R. \end{cases} \tag{10}$$

[38] It's just convenient to state it for non-negative random variables. It's not that you can't state it for other random variables.

Now, we can compute

$$\mathbb{E}X_R = \int_0^R x \; dF_R(x)$$
$$\overset{almost}{=} \int_0^R x \; dF(x).$$

This concludes the first step: Everything became the same except the range became finite.

2. Now we integrate by parts:

$$\mathbb{E}X_R = x \; F_R(x)|_0^R - \int_0^R F_R(x)$$

By (10), we know the value of F_R at the endpoints:

$$= (R \cdot 1 - 0 \cdot 0) - \int_0^R F(x) \; dx = \int_0^R (1 - F(x)) \; dx = \int_0^R \mathbb{P}(X > x) \; dx.$$

and we know that this $\to \int_0^\infty \mathbb{P}(X > x) \; dx$, as $R \to \infty$.

This completes the proof. □

These issues about the infinite ranges may cloud issues a bit, but the heart of the proof is in the integration by parts.

These are some exercises to see how this is applied.

8.1.1 Exercises

1. Uniform random variable on $[0, 1]$. The density is

$$f(x) = \begin{cases} 1, & x \in [0, 1] \\ 0, & \text{otherwise.} \end{cases}$$

(An example is the delayed flight.)

Let's compute the expectation. In this case, we know the density function f, so we don't integrate by parts:

$$\mathbb{E}X = \int_0^\infty x \; f(x) \; dx = \int_0^1 x \; dx = \frac{x^2}{2}|_0^1 = \frac{1}{2}$$

2. Standard normal random variable. The density was

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for all x. Then, the expectation is

$$\mathbb{E}X = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x e^{-x^2/2} \; dx$$

The integrand is an odd function, so the integral is zero.

3. More generally, for symmetric random variables.

   Definition 8.8. *We call a random variable X symmetric if $X \overset{d.}{=} -X$. In words, $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in [-b,-a])$.*

   The standard normal random variable is symmetric. The density function f of a symmetric random variable is even. Then $x \times f(x)$ is odd, so the expectation is zero (assuming $\mathbb{E}X$ exists). (Prove this for general symmetric X, without assuming it has a density function f.)[39]

4. Exponential random variable X with parameter $\lambda > 0$. It has the distribution function

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0. \end{cases} \tag{11}$$

   We can apply Corollary 8.7. We get

$$\mathbb{E}X = \int_0^\infty (1 - F(x))\, dx = \int_0^\infty e^{-\lambda x}\, dx = \frac{1}{\lambda}.$$

9 Variance

Every random variable has two parameters. One is expectation. The other is variance.

Definition 9.1. *The variance of X is* $\operatorname{var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$.

The variance of X measures how spread the values of X are. It shows the "typical distance" to the center of the distribution. We also define the *standard deviation* of X:

$$\text{st. dev}(X) = \left(\mathbb{E}(X - \mathbb{E}X)^2\right) \overset{1/2}{=} \sqrt{\operatorname{var}(X)}.$$

The standard deviation of X shows how far X is typically from its mean.

Proposition 9.2. $\operatorname{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$

Proof. $\operatorname{var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2) = \mathbb{E}X^2 - 2(\mathbb{E}X)(\mathbb{E}X) + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$. □

Proposition 9.3 (Dilation / Translation)**.** *1.* $\operatorname{var}(X + b) = \operatorname{var}(X)$.

2. $\operatorname{var}(aX) = a^2 \operatorname{var}(X)$.

Proof. Exercise. □

[39] We could actually do example 1 here, by shifting the function.

Corollary 9.4 (Normalization)**.** *Let X be a random variable with mean*[40] *μ and variance σ^2. So, we have a general random variable X. Now, we want to create a standard random variable (with mean 0 and variance 1). How do we do this? First, we create mean 0, by shifting:*

$$X - \mu$$

Now, to get the variance to be 1, we divide by σ.

$$\frac{X-\mu}{\sigma}.$$

This random variable has mean 0 and variance 1.

Example 9.5.

1. *Uniform random variable on $[0,1]$. What does it take to compute the variance? We need to know the mean $\mathbb{E}X$, and we need to know the mean of X^2 (by the proposition). We already have computed the mean as $\frac{1}{2}$. Now, we need to compute the expectation of X^2. we do it by the usual formula:*

$$\mathbb{E}X^2 = \int_{-\infty}^{\infty} x^2 \ f(x) \ dx = \int_0^1 x^2 \ dx = \frac{x^3}{3}|_0^1 = \frac{1}{3}.$$

So then the variance is $\operatorname{var}(X) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$. Then, the $\operatorname{st.dev}(X) = \frac{1}{\sqrt{12}} \approx 0.29$.

2. *Standard normal random variable. We already know $\mathbb{E}X = 0$. The expectation of X^2? We apply the same formula:*

$$EX^2 = \int_{-\infty}^{\infty} x^2 \ f(x) \ dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} \ dx = 1.$$

Then $\operatorname{var}(X) = 1 = \operatorname{st.dev}(X)$.

3. *General normal random variable $Y = \sigma X + \mu$. Now Y has mean μ and variance σ^2, be the proposition. The density of Y is*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

This distribution is denoted $N(\mu, \sigma^2)$. (N for normal.)

NOVEMBER 5, 2007

One announcement. So there will be no office hour today. I think we'll work out an example with a specific distribution of how to compute the expectation and variance. We've worked with uniform, standard normal, normal, exponential (HW4 Exercise) distributions.

[40] Expectation of X is mean of X

9.1 Bernoulli Distribution

We'll do now some examples of discrete distributions. The simplest example is the *Bernoulli distribution with parameter p*[41]. Here $X \in \{0, 1\}$ given by the probability $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

An example of a Bernoulli random variable is a coin toss. In the coin toss, $X = 1$ if H and $X = 0$ if T. In this case, if the coin is fair, then the parameter $p = \frac{1}{2}$.

Recall that for discrete random variables X, the expectation is just the series, where the series is over all values that it can take, with weight:

$$\mathbb{E}X = \sum_k x_k \mathbb{P}(x = x_k).$$

In our case, $\mathbb{E}X = 1 \cdot p + 0 \cdot (1 - p) = p$. To compute the variance, we use the formula $\text{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. We first compute

$$\mathbb{E}X^2 = \sum_k x_k^2 \mathbb{P}(X = x_k)$$

to get $\mathbb{E}X^2 = 1^2 \cdot p + 0^2 \cdot (1 - p) = p$. So, the variance is the difference

$$\text{var}(X) = p - p^2 = p(1 - p). \tag{12}$$

In particular, $\text{st. dev}(X) = \sqrt{p(1-p)}$, and you can think of this as a geometric mean. For a fair coin, $\text{st. dev}(X) = \frac{1}{2}$. That's what you should expect, because the mean has to be the center. The deviation is always $\frac{1}{2}$, so the standard deviation is $\frac{1}{2}$. This is the simplest example.

9.2 Binomial Distribution

The next simplest example is the *Binomial Distribution with parameters* (n, p). Assume we perform n independent trials, with the probability of success in each trial equal to p. So the random variable X will be the number of successes. X is then called a *Binomial random variable* (with these two parameters).

The example is the number of heads in n independent tosses of a fair coin. Here, $p = \frac{1}{2}$. The Binomial distribution is a little more complicated, because as opposed to the Bernoulli distribution, $X \in \{0, 1, 2, \ldots, n\}$. To compute its parameters like expectation and variance, we need to compute the distribution. How do you compute $\mathbb{P}(X = k)$, the probability of k successes? How many ways are there to find k successes among n trials? $\binom{n}{k}$.

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

(We have to compute the probability of each of the $\binom{n}{k}$ times as k successes and $n - k$ non-successes. So, we have a formula. This is bad, because of the algebra.

[41] Sometimes, people take Bernoulli random variables with values ± 1, but there are two conventions.

So, here's how we'll do this. We'll have a sum of random variables. Then, the expectation becomes a simpler problem, because of the linearity of expectation.

We will represent X as a sum of random variables

$$X = \sum_{j=1}^{n} X_j$$

over all of the trials j, where

$$X_j = \begin{cases} 1 & \text{if success in } j\text{-th trial} \\ 0 & \text{otherwise} \end{cases}$$

Then every X_j is a Bernoulli random variable with parameter p. That's good, because then we know the expectation of every X_j.

$$\mathbb{E}X_j = p.$$

Then, we just use the linearity of expectation. We get

$$\mathbb{E}X = \sum_{j=1}^{n} \mathbb{E}X_j = np.$$

We haven't used independence, actually. This is a good example that you can sometimes to probability theory without independence. So, this holds also for arbitrary trials. You should still expect the same for dependents.

Let's go for the variance now. We have to compute $\mathbb{E}X^2$. Again, we do not want to use the distribution (with its wild binomial coefficients). So, to simpify,

$$\mathbb{E}X^2 = \mathbb{E}\left(\sum_{j=1}^{n} X_j\right)^2 = \mathbb{E}\left(\sum_{j=1}^{n} X_j^2 + \sum_{i \neq j} X_i X_j\right).$$

We again use the linearity of expectation, which allows us to move the expectation inside, so the problem reduces to computing the expectations of these individual terms.

The first type of term is easy: $\mathbb{E}X_j^2 = \mathbb{E}X_j = p$, so this was easy. But what about $\mathbb{E}(X_i X_j)$? Each of them is either zero or one, so the product is zero or one. When is it 1? This is when we have **both**:

$$X_i X_j = \begin{cases} 1 & \text{if success on } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

By independence, $\mathbb{P}(\text{successes in } i \text{ and } j) = p^2 \implies \mathbb{E}X_i X_j = p^2$. So, now, we have everything we need:

$$\mathbb{E}X^2 = n \cdot p + (n^2 - n) \cdot p^2.$$

Then, the variance is

$$\text{var}(X) = np + (n^2 - n)p^2 - (np)^2 = np - np^2 = np(1 - p).$$

Now a little surprise here. The variance of X_j was $p(1-p)$, by (12). We see that

$$\operatorname{var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \operatorname{var}(X_j)$$

This is not an accident. This always happens for indpendent X_j. We will show this later, and it would have allowed us to circumvent all of this calculation. Then st. dev$(X) = \sqrt{np(1-p)}$. For a fair coin, $= \frac{\sqrt{n}}{2}$.

Corollary 9.6. *The difference between the number of heads and tails in n independent tosses of a fair coin is $\sqrt{n}$ (typically).*

For example, if you have 100 tosses, you should expect a difference of 10. If you increase n, then the portion of difference, the proportion will go to 0. So, this predicts the rate at which.

9.3 Poisson Distribution

This is one of the remarkable distributions in probability that does not come up naturally. The other is the standard normal. This is a limiting law. The random variable X is a Poisson random variable with parameter λ if:

- $X \in \{0, 1, 2, \ldots\}$.
- $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \qquad k = 0, 1, 2, \ldots.$

This strange distribution is explained by the fact that the Poisson distribution is a good approximation of the Binomial distribution, with parameters n and p, for $\lambda = np$.

The approximation is good (this is important) if n is large while the mean np is moderate. We have very many trials, like 1000 trials, for example. But you expect only 5 of them to be successful. In particular, if you let n go to infinity but let np stay constant, then this is the setup.

What do we mean by large? We should see the proof of this. $n = 20$ is large and $p = 0.05$ (thus $np = 1$). Let's do the proof, since it's very easy.

Proof. Let Y be Binomial with parameters n and p. We want to approximate

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Here, we want to see $\lambda^k/k!$. So we isolate this, and we estimate the rest:

$$= \frac{\lambda^k}{k!} + \frac{n(n-1)\cdots(n-k+1)}{n \cdot n \cdots n}\left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k},$$

The third term ($e^{-\lambda}$) is good to keep. For the second factor, the ratio is ≈ 1. Finally, if k is constant, then the fourth factor is ≈ 1. Thus, $\mathbb{P}(Y = k) \approx e^{-\lambda}\frac{\lambda^k}{k!}$. □

The Central Limit Theorem will take over if np is large. The Poisson distribution is often used for "$n = \infty$". This is not a rigorous thing. But, if n goes to infinity, and the number of successes is to remain constant (accidents occur with fixed average rate λ in number of occurances per unit of time), then the Poisson distribution measures how many accidents occur in a unit interval of time. (n plays no role in the final formula, because there is no n in the end.) There is a big history of this Poisson distribution and how it is used. I like Wikipedia. Wikipedia has a good article on Poisson distributions.

Okay, so the expectation of X. Again, this is a sum

$$\mathbb{E}X = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

To compute $\mathbb{E}X^2$, we compute $\mathbb{E}X(X-1)$, because that will allow us to do the same trick as before. Namely, pull $e^{-\lambda}$ outside.

$$\mathbb{E}X(X-1) = \sum_{k=0}^{\infty} k(k-1) e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} =$$

$$= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2$$

But, we didn't compute $\mathbb{E}X^2$, we computed $\mathbb{E}X(X-1)$, but now can just subtract. So, $\mathbb{E}X^2 - \mathbb{E}X = \lambda^2$, thus $\mathbb{E}X^2 = \mathbb{E}X + \lambda^2 = \lambda + \lambda^2$. So $\mathrm{var}(X) = (\lambda + \lambda^2) - \lambda^2 = \lambda$, and $\mathrm{st.\,dev}(X) = \sqrt{\lambda}$.

Go check out the Wikipedia page!

NOVEMBER 7, 2007

How was the exam? Did anybody have success on the bonus problem? Usually, when you have this something starts to be bad at $p = 2$, zzz.

We'll now do the first real probability inequality.

10 Markov's and Chebychev's Inequalities

If we know the tails $\mathbb{P}(X > x)$ of a random variable $X \geq 0$ for every x, then we know the expectation. Then we can compute the expectation. We know this is $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > x)\, dx$. So if you know how a probability settles down for large values of X, then you know the expectation[42].

Very often we have the converse problem. What if we know $\mathbb{E}X$, and want to compute the tails $\mathbb{P}(X > x)$. Sometimes, the expectation is easier to compute than the tail (for example, in the binomial distribution). The tails for binomial distributions is not easy to compute.

[42] Yes, this is for positive random variables. We know what to do with negatives.

One bad thing and one good thing: First, we can not compute it exactly. If you only know $X \geq 0$ and $\mathbb{E}X$, then it may be a Gaussian random variable (which settles down really fast) or it may be a "heavier" random variable in the tails. But, we do have two inequalities. The first is Markov's inequality. It is not sharp, but it is useful in many cases.

Theorem 10.1 (Markov's Inequality)**.** *Suppose $X \geq 0$ is a random variable. Then,*

$$\mathbb{P}(X > x) \leq \frac{\mathbb{E}X}{x} \quad \textit{for all } x > 0. \tag{13}$$

So if you think of the expectation is a constant, we can say that the tail is "below" $\frac{1}{x}$. So, we know this is true for Gauss.

In particular, with probability at least $\frac{1}{2}$, the random variable X does not exceed $2\mathbb{E}X$. Just let $x = 2\mathbb{E}X$.

Let's make a picture. It's not needed for the formal proof. Think of $\Omega = [0, 1]$. For every ω, we graph $X(\omega)$. Then we do a truncation of X at x. So, we want to estimate the interval $I = \{X \geq x\}$ of $[0, 1]$. We define a new random variable Y which will take value x in I and 0 otherwise.

The area under the whole graph is $\mathbb{E}X$. The little box is $\mathbb{E}Y$. So $\mathbb{E}X > \mathbb{E}Y$.

Proof. Define $Y := x\mathbf{1}_{\{X \geq x\}}$. First, we say that $Y \leq X$ pointwise. (simple, just consider the two cases: when $X \geq x$ and $X < x$.) Therefore,

$$\mathbb{E}Y \leq \mathbb{E}Y. \tag{14}$$

$$\mathbb{E}Y = x \cdot \mathbb{P}(X \geq x).. \tag{15}$$

Then (14) and (15) imply Markov's inequality. □

Sometimes people say: $\frac{1}{x}$ is good, but not really good, because we cant integrate this. One happens when we integrate both sides of (13)? On the left, we get expectation. On the right, we have $\mathbb{E}X$ times a non-integrable function.

There is a fix to this problem.

Theorem 10.2 (Chebychev's Inequality)**.** *Suppose $X \geq 0$ is a random variable, and $p > 0$ (usually 1). Then*

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X^p}{x^p} \textit{ for all } x > 0. \tag{16}$$

So that fact that X^p has expectation is a stronger requirement than X has expectation. If you increase p, it will be harder for X^p to have expectation. But in the end, you will be rewarded with an integrable right hand side.

The proof is an easy reduction to Markov's Inequality.

Proof. $\mathbb{P}(X \geq x) = \mathbb{P}(X^p \geq x^p) \leq \frac{\mathbb{E}X^p}{x^p}$, by Markov's inequality. □

Remark 10.3. *This power p is not important in this proof. You can have an arbitrary increasing function.*

We use this a lot.

One corollary to Chebychev's inequality is:

Corollary 10.4 (Variance)**.** *Suppose X is a random variable with mean μ and variance σ^2. Then*

$$\mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2} \text{ for any } t > 0. \tag{17}$$

While Chebychev's says something about expectation, the corollary will say something about the variance. As soon as we know that the random variable is smaller than average, we can say that the value is small (with high probability). Here, we say something about the variance. Suppose X has a small variance σ^2. What is the probability that x will be far away from its mean?

You can say with high probability that x is; close to its mean. The proof is simple by Chebychev's inequality

Proof. Use Chebychev's Inequality for the random variable $|X - \mu| \geq 0$. Then we obviously will set $x = t\sigma$ and $p = 2$. Then $\mathbb{E}|X - \mu|^2 = \sigma^2$, and we have by Chebychev's inequality,

$$\mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{\sigma^2}{(t\sigma)^2} = \frac{1}{t^2}.$$

□

In particular, for $t = \sqrt{2}$, every random variable X with variance σ^2 is within $\sqrt{2} \cdot \sigma$ from its mean, with probability $\geq \frac{1}{2}$. So every random variable is close to its mean, and the units are (so to speak) the standard deviation.

The next section, we'll try to understand what happens with more than one random variable. This is a big section on independent random variables.

11 Independent random variables

Recall that before this, we only had independent events. The first thing to understand is the *joint distribution*, as a way to handle more than one random variable at once. This is familiar from last lectures.

Suppose $X_1, \ldots, X_n$ are random variables. We will look at them jointly as a random vector $X = (X_1, \ldots, X_n)$. So again, $X(\omega) = (X_1(\omega), \ldots, X_n(\omega)) \in \mathbb{R}^n$. (It is clear why this is called a random vector: it's like a random point in space.)

As before, the random vector X induces a distribution P on $\mathbb{P}^n$. Recall, we assigned

$$P(A) = \mathbb{P}(X \in A)$$

for a Borel set A in $\mathbb{R}$. We do the same thing. We assign

$$P(A) = \mathbb{P}(X \in A) \tag{18}$$

for a Borel set $A \subseteq \mathbb{R}^n$. The P in (18) is called the *joint distribution* of random variables $X_1, \ldots, X_n$.

What did we do? We said, this is a good thing, but there's too many sets. So, we focused on sets $X \leq x$. Similarly to the case of random variables, the joint distribution, the joint distribution is determined by half-infinite boxes $\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$. Namely,

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \tag{19}$$

the *joint distribution function* of $X_1, \ldots X_n$. We didn't really do anything that we haven't already done.

Similarly to random variables, the joint distribution is *absolutely continuous* if $\exists\ f : \mathbb{R}^n \to \mathbb{R}$, called the *joint density* of $X_1, \ldots, X_n$, and such that

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int_{y \leq x} f(y_1, \ldots, y_n)\ dy. \tag{20}$$

Here, $\{y \leq x\} := \{y \in \mathbb{R}^n : y_1 \leq x_1, \ldots, y_n \leq x_n\}$. Then,

$$\mathbb{P}(X \in A) = \int_A f(x)\ dx \text{ for Borel set } A \subseteq \mathbb{R}^n.$$

The application of this is that now we can define independence. We have to know how to look at two random variables at once.

11.1 Defining independence of random variables

Let's go back to the case of events and recall that events A, B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

We call random variables X, Y *independent* if[43]

$$\mathbb{P}(X \in C, Y \in D) = \mathbb{P}(X \in C) \cdot \mathbb{P}(Y \in D)$$

for all Borel sets C, D in $\mathbb{R}$.

Question: Is this the same thing as the expectations the same? zzz

- **Answer:** It's not the same, but it will be if ... zzz. Those are called uncorrelated.

In other words, what we just wrote here are that the events $\{X \in C\}$ and $\{Y \in D\}$ are independent. The collection of $\{X \in C\}$ is an induced σ-algebra. So, equivalently, $\sigma(X)$ and $\sigma(Y)$ are independent.

The general definition (for more than two random variables is):

[43] if the fact that X takes on a specific value is independent of Y taking another value. Those are two events. Therefore, we define the independence of variables by independence of events.

Definition 11.1. *Random variables $X_1, \ldots, X_n$ are independent if*

$$\mathbb{P}(X_1 \in B_1, \ldots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \cdot \cdots \cdot \mathbb{P}(X_n \in B_n)$$

for every Borel sets $X_1, \ldots, B_n$ in $\mathbb{R}$.

Question: Do we need to say for every subsets?

- **Answer:** Actually no. Why, because we can just choose B_1 to be the full line, for example. So the definition differs in that we do not talk about different subsets (because we can handle them by allowing the appropriate B_i to be the full line).

So equivalently (to Definition 11.1), $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.

Example 11.2. *Let X and Y be two independent random variables in $[0,1]$. Then, we can compute*

$$\mathbb{P}\left(X \leq \frac{1}{2}, Y \leq \frac{1}{2}\right) = \mathbb{P}\left(X \leq \frac{1}{2}\right) \cdot \mathbb{P}\left(Y \leq \frac{1}{2}\right) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

There are too many Borel sets, so this is not a practical way to verify independence!

11.2 Verifying Independence

We can check just on generators (half-intervals).

Theorem 11.3. *$X_1, \ldots, X_n$ are independent if and only if*

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdot \cdots \cdot \mathbb{P}(X_n \leq x_n)$$

for every $x_1, \ldots, x_n \in \mathbb{R}$.

Proof. ($\Longrightarrow$) trivial.

($\Longleftarrow$) Recall the theorem on October 12th: namely, Theorem 3.8.

We use Theorem 3.8 for $\mathcal{A} = \{\text{sets of the form } \{X \leq x\}, x \in \mathbb{R}\}$, and $\mathcal{B} = \{\text{sets of the form } \{Y \leq y\}, y \in \mathbb{R}\}$. We know these are independent, by assumption, and closed under intersection (because they are half-intervals).

Finally, $\sigma(\mathcal{A}) = \sigma(X) = \{\text{sets } \{X \in A\}, \text{ Borel } A\}$, and same for $\sigma(\mathcal{B})$. So $\sigma(X), \sigma(Y)$ are independent. □

NOVEMBER 9, 2007

Here, we will assume the f and g that follow are measurable.

Proposition 11.4. *If $X_1, \ldots, X_n$ are independent, then $f_1(X_1), \ldots, f_n(X_n)$ are independent.*

For example, if X, Y are independent, then X^2, Y^2 independent.

Proof. First, observe

$$\mathbb{P}(f_1(X_1) \in B_1, \ldots, f_n(X_n) \in B_k) = \mathbb{P}(f_1(X_1) \in B_1) \cdots \mathbb{P}(f_n(X_n) \in B_n). \quad (21)$$

$f_k(X_k) \in B_k$ is equivalent to $X_k \in f^{-1}(B_k)$.

Use this in (21) $\Longrightarrow$ □

Remark 11.5. *All of the above (definition and results on independent) are true for random vectors.*

Corollary 11.6. *If $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent random variables, then $f(X_1, \ldots, X_n)$ and $g(Y_1, \ldots, Y_m)$ are independent.*

Why is this immediate? It's not. It is if you think of random vectors.

Proof. Consider random vectors $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_m)$. Apply the proposition for two $f(X)$, $g(Y)$. □

(So, the usual way people state this is: If $X_1, \ldots, X_{2n}$ are independent, then $X_1 + X_2 + \cdots X_n$ and $X_{n+1} + \cdots X_{2n}$ are independent.)

How can we consider tossing a coin and rolling a dice at the same time? These are defined in different probability spaces. In this definition of independence, we assume they are defined on the same probability space. Is there a way to do this? We usually combine these things by studying product spaces.

12 Product Probability Spaces

The product probability space will allow us to combine different probability spaces into one space (such as throwing a coin and rolling a dice), and also, this will allow us to construct independent events and random variables. This is the goal, and it's easily achieved.

We just consider (for simplicity), just two different probability spaces. THe construction is easily generalized to more than two.

Consider two probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$. We are going to define their *product* $(\Omega, \mathcal{F}, \mathbb{P})$.

The simplest thing is to define the product of the Omegas:

$$\Omega := \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}.$$

So far, there is no structure. This is just a set.

As an example, consider $\Omega_1 = [0, 1]$ and $\Omega_2 = [0, 1]$. Then Ω is the square $[0, 1]^2$.

Why don't we take

$$\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2?$$

This won't have things such as circles. So,

$$\mathcal{F} \neq \mathcal{F}_1 \times \mathcal{F}_2.$$

So, let

$$\mathcal{A} = \{A_1 \times A_2 : A \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\},$$

and let $\mathcal{F} := \sigma(\mathcal{A})$. Then you do have circles.

There is also a twist in how to define the measure $\mathbb{P}$. We define the measure

$$\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1) \times \mathbb{P}_2(A_2) \text{ for } A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$$

Think of boxes.

We define the measure $\mathbb{P}$ on the boxes first, and then we extend it (since $\mathbb{P}$ was defined on a collection of generators). It can be uniquely extended, but this is not a trivial fact (called the Caratheodory's Extension Theorem). This theorem is one-half of the construction of Lebesgue measure. Why is it non-trivial? Apply it to intervals in the plane. For shorthand, we write $\mathbb{P}_1 \times \mathbb{P}_2$ for $\mathbb{P}$.

Then, the *product space* $(\Omega, \mathcal{F}, \mathbb{P})$ is defined.

Example 12.1. *The product of* $[0,1]$ *and* $[0,1]$ *is* $[0,1]^2$ *with Lebesgue measure*[44].

Question: It seems like we got independence for free.

- **Answer:** Yes, because we defined our measure the way that we did.

This way, we define independence for events. So I'll just mention one powerful result. You've probably heard of it.

Remark 12.2. *We never used the fact that these are probability spaces. The construction holds for arbitrary measure spaces, in which* $\mathbb{P}(\Omega \neq 1$.

Theorem 12.3 (Fubini's Theorem)**.** *If* $f(\omega_1, \omega_2)$ *is integrable on a product space* $\Omega_1 \times \Omega_2$ *with product measure* $\mu = \mu_1 \times \mu_2$*, then*

$$\int_{\Omega_1 \times \Omega_2} f \, d\mu = \int_{\Omega_1} \left(\int_{\Omega_2} f \, d\mu_2 \right) d\mu_1 = \int_{\Omega_2} \left(\int_{\Omega_1} f \, d\mu_1 \right) d\mu_2.$$

This double integral is equal to the iterated integrals.

Random variables can be dependent or independent. Regardless of that, they have a joint distribution. The joint distribution will tell you about that phenomenon. So, let's understand the joint distribution of independent random variables.

13 Joint Distribution of Independent Random Variables

Just from the definition of the independence (just the fact that the probability factors into a product), we see that a joint distribution function

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \times \cdots \times F_{X_n}(x_n),$$

[44] two-dimensional

so the distribution function factors into a product of previous distribution functions.

Theorem 13.1. *The random variables $X_1, \ldots, X_n$ are independent if and only if their joint distribution $\mathbb{P}$ on $\mathbb{R}^n$ is the product of the distributions P_k of X_k on $\mathbb{R}$.*

Thus, the only way random variables can be independent is when they are defined in a product space. The product space is the only way to think about independent random variables.

Proof. In the $\Longrightarrow$ direction, by the uniqueness theorem of measure theory (measure uniquely defined on the boxes), it is enough to check that $P(A) = (P_1 \times \cdots \times P_n)(A)$ for $A \in \mathcal{A}$.

As we know, every $A \in \mathcal{A}$ has the form $A = A_1 \times \cdots A_n$, where A_k is Borel. Then, by independence,

$$\begin{aligned} P(A_1 \times \cdots A_n) &= \mathbb{P}((X_1, \ldots, X_n) \in A_1 \times \cdots \times A_n \\ &= \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) \\ &= \mathbb{P}(X_1 \in A_1) \times \cdots \times \mathbb{P}(X_n \in A_n) \\ &= P_1(A_1) \times \cdots \times P_n(A_n). \end{aligned}$$

In the other direction ($\Longleftarrow$), from last lecture, to check independence of $X_1, \ldots, X_n$, it is enough to handle half-intervals. Namely, if

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

factors, we are done. But this (def of joint distribution) is

$$P((-\infty, x_1] \times \cdots \times (-\infty, x_n]).$$

By the assumption, we know this factors:

$$\begin{aligned} &= P_1(-\infty, x_1] \times \cdots \times P_n(-\infty, x_n] \\ &= \mathbb{P}(X_1 \leq x_1) \times \cdots \times \mathbb{P}(X_n \leq x_n). \end{aligned}$$

And so we are done. □

The moral of today's lecture is: The only way to work with independence is to work through product spaces.

NOVEMBER 14, 2007

From last time:

Theorem 13.2. *Random variables $X_1, \ldots, X_n$ are independent $\Longleftrightarrow$ their joint distribution P (on $\mathbb{R}^n$) is the product of distributions P_k of X_k (on $\mathbb{R}$).*

Professor Vershynin is away, so that's why zzz. What you started doing at the end of last time: here is a theorem about random variables. It's not part of the definition of it, but it's proved from the definition of it. It says that Random variables $X_1, \dots, X_n$ are independent if and only if their joint distribution P (on $\mathbb{R}^n$) is the product of distributions individual distributions.

Let's start with a corollary:

Corollary 13.3. *Given arbitrary probability measures $P_1, \dots, P_n$ on $\mathbb{R}$, there exists independent random variables $X_1, \dots, X_n$ with distributions $P_1, \dots, P_n$.*

So this is an existence theorem.

Proof. Let $P = P_1 \times \cdots \times P_n$ be the product probability measure on $\mathbb{R}^n$. Every probability measure on $\mathbb{R}^n$ is a distribution of some random variable. So you have to produce a random variable (or really, random vector). How will we define it? Take the probability space to be ... what? Okay, so now we have to think a little bit. We need a triple $(\Omega, \mathcal{F}, \mathbb{P})$. Take the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to be $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P)$. So, that's the probability space. Here, $\mathcal{B}$ sounds for the Borel σ-algebra.

So now, we need to define a random variable X which is a function of this set. What will the values of $X(x)$ be? What is going to be the image of a point x for this to work? Any guesses? Yes $X(x) = x$. This is a well-defined function from your probability space, which is $\mathbb{R}^n$. It's defined on the correct probability space, and defined into the correct codomain space.

Now, let's check? We have $P(X \in B)$ is what? Well, X is just the identity function, so $X \in B$ if and only if $x \in B$. So $P(X \in B) = P(B)$, so the distribution of X is equal to P.

Furthermore, the $X_1, \dots, X_n$ are independent, by the previous theorem (because their distribution is the product distribution of the individual ones). $\square$

I should say, that as a remark, this corollary also holds for a countable collection of random variables. So if you go on and on with these probability measures, and you have countable number of them, you can find a sequence of random variables. So, that's called the Kolmogorov's Extension Theorem. You have to work considerably harder to prove it.

Second of all, I should say, so that this construction is done so that it's trivial do to it. However, if you fiddle with this construction a little bit, you can make it so that all your random variables are functions from the unit interval. These are standard variables. It takes a bit more work (you have to somehow invert the distribution).

The one thing you should (if you're dealing with this the first time) think about: many many different random variables have the same distribution. The distribution tells you how to compute probabilities, but it does not tell you everything.

How does all of this reflect to densities? Let's recall a few facts about densities.

1. First, a positive measurable function $f \geq 0$ is the density of a random variable X if $\mathbb{P}(X \in A) = \int_A f(x)\ dx$, for every Borel set $A \subset \mathbb{R}$. Of course, the density is only defined a. e. If you change the density on a set of measure zero, then nothing changes.

2. Two, if you have a random variable instead a random vector: the definition is the same if X is a random vector except that f is define on $\mathbb{R}^n$ and A is a Borel set in n dimensions.

3. Joint density of n random variables $X_1, \ldots, X_n$ is the density of the random vector $(X_1, \ldots, X_n)$.

Theorem 13.4. *Let $X_1, \ldots, X_n$ be random variables with densities $f_1, \ldots, f_n$ and joint density f. Then $X_1, \ldots, X_n$ are independent $\iff$ $f(x_1, \ldots, x_n) = f_1(x_1) \cdot \cdots \cdot f_n(x_n)$ for almost all $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$.*

There are some technical difficulties. In order to get a density from a distribution, we need to differentiate. If you have functions which are not continuous, then differentiation is kind of a problem. We need the following (not-so-easy-to-prove) result from measure theory:

Theorem 13.5 (Lebesgue differentiation theorem)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a Lebesgue integrable function. Then*

$$\frac{1}{\epsilon} \int_x^{x+\epsilon} f(y)\ dy$$

is going to converge (known from calculus) to $f(x)$, for almost all $x \in \mathbb{R}$.

This is a one-dimensional version of the Lebesgue differentiation theorem. You also have the n-dimensional version. So,

Theorem 13.6. *Moreover, if $f : \mathbb{R}^n \to \mathbb{R}$ and is Lebesgue integrable, then*

$$\frac{1}{vol(Q)} \int_Q (y)\ dy \to f(x)$$

for almost every x, where Q is a cube containing[45] x.

Proof of Theorem 13.4. Without loss of generality, we can assume $n = 2$. (Make it an exercise to check this).

X has density f, Y has density g, and (X, Y) has density φ. This means that $P(X \in A) = \int_A f(x)\ dx$, $P(Y \in B) = \int_B g(y)\ dy$, and $P((X, Y) \in A \times B) = \int_{A \times B} \varphi(x, y)\ dxdy$

So let's start proving $\Rightarrow$. Then X and Y are supposed independent. Then $\int_{A \times B} \varphi(x, y)\ dxdy$ must be equal to, by definition, the product of the individual probabilities $\int_A f(x)\ dx \times \int_B g(y)\ dy$. Let $A = [x_0, x_0 + \epsilon]$ and $B = [y_0, y_0 + \epsilon]$.

[45] x can be anywhere in the cube, including the boundary.

Then $A \times B$ is exactly a cube Q which contains (x_0, y_0). The situation is ripe for using exactly the theorem.

Send $\epsilon \to 0$, to get: the LHS (by the Theorem) will be $\phi(x_0, y_0)$ and the RHS is $f(x_0)g(y_0)$ for almost all (x_0, y_0). If you want to be a lot more careful, you can think about sets of measure zero. In particular, you can identify what your set of measure zero is, and what your final set of measure zero is.

For the direction ($\Longleftarrow$), it's even easier. (Of course, the other direction was easy to do by use of a sledgehammer). In this direction, we assume that the joint density is the product of the two densities for almost all $(x, y) \in \mathbb{R}^2$. Now, we use Fubini's theorem.

$$\begin{aligned} P(X \in A, Y \in B) &= \int_{A \times B} \varphi(x, y)\ dxdy \\ &= \int_{A \times B} f(x)g(y)\ dxdy \\ &= \int_A f(x)\ dx \int_B g(y)\ dy, \text{ by Fubini} \\ &= P(X \in A) \cdot P(Y \in B). \end{aligned}$$

So, by definition, X and Y are independent. □

Let's look at a few examples. The first example we should look at is the uniform distribution. Let X and Y be independent random variables, uniformly distributed on $[0, 1]$. This is two separate calls to your computer's uniform random number generator. So, the density of both X and Y is given by the density function

$$f(x) = \begin{cases} 1, & x \in [0, 1] \\ 0, & \text{otherwise.} \end{cases}$$

So the joint density of X and Y is

$$\varphi(x, y) = \begin{cases} 1, & \text{if } (x, y) \in [0, 1] \times [0, 1] \\ 0, & \text{otherwise.} \end{cases}$$

I should also say (as an unofficial exercise) try to prove that if you have a random vector uniformly distributed on a set $S \subset \mathbb{R}^2$, then its coordinates are independent if and only if $S = A \times B$.

Question: What is our infinitesimal integration?

- **Answer:** This is our two-dimensional Lebesgue measure. It doesn't say anything about the fact that the two dimensional function is corresponding to a random vector.

You could in principle, talk about densities with respect to other underlying measures, but we won't talk about that in this course.

Second example: the Gaussian distribution on $\mathbb{R}^n$. Also called the multivariate normal distribution. The density of each coordinate X_k is given by the

standard one-dimensional Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, x \in \mathbb{R}$$

and the coordinates are independent, so the density of $(X_1, \ldots, X_n)$ is

$$\varphi(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}|x|^2},$$

where $|x| = x_1^2 + x_2^2 + \cdots + x_n^2$. See you on Friday.

NOVEMBER 16, 2007

Today's the last day. It seems to be working now.

Theorem 13.7.

$$\mathbb{E}(XY) = \mathbb{E}X \times \mathbb{E}Y.$$

This is not hard to see.

Proof. let P_X and P_Y be the distributions of X and Y. Then these are the measures on $\mathbb{R}$ (because these are random variables). Then

$$\mathbb{E}X = \int x \, dP_X$$

and

$$\mathbb{E}Y = \int y \, dP_Y,$$

and on the other hand

$$\mathbb{E}(XY) = \int xy \, d(P_X \times P_Y)$$

since X and Y are independent. By Fubini, you can write this as an iterated integral (which falls into a product), which is exactly $\mathbb{E}X \times \mathbb{E}Y$.

But Fubini is not valid without any assumptions whatsoever. In this case, this is true because you can apply everything to $|X|$ and $|Y|$ instead of X and Y. (To justify Fubini, do the absolute values first.) □

By induction,

Corollary 13.8. *If $X_1, \ldots, X_n$ are independent random variables with finite expectations, then $\mathbb{E}(X_1 X_2 \cdots X_n) = \mathbb{E}X_1 \times \cdots \times \mathbb{E}X_n$.*

In general, two random variables X and Y for which $\mathbb{E}(X \cdot Y) = \mathbb{E}X \times \mathbb{E}Y$ are called *uncorrelated*.

So independent random variables are uncorrelated. It's not necessarily true that uncorrelated random variables are independent. For an example, let's look at $\Omega = [-1, 1]$ with uniform probability. Let's look at the two random variables

$X(x) = x$ and $Y(x) = 3x^2 - 1$. Then $\mathbb{E}X = 0$ and $\mathbb{E}(XY) = 0$. It also happens to be that $\mathbb{E}Y = 0$, but that doesn't matter. Then X and Y are uncorrelated, but not independent. How can you see this immediately? Prove by definition, sure, but how can you see if there is any sense in this notion of independence, they can be independent. Well, Y is a function of X, that's why! They are in functional relation to each other.

These two polynomials are the first two uncorrelated polynomials in this family called the *Legendre polynomials*. These come up in many areas of probability, combinatorics, differential equations, and so on and so forth.

13.1 Sums of independent random variables

Let's start with an example. If X and Y are random variables with known distributions, is the distribution of $X+Y$ determined? Certainly the expectation is determined (by the sum rule). But is the entire distribution determined? It isn't. Here is an example: Take any random variable X such that $X \overset{d.}{=} -X$. One example is flipping a coin, using heads for 1 and tails for -1. Another is $X(x) = x$. Then $X + X$ and $X + (-X)$ are what? All four summands have the same distribution. The first is equal to $2X$ and the second is equal to 0.

So, knowing the distribution of the summands is not enough. You actually need the joint distribution. If X and Y are independent, then we can compute the distribution of the sum.

So, if X and Y are independent, how can we compute the distribution of $X + Y$? Well, we'll see this next. So this is our next theorem.

Theorem 13.9. *X and Y are independent random variables with distribution functions G and G, respectively. Then, the distribution function of the sum $X + Y$ is given by*

$$H(z) = \int_{\mathbb{R}} F(z - y)\ dG(y) = \int_{\mathbb{R}} G(z - y)\ dF(y).$$

Before we go on to prove this theorem, let's just try a little bit to understand what's going on here.

Corollary 13.10. *If X and Y have densities f and g, respectively, then $X + Y$ has density*

$$h(x) = \int_{\mathbb{R}} f(x - y)g(y)\ dy.$$

This $h = f \star g$ is known has the *convolution* of f and g.

Proof of the Corollary. We have this formula here, so if you rite $F(z - y)$, this is of course

$$\int_{-\infty}^{z-y} f(x)\ dx,$$

which is by change of variables

$$\int_{-\infty}^{z} f(x-y)\ dx$$

(just replacing x with $x-y$). Now, we look at the formula H to get

$$H(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z} f(x-y)\ dx \right) g(y)\ dy$$

Now, we reverse the order in our double integral:

$$H(z) = \int_{-\infty}^{z} dx \int_{-\infty}^{\infty} f(x-y)g(y)\ dy$$

Therefore, we take derivative, and h is equal to the convolution. □

Proof of Theorem. The joint distribution of X and Y has distribution $P_X times P_X$, since theyr'e independent. Then $H(z) = P(X+Y \leq z)$ is equal to $\mathbb{P}((X,Y) \in \{(x,y) \in \mathbb{R}^2 : x+y \leq z\})$. So this is really saying nothing, but saying what we want to do a little bit clearer.

When you have a distribution, you can compute that probability by integrating over this set. So, we have

$$\int_{\{(x,y)\in\mathbb{R}^2 : x+y\leq z\}} d(P_X \times P_Y)$$

In the next step, instead of integrating over this set, we integrate over the entire plane and take the indicator of this set.

$$\int \int_{\mathbb{R}} \mathbf{1}_{\{(x,y)\in\mathbb{R}^2 : x+y\leq z\}} d(P_X \times P_Y)$$

By Fubini, we can rewrite this as an iterated integral

$$\begin{aligned}
&= \int_{\mathbb{R}} dP_Y \cdot \int_{\mathbb{R}} \mathbf{1}_{\{(x,y)\in\mathbb{R}^2 : x+y\leq z\}} dP_X \\
&= \int_{\mathbb{R}} dP_Y \cdot \int_{\mathbb{R}} \mathbf{1}_{\{(x,y)\in\mathbb{R}^2 : x\leq z-y\}} dP_X \\
&= \int_{\mathbb{R}} dP_Y \cdot \int_{\mathbb{R}} P_X(X \in (-\infty, z-y])dP_X \\
&= \int_{\mathbb{R}} F(z-y)\ dP_Y(y)
\end{aligned}$$

The only thing you need to remember that the integral with respect to a measure is the same as the Stieltjes integral with respect to a distribution function.

So, this is $\int_{\mathbb{R}} F(z-y)G(y)$. You just have to see things from earlier in the class. □

It's used only when X and Y have densities. The convolution are truly what's used the most. It's also used in the discrete case as well. Then, the formula is much easier then. You can do it as an exercise to write down the formulas in the case of discrete random variables:

Example 13.11. *Let X and Y be independent uniform random variables on $[0,1]$. Compute the sum of the two calls. What do you get? You get the triangle.*

$$f = \text{density of } X = \text{density of } Y$$

So the density of $X+Y$ is

$$h(x) = \int_{\mathbb{R}} \mathbf{1}_{[0,1]}(x-y)\mathbf{1}_{[0,1]}(y)$$

So $0 \le y \le 1$, but also $0 \le x-y \le 1$, so $y \le x$ and $y \ge x-1$.

What are the restrictions on x? The sum will have values between 0 and 2. It's best to not sort things out from the integral. So $x \in [0,2]$. If $x \in [0,1]$, then you will have two different restrictions in y. The only condition you have if $x \in [0,1]$ is $y \le x$, so you have $\int_0^x dy = x$. If $x \in [1,2]$, then the other condition is worthless, so you have $\int_{x-1}^1 dy = 2-x$. So, indeed, you get the triangle (or "hat" function) for h.

The final theorem that we'll prove today is:

Theorem 13.12. *If $X_1, \dots, X_n$ are independent random variables with finite variances, then*

$$\operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n).$$

This is very very important. This is because the expectation and the variance increase at the same rate. So, the deviation from the expectation increases at a slower rate.

Proof. Recall the formula for variance. The one that's most useful here is

$$\operatorname{var}(X) = \mathbb{E}((X - \mathbb{E}X)^2).$$

So,

$$\begin{aligned}
\operatorname{var}(X_1 + \cdots + X_n) &= \mathbb{E}((\sum_k X_k - \mathbb{E}X_k))^2) \\
&= \mathbb{E}(\sum_{k,\ell=1}^n (X_k - \mathbb{E}X_k)(X_\ell - \mathbb{E}X_\ell)) \\
&= \sum_{k,\ell=1}^n \mathbb{E}((X_k - \mathbb{E}X_k)(X_\ell - \mathbb{E}X_\ell)) \\
&= \sum_{k=1}^n \mathbb{E}((X_k - \mathbb{E}X_k)^2).
\end{aligned}$$

□

Okay, that's it for me. Good luck.

NOVEMBER 19, 2007

I'll put up the next homework today, so that you'll have more than a week to do it. So Janko Gravner talked on variance, independence, density of a sum of independent random variables.

There are two classical parameters associated to random variables: the mean and variance. I will show you more ways to control a random variable (i.e. decay at infinity). These decays (read, parameters) are best studied through moments:

14 Moments, L^p spaces, Inequalities

Let X be a random variable. We call $\mathbb{E}X^p$ the *p^{th} moment of X*. In almost all cases, we can safely assume that $p > 0$ and $p \geq 1$.

Often, $X \geq 0$ is the set up in which we consider these random variables and moments. Often, it is convenient to consider the *p^{th} absolute moment* $\mathbb{E}|X|^p$. If $p = 1$, we are talking about the expectation of X. If the expectation of X is zero, then the second moment will be the variance. The third moment will measure even more how X is spread on $\mathbb{R}$. So the mean shows the location of X. The variance shows how it's spread. The third moment shows even more how it's spread.

The most convenient way to do this is to map it into the theory of L^p spaces in analysis.

14.1 L^p Spaces

Let $f : \Omega \to \mathbb{R}$ be a measurable function (on a probability space). If we have functions that are equal almost everywhere, we will identify them.

Instead of working with individual functions f, we will work with the whole "class of equivalence" of the functions

$$\{g : g = f \text{ almost everywhere}\}.$$

We will work with the classes of equivalence, but we'll just verbally say that we work with functions. We do not want to distinguish a class from its representative.

The set of all functions (classes) for which the integral

$$|f|^p \, d\mu < \infty$$

is called *$L^p = L^p(\Omega) = L^p(\Omega, \mu) = L^p(\mu)$, the L^p space*.

Being in an L^p space is a question about the growth of a function. We measure the "size" of these functions with the *L^p norm*

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{1/p}. \qquad (p \geq 1).$$

Even if $p = \infty$, this makes sense (because we can think of the limit as p goes to ∞). Then (in analysis), this converges to the sup:

$$\|f\|_\infty = \sup_{\omega \in \Omega} |f(\omega)|. \qquad (p = \infty)$$

14.2 Inequalities

What's remarkable here is that there are a whole bunch of inequalities that govern these L^p spaces. One of these is the Minkowski inequalitiy.

Proposition 14.1 (Minkowski's Inequality)**.**

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \qquad (1 \leq p \leq \infty)$$

This is clear for ∞ and also clear for 1.

The following is easy intuitively, but it takes some time to prove it.

Proposition 14.2 (Jensen's Inequality)**.** *Consider any convex*[46] *function* φ*. Then*

$$\varphi\left(\int f \, d\mu\right) \leq \int \varphi(f) \, d\mu.$$

How should you think about this? The convexity of a function is a two-point condition. The function applied to the average is smaller than the average applied to the function. The proof of this reflects that.

If you take $\varphi(x)$ to be the absolute value, one gets

Corollary 14.3. $|\int f \, d\mu| \leq \int |f| \, d\mu$.

Proof. Take $\varphi(x) = |x|$. □

If you take $\varphi(x) = |x|^p$, then you get

Corollary 14.4. $|\int f \, d\mu|^p \leq \int |f|^p \, d\mu$ *for* $1 \leq p \leq \infty$.

Proof. Take $\varphi(x) = |x|^p$. For $p > 1$, φ is convex. □

In particular, if we take the p^{th} root of both sides, we can get that

$$\|f\|_1 \leq \|f\|_p \text{ for } p \geq 1.$$

So this immediately gets you an inequality about the norms. More generally, the proposition we have is that the norms are "well-ordered":

Proposition 14.5. $\|f\|_p \leq \|f\|_q$ *if* $0 \leq p \leq q \leq \infty$.

There, there is a whole range of norms. It's harder to control the ∞-norm than any other norm. This inequality contains much information: the average is smaller than the infinity [norm], etc.

Finally, there's

[46] μ is convex if $\varphi(\lambda x + \mu y) \leq \lambda\varphi(x) + \mu\varphi(y)$ if $\lambda + \mu = 1$ and $\lambda, \mu \geq 0$.

Proposition 14.6 (Hölder's Inequality)**.** $\int |fg|\, d\mu \leq (\int |f|^p\, d\mu)^{1/p} (\int |g|^q\, d\mu)^{1/q}$ *for* $\frac{1}{p} + \frac{1}{q} = 1$ *connected by this conjugacy.*

In other words, $\|fg\|_1 \leq \|f\|_p \|g\|_q$. Why is this useful? You've seen sometimes when you can integrate one factor of a product, but you don't know how to integrate the product (and integration by parts fails). Then you can use this to estimate the integral of the product. One partial case of this is where $p = q = \frac{1}{2}$ where you get the Cauchy-Schwarz inequality.

Corollary 14.7 (Cauchy-Schwarz inequality)**.** $\|fg\|_1 \leq \|f\|_2 \|g\|_2$.

Question: Is there any problem with the inequalities that you have here? It seems that if $p = 1$, then the triangle inequality doesn't hold. It seems that we require φ to be convex.

- **Answer:** Suppose that we know the corollary for p. It's actualy a good exercise to go from $\|f\|_1 \leq \|f\|_p$ to the proposition $\|f\|_q \leq \|f\|_q$. You can do this by a preprocessing. So it's true.

Question: How does this proof work?

- **Answer:** Apply the formula not to f, but to $|f|$.

We will not use all of these inequalities at once. I just wanted to collect them all at once so that you could see them. There are many consequences (I should say interpretations) of these inequalities in probability theory. For example

$$(\mathbb{E}|X|^p)^{1/p} \leq (\mathbb{E}|X|^q)^{1/q} \qquad \text{if } 0 < p \leq q \leq \infty$$

So, if you bound $(\mathbb{E}|X|^q)^{1/q}$, then you know all moments below it.

And so on. So, for every inequality above, we get an inequality about moments. So this was a break into analysis.

15 Limit Theorems

We will now go to probability theory. In fact, the first main result of probability theory: the limit theorems. The first is the Law of Large Numbers.

15.1 The Weak Law of Large Numbers

Laws of large numbers tell us: The frequency of successes in large trials. I expect that 50% I'll be hitting heads and 50% tails. The Law of Large Numbers will tell us the probability. In words, the theorem tells us with probability going to one, we have half heads and half tails. The question is what is the probability of this?

It will tell us that some frequencies will converge to $\frac{1}{2}$. We need some notion (modes) of convergence. There are many modes of convergence, and one of them is the convergence in probability.

Definition 15.1 (Convergence in probability)**.** *A sequence of random variables $X_1, X_2, \ldots$ converges in probability to a random variable X if*

$$\forall \epsilon > 0 \quad \mathbb{P}(|X_n - X| > c) \to 0 \text{ as } n \to \infty.$$

Then we write $X_n \xrightarrow{p} X$.

It's a very natural mode of convergence. We say that whatever tolerance we choose, we say that X_n will not deviate by much.

There is a stronger convergence: the convergence in L^p $(p > 0)$.

Definition 15.2. *A sequence of random variables $X_1, X_2, \ldots$ converges in L^p to a random variable X if*

$$\|X_n - X\|_p \to 0 \text{ as } n \to \infty.$$

In other words,

$$\mathbb{E}|X_n - X|^p \to 0 \text{ as } n \to \infty.$$

Why is this stronger?

Proposition 15.3. *$X_n \xrightarrow{L^p} X$ implies $X_n \xrightarrow{p} X$.*

Proof. We know some information about the expectation. The expectation of the difference is small. We need that the probability (or the tail of the difference) is small. If you know the expectation, we need to say something about the tail. What is that? Chebychev's Inequality[47]. We apply Chebychev's Inequality.

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \mathbb{P}(|X_n - X|^p \geq \epsilon^p) \leq \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p} \to 0.$$

□

Theorem 15.4 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}X_k = \mu$ for all k, and $\operatorname{var}(X_k) \leq c < \infty$, by some common[48] constant c. Consider $S_n = X_1 + X_2 + \cdots + X_n$.*

Then

$$\frac{S_n}{n} \to \mu$$

in probability (as $n \to \infty$).

So the average of n independent random variables converges to the common mean. Given all of this theory, the proof is simple.

[47] There's a confusion in the literature. Markov was a student of Chebychev.

[48] Don't be too concerned about the variance. The most important thing is their common mean.

Proof. Given the proposition, we will prove that $\frac{S_n}{n} \xrightarrow{L^2} \mu$, and that will be enough. So, is

$$\mathbb{E}\left|\frac{S_n}{n} - \mu\right|^2 \to 0? \tag{22}$$

What is μ? Note that the mean of $\frac{S_n}{n}$ is μ (by the linearity of expectation). So $\mathbb{E}(\frac{S_n}{n}) = \mu$.

Then (22) $= \operatorname{var}(\frac{S_n}{n}) = \frac{1}{n^2}\operatorname{var}(S_n) = \frac{1}{n^2}(\operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n)) \leq \frac{cn}{n^2} = \frac{c}{n} \to 0$. □

We actually even know the rate ($\frac{1}{n}$) of convergence.

NOVEMBER 21, 2007

I have graded your "midterm." It's good over all, but I need it back. Typically, the bargaining case for the final grade is based on overall performance. You can look at your grade on SisWEB. You are welcome to look at it for an hour or so.

We were studying the first major theorem of probability theory, which is the Weak Law of Large Numbers. See Theorem 15.4. This is a very simple theorem to prove, and it's very flexible.

Remark 15.5. *1. The convergence holds in L^2, rather than simply in probability (which is the expectation of $X_1 + X_2 + \cdots + X_n$ squared, which is just the variance of the variable $\frac{S_n}{n}$.*

2. The theorem holds for uncorrelated[49] X_k. The only place where we needed the uncorrelated was where the sums of the variances were the variance of the sum.

It has an interpretation in statistics (in statistical learning theory) as follows: So suppose X is a random variable with unknown mean (it maybe an outcome of the experiment); We want to estimate/compute the mean $\mu = \mathbb{E}X$. What access do we have to this random variable? We can make a finite number of experiments. We can make n experiments. Take n independent observations of X. These are independent random variables $X_1, \ldots, X_n$ with the same distribution as X. They are called "independent copies of X." We do not know the random variable X, because we do not know the probability space. So we take these many samples. The Weak Law of Large Numbers guarantees that the unknown mean μ of X can be well-approximated by $\frac{X_1+\cdots+X_n}{n}$, the known arithmetic mean of the sample. With any fixed tolerance ϵ, we can "observe" the mean. That's the power of the Weak Law of Large Numbers. For example, you are tossing a coin and you don't know if it's fair. You make an experiment a million times. If you see heads more than one-half, then you don't have a fair coin.

[49] rather than independent

One particular case for this was the very old Bernoulli Law of Large Numbers (1713). Bernoulli proved the statement for Bernoulli random variables:

Theorem 15.6 (Bernoulli Law of Large Numbers)**.** *Let S_n be a binomial random variable with parameters (n, p). S_n is the number of successes in n independent trials, where the probability of success in each trial is p.*

Then S_n is a sum of independent random variables, where X_k are each Bernoulli random variables with parameter p. Then this Law of Large Numbers applies, so

$$\frac{S_n}{n} \to p$$

in probability, as $n \to \infty$.

This ratio $\frac{S_n}{n}$ is the fraction of successes.

Bernoulli was one of the most well-known mathematical clan. They had around 12 Bernoulli spread around 200 years. It's hard to know now who did what. There were multiple Jacob's. As they moved around, they even changed names.

Example 15.7. *Suppose you want to know how many women and how many men are in a given population. Take a random sample of people from a given population. Then, with high probability, the sample proportion of women is very close to the true proportion in the population.*

Example 15.8. *Tossing a fair coin n times: We're looking at the number of heads. Then*

$$\mathbb{P}(0.49n \leq \text{ number of heads } \leq 0.51n) \to 1 \quad \text{as } n \to \infty.$$

Question: How large do sample sizes need to be?

- **Answer:** This is the beginning of the statistical learning theory. It's a big question to ask and answer how big the sample needs to be. To make a long story short, the next big theorem in probability is the Central Limit Theorem. This says that we can treat a random variable like a Gaussian.

  So, it will become eventually that you can't beat the variance, for one. For coins, we the variance can not be beat by $\sqrt{n}$.

  $$\mathbb{P}(n - t\sqrt{n} \leq \text{ number of heads } \leq n + t\sqrt{n}) \to 1 \quad \text{as } n \geq 1 - e^{-t^2/2}.$$

  We have a parameter t.

Now we will try to improve on the Weak Law of Large Numbers, because of its importance. We will try to get rid of the finite variance condition (because it's actually not needed). But then, we will throw into the conditions that the X_k are indentically distributed. In the exercise, you will get rid of this requirement. You need some way to control what happens at tails. Another way is something like $\mathbb{E}|X_k| = \mu$. Any kind of condition you put on the uniform control, you'll get a Law of Large Numbers. Everything can be weakened. So we'll try to get rid of the most crucial restriction, which is the variance.

Remark 15.9. *There exist random variables with finite mean and infinite variance. That was one of the exam problems ($\mathbb{E}X < \infty, \mathbb{E}X^2 = \infty$).*

If there is no finite variance, the only good thing that you probably have going for you is truncation. So, this is a method that goes back to Markov.

Theorem 15.10 (Markov's Truncation Method)**.** *Take X any random variable, with finite mean. Let $M > 1$ be a level (of truncation). A method for truncation: take*

$$X^{(M)} = \left\{ \begin{array}{ll} X, & \text{if } |X| \le M \\ 0, & \text{if } |X| > M \end{array} \right\} = X \cdot \mathbf{1}_{\{|X| \le M\}}.$$

Then $X^{(M)} \to X$ pointwise as $M \to \infty$, and $|X^{(M)}| \le M$ pointwise, and $|X^{(M)}| \le |X|$. In particular, you have not only the variance, but all moments. These two imply, by Dominated Convergence Theorem, that $\mathbb{E}X^{(M)} \to \mathbb{E}X$ as $M \to \infty$. Moreover,

$$\mathbb{E}|X^{(M)} - X| \to 0 \text{ as } M \to \infty. \tag{23}$$

An exercise is to prove this last line!

What's the difference in (23)? In $X - X^{(M)}$, we have just the part that has been truncated, so

$$X - X(M) = X \cdot \mathbf{1}_{\{|X| > M\}}.$$

So we just proved the corollary (which has nothing to do with truncation) that

Corollary 15.11. *For a random variable X with finite mean,*

$$\mathbb{E}|X| \cdot \mathbf{1}_{\{|X| > M\}} \to 0 \text{ as } M \to \infty.$$

X is finite mean, that ensures that there are "few" "big" values (but they are still big).

So, now we'll prove the theorem for independent identically distributed variables without the finite variance condition.

Proof of Theorem. Truncation. Let $M > 1$, consider

$$X_k^{(M)} := X_k \cdot \mathbf{1}_{\{|X_k| \le M\}}, \qquad S_n^{(M)} = X_1^{(M)} + \cdots + X_n^{(M)}.$$

Then

$$\operatorname{var}(X_k^{(M)}) = \mathbb{E}((X_k^{(M)})^2 - (\mathbb{E}X_k^{(M)})^2 \le M^2.$$

Then, we can apply the old version of the Law of Large Numbers with finite variance:

$$\frac{S_n^{(M)}}{n} \to \mathbb{E}X_1^{(M)}$$

in L^2 as $n \to \infty$. If L^2 norm goes to zero, then L^1 norm goes to zero. Thus, this convergence is also in L^1, therefore also in probability. What does this mean? It says

$$\mathbb{E}\left| \frac{S_n^{(M)}}{n} - \mathbb{E}X_1^{(M)} \right| \to 0 \text{ as } n \to \infty.$$

We haven't done anything yet. This is just old stuff. Nothing happened so far. Something will happen now. What will we need? We'll need to know this without the truncation? The truncated random variable is close to the non-truncated one. Perhaps same for the other term. Perhaps the integration will take care of the small difference.

Now we approximate $S_n^{(M)}$ by S_n and $X_1^{(M)}$ by X_1. By the triangle inequality,

$$\mathbb{E}\left|\frac{S_n}{n} - \mathbb{E}X_1\right| \leq \mathbb{E}\left|\frac{S_n^{(M)}}{n} - \mathbb{E}X_1^{(M)}\right| + \mathbb{E}\left|\frac{S_n - S_n^{(M)}}{n}\right| + \mathbb{E}|X_1 - X_1^{(M)}|.$$

Now, we estimate the first term:

$$\mathbb{E}\left|\frac{S_n - S_n^{(M)}}{n}\right| \leq \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}|X_k - X_k^{(n)}| = \mathbb{E}|X_1 - X_1^{(M)}|.$$

Thus

$$\mathbb{E}\left|\frac{S_n}{n} - |\mathbb{E}X_1\right| \leq 2\mathbb{E}|X_1 - X_1^{(M)}| + \mathbb{E}\left|\frac{S_n^{(M)}}{n} - \mathbb{E}X_1^{(M)}\right|$$

Now we let $n \to \infty$. Then

$$\limsup \mathbb{E}\left|\frac{S_n}{n} - \mathbb{E}X_1\right| \leq 2\mathbb{E}|X_1 - X_1^{(M)}|.$$

Let $M \to \infty$. Then $\mathbb{E}|X_1 - X_1^{(M)}| \to 0$. □

What if you know the variance is finite, but not uniformly bounded? What advantage can you get if you don't know that it's uniformly bounded? It's a quantatative statement.

NOVEMBER 26, 2007

15.2 Applications of the Weak Law of Large Numbers

We'll look at two elegant applications of the weak law of large numbers. The first application is called *Monte-Calro integration*. This is an application in basic scientific computing.

Maybe I'll just remind you what is the Weak Law of Large Numbers (WLLN).

Theorem 15.12. *If $X_1, \ldots, X_n$ are independent identically distributed random variables with the common mean $\mu < \infty$. Consider their sum $S_n = X_1 + \cdots + X_n$. Then $\frac{S_n}{n} \to \mu$ in probability.*

Recall that convergence in probability here means that

$$\mathbb{P}(|\frac{S_n}{n} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

for every $\epsilon > 0$.

15.2.1 Monte-Carlo Integration

So here's a problem that has no apparent connection to probability. Consider an integrable function

$$f : [0, 1] \to \mathbb{R}.$$

As you know, there are many functions that are not analytically integrable (no closed form formula for the integral). So the problem is numerically integratie f. That is, compute $\int_0^1 f(x)\ dx$. The use of $[0, 1]$ here is just arbitrary.

The first thing you might do is to approximate this by Riemann sums. We can take an equidistributed list of x-value points. The first naïve approach is to take equidistant points $x_1, \ldots, x_n$. At x_k, evaluate the function. Then hope that the integral can be well-approximated by the expected mean:

$$\int_0^1 f(x)\ dx \approx \frac{1}{n} \sum_{k=1}^{n} f(x_k).$$

This is a Riemann integral, because the mesh $\frac{1}{n}$ goes to zero. In very simple situations it works. But it almost always fails in the hard problems of scientific computing. Why is that? There is one problem with this approach.

1. We do not know anything about this function except that it is integrable. In particular, if f has a lot of structure, we may just pick the wrong points every time. Our function might just vanish at the points that we pick. We want a result that has "zero prior knowledge." The structure of f may lead to wrong information about the integral from the points $f(x_k)$. There are many such "oscillatory" functions (of sines and cosines).

The result we are going to prove allows you to integrate arbitrary functions f without prior knowledge. We will not take equ-distributed points. Why? Because the structure of the points might align with the structure of the function f. Instead, we use random points. Randomness goes against structure.

So, the solution to this problem is: instead of equidistributed points x_k, take random points. So let's do that.

We consider random variables $x_1, \ldots, x_n$ independent and uniformly distributed on $[0, 1]$. (When you see this for the first time, you look at this with a negative approach, because you worry about gaps. The gaps will be small, if n increases. The average gap will still be $\frac{1}{n}$.)

Then we hope that the arithmetic mean

$$I_n = \frac{1}{n} \sum_{k=1}^{n} f(x_k)$$

is a good approximation to the true mean $\int_0^1 f(x)\ dx$.

Well, this is just a reformulation of the Weak Law of Large Numbers. In words, each x_k is a uniform independent random variable.

Theorem 15.13 (Monte-Carlo Integration)**.** *$I_n \to \int_0^1 f(x)\ dx$ in probability.*

Monte-Carlo Integration is a random method. It is a randomized algorithm. So, in probability, we concern the I_n's. This is the theorem, and the proof is just a reinterpretation of the WLLN in this context.

Proof. We view f as a random variable[50] on the probability space $\Omega = [0,1]$ with Borel σ-algebra and the uniform probability measure.

Then $f \stackrel{d.}{=} f(x_k)$ for all k. (Why is the distribution the same? It says that the measure of $\{x : f(x) < a\}$ has to be the same as the probability $\mathbb{P}(f(x_k) < a\}$, but these are both uniform measures.)

Once we see this is true, then $\mathbb{E}f = \int_0^1 f(x)\ dx$. Then, the WLLN applied to the random variables $f(x_1), \ldots, f(x_n)$ completes the proof. □

This is one of the first randomized algorithms in scientific computation. The main strength of the method is that it does not require any knowledge about f. There may be no structure. It's just that f is integrable.

15.2.2 Weierstrass Approximation Theorem

Another application of the Weak Law of Large Numbers will be the classical analysis. This is one of the major results in analysis that says an arbitrary function can be well-approximated. We will prove this with probability theory.

The result is "Every continuous function f on $[0,1]$ can be well approximated by a polynomial f_n in the sup-norm." Namely, point-wise, f_n will be close to f, but also over the whole interval, this will be true as well.

Theorem 15.14 (Weierstrass Approximation Theorem)**.** *Let f be a continuous function on $[0,1]$. Then there exist polynomials f_n such that*

$$\|f - f_n\|_\infty = \sup_x |f(x) - f_n(x)| \to 0 \text{ as } n \to \infty.$$

This is a very useful theorem because it tells you that independent of the "wildness of the functions," we can think of arbitrary continuous functions as polynomials (if we can accept a little bit of error). There are many proofs, constructive and non-constructive:

One proof is due to Bernstein (1913) who gave an explicit formula for f_n. We will use probability theory to construct these Berstein polynomials f_n.

The first naïve approach is to interpolate f between equidistributed nodes x_k and compute the values $f(x_k)$ of f at these nodes. This is a classical interpolation method. There are two problems to the approach of interpolation:

1. The structure problem will still be there.
2. The second problem is that if you do this, the polynomials will be okay in the middle and then they will start to oscillate near the endpoints. There will be very huge oscillations of f_n near the edges. This is called *Runge's phenomenon*. While we can control the function at the nodes,

[50] after all, we view random variables as measureable functions, and f is a function!

we can no longer control the function between the nodes (because it is a polynomial). There is a way to break down this oscillation problem using random points.

3. But even then, there is a problem for an explicit formula for f_n. You will need to solve a system of linear equations given by the [unstable] Vandermonde matrix.

This is one of the very elegant proofs in analysis.

Proof of Weierstrass Approximation Theorem after Berstein. Think of this problem for a fixed x. Fix $x \in [0,1]$. We want to find a polynomial f_n such that for this particular x,

$$f_n(x) \approx f(x).$$

Where to look for such a polynomial? So this is the brilliant idea of Bernstein. We give up thinking of x as a fixed point, but we make a random variable that is tightly distributed around x. We will replace x by a cloud around x, the cloud being the values of the random variable around x. So, we will replace x by a random variable that's concentrated about x, and then compute its mean. So, how do we do that?

What's the simplest thing we can do? Take a binomial random variable. Consider S_n, the Binomial random variable with parameters n and x. Recall that the Binomial random variable is the sum of 0s or 1s, each taken with probability x. So, what is $\mathbb{E}S_n$? It is nx. So $\frac{S_n}{n} \to x$ in probability by the Weak Law of Large Numbers.

So, we'll replace x by this random variable. And then we'll see what happens. We compute $\mathbb{E}f(\frac{S_n}{n})$. This is a discrete random variable, so

$$\begin{aligned}\mathbb{E}f\left(\frac{S_n}{n}\right) &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\mathbb{P}(S_n = k)\\ &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\cdot\binom{n}{k}(1-x)^{n-k}x^k.\end{aligned}$$

What is this? This is a polynomial. Call this *Bernstein's polynomial* $f_n(x)$.

We have not proved the theorem yet. It is now highly plausible that $f_n \approx f$. Why? This $\frac{S_n}{n}$ is close to x. So is f of one is near f of the other? That's the heuristic reason for this to hold. The arguments are close to each other, but we do not know what happens once we apply the function?

We claim that $f_n \to f$ in the sup-norm. Fix $\epsilon > 0$. We want to show that

$$\exists\ n_0\ \forall\ n > n_0 : \left|\mathbb{E}f\left(\frac{S_n}{n}\right) - f(x)\right| < \epsilon \text{ for every } x \in [0,1]. \tag{24}$$

By WLLN, we know that $\frac{S_n}{n} \approx x$.

We use the (uniform) continuity of f, thus

$$\exists \delta > 0 \text{ such that } |x-y| < \delta \ \Rightarrow\ |f(x) - f(y)| < \epsilon.$$

We need to formulate the WLLN here a little bit. We consider the "good event"

$$A_n = \left\{ \left| \frac{S_n}{n} - x \right| < \delta \right\}.$$

By WLLN, $\mathbb{P}(A_n) \to 1$ as $n \to \infty$.

We don't have f convex, so we split into two parts (the good part and the bad part):

$$\begin{aligned} \mathbb{E}\left| f\left(\frac{S_n}{n}\right) - f(x) \right| &\le \mathbb{E}\left| f\left(\frac{S_n}{n}\right) - f(x) \right| \mathbf{1}_{A_n} + \left| \mathbb{E} f\left(\frac{S_n}{n}\right) - f(x) \right| \mathbf{1}_{A_n^c} \\ &= \text{Good} + \text{Bad}. \end{aligned}$$

On A_n, $\left|\frac{S_n}{n} - x\right| < \delta$. By continutity, $\left|f\left(\frac{S_n}{n}\right) - f(x)\right| < \epsilon$. So "Good" $\le \epsilon$.

On $[0,1]$, $|f(x) \le M$ (since f is bounded). Then, $\left|f\left(\frac{S_n}{n}\right) - f(x)\right| \le 2M$. So "Bad" $\le 2M \cdot \mathbb{P}(A_n^C) \to 0$ as $n \to \infty$. So "Bad" $\le \epsilon$ for all large n.

Thus Good + Bad $\le 2\epsilon$. This is not quite right. We indeed proved (24) for an individual x. □

NOVEMBER 28, 2007

You can use anything, but in the end, your final submission should only use results from class and the exercises. Also, I will be away next week, so no office hours, though the TA will have office hours.

We'll try to do more work on Kolmogorov's 0/1 Law. The "bad thing" is that we don't know if the probability is 0 or 1.

16 Borel-Cantelli Lemmas

This corresponds to Chapter 2 section 18 in the text. The Kolmogorov 0-1 Law implies that: if $A_1, A_2, \ldots$ are independent, then all the tail events have probability either 0 or 1. Recall, a tail event is independent of any first k (k finite) events A_i. One such is A_n occurs i.o. (infinitely often). So,

$$\mathbb{P}(A_n \text{ i.o.}) = 0 \text{ or } 1.$$

The Borel-Cantelli Lemmas allow us to decide which one we have, and the conditions are very simple.

Theorem 16.1 (Borel-Cantelli Lemmas)**.** *Let $A_1, A_2, \ldots$ be events.*

1. *If $\sum_k \mathbb{P}(A_k) < \infty$ then $\mathbb{P}(A_n \text{ i.o.}) = 0$.*
2. *Suppose $A_1, A_2, \ldots$ are independent. If $\sum_n \mathbb{P}(A_n) = \infty$ then $\mathbb{P}(A_n i.o.) = 1$.*

Remark 16.2. *In the second part, independence is needed. Just take one event A and let $A_n = A$ for all n.*

Corollary 16.3 (0-1 Law)**.** *Let $A_1, A_2, \ldots$ be independent events. Then*

$$\mathbb{P}(A_n \text{ } i.o.) = \begin{cases} 0 & \text{if } \sum_n \mathbb{P}(A_n) < \infty \\ 1 & \text{if } \sum_n \mathbb{P}(A_n) = \infty. \end{cases}$$

Of course, the Kolmogorov Law told you about more than this specific tail event, but here's a computation for this specific tail event. Now, the proof of the lemmas in Theorem 16.1:

Proof. 1. We can write the probability than any A_n occurs as an expectation:

$$\mathbb{P}(A_n) = \mathbb{E}\mathbf{1}_{A_n}.$$

Then

$$\sum_n \mathbb{P}(A_n) = \mathbb{E}\sum_n \mathbf{1}_{A_n}.$$

We should justify this step above with some limit theorem. We know that the LHS is $\sum_n \mathbb{E}\mathbf{1}_{A_n}$ and we can pull the expectation in front of the sum by Monotone Convergence Theorem. Notice that

$$\sum_n \mathbf{1}_{A_n}$$

is the number of events A_n that occur. Let's denote it by N. N is a random variable. This is close to what we actually need, because the conclusion of the lemma is about how many events occur. From the above argument, we know that N has finite mean:

$$\mathbb{E}N < \infty.$$

Therefore,

$$\mathbb{P}(N < \infty) = 1.$$

In other words, N is finite a.s. This is exactly the conclusion of part 1 of the lemma[51]. To restate and make our statement closer to the statement of the theorem, $\mathbb{P}(N = \infty) = 0$.

2. We need to use independence. What is the event that $\{A_n \text{ i.o.}\}$? This is

$$\{A_n \text{ } texti.o.\} = \limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_n.$$

For every event n, there is a further event that occurs. That causes the "infinitely often" to be satisfied. Since the sequence is decreasing, we can also write that this is

$$\lim_n \bigcup_{k \geq n} A_n.$$

[51] We go to random variables, introducing N, even though the original statement doesn't speak of random variables. N "counts" the number of events that occur.

We want to compute the probability. To show the probability of $\{A_n \text{ i.o.}\}$ is 1, it is necessary and sufficient to show that

$$\mathbb{P}\left(\bigcup_{k\geq n} A_k\right) \to 1 \text{ as } n \to \infty.$$

We will actually show that it is 1. We don't know about unions of indepnedent events, so we use De Morgan's law to make it an intersection:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k\geq n} A_k\right) &= 1 - \mathbb{P}\left(\bigcap_{k\geq n} A_k^c\right) \\
&= 1 - \prod_{k\geq n} \mathbb{P}(A_k^c) \text{ by independence and continuity} \\
&= 1 - \prod_{k\geq n} (1 - \mathbb{P}(A_k)) \\
&= 1,
\end{aligned}$$

because $\sum_{k\geq n} \mathbb{P}(A_k)$ diverges. Here, we used a fact from analysis:

$$\sum_k a_k = \infty \Longrightarrow \prod_k (1 - a_k) = 0.$$

This is an exercise for which you need to take logarithms.

□

16.1 Head runs

There are lots of applications of Borel-Cantelli Lemmas. One application that we will do in detail will be for Head runs.

Toss a coin infinitely many times. You seed heads or tails come up. Sooner or later, you will see HH in a row. Or 20 heads in a row. If you have a lot of patience, you will see 100 heads come up in a row. The question is how long is it going to take for you to see this 1000000 heads in a row? How often do the patterns appear in random structures? The Borel-Cantelli Lemmas are the right tool to study these questions.

We toss a fair coin. We want to observe "HHH$\cdots$H", and we want to study their lengths. Let $\ell(n)$ = the number of consecutive Hs starting from n^{th} toss. (So $\ell(n) = 0$ if T in n^{th} toss.)

An exercise is to show that

$$\mathbb{P}(\ell(n) \geq 100 \text{ i.o.}) = 1.$$

This you can do with a little analysis. The very interesting question is can you increase this to an arbitrary constant? How fast will this sequence of heads grow as you toss the coin. It will grow like $\log n$.

Theorem 16.4. *1. For every $\epsilon > 0$, $\mathbb{P}(\ell(n) \geq (1+\epsilon)\log_2 n \text{ i.o.}) = 0$.*

2. $\mathbb{P}(\ell(n) \geq \log_2 n \text{ i.o.}) = 1$.

Proof. 1. $\mathbb{P}(\ell(n) \geq r) = \frac{1}{2^r}$. Now we do this probability by Borel-Cantelli. Let's just substitute $(1+\epsilon)\log_2 n$.

$$\mathbb{P}(\ell(n) \geq (1+\epsilon)\log_2 n) = \frac{1}{n^{1+\epsilon}}.$$

This series converges, and by part one of Theorem 16.1, the proof is complete.

2. The second part is more interesting. It gives you a common technique in probability theory. Why is the second part hard? The problem is that the events $\{\ell(n) \geq \log_2 n\}$ are not independent, so we can't just apply the Borel-Cantelli lemma right away.

We can "prune" these events so that we're only talking about non-overlapping events. Let's denote $r(n) := \log_2 n$.

We look at $n_1 = 1$. Then we look at the next n "after" $\log n_1$, so $n_2 = n_1 + r(1)$. In general,

$$\begin{aligned} n_1 &= 1 \\ n_{k+1} &= n_k + r(n_k). \end{aligned}$$

Then, the events $A_k = \{\ell(n_k) \geq r(n_k)\}$ are independent. We should somehow "fix" this so that the recurrence starts > 1, simply because $\log 1 = 0$. Back to the lemma:

These events are independent because A_k^c involve non-overlapping indices. We just wanted to "separate" events somehow. So we have

$$\mathbb{P}(A_k) = \frac{1}{2^{r(n_k)}}.$$

Now, we just want to see that the sum of this series diverges. We'll do it in a second, but now there's no probability left. This is calculus now. So, we need that

$$\sum_{k=1}^{\infty} \frac{1}{2^{r(n_k)}}$$

is defined iteratively. So, we fill the gaps in the sum by doing this:

$$\sum_{k=1}^{\infty} \frac{1}{2^{r(n_k)}} = \sum_{k=1}^{\infty} \frac{1}{2^{r(n_k)}} \cdot \frac{n_{k+1} - n_k}{r(n_k)}.$$

Now, we can interpret this jump $n_{k+1} - n_k$ as a sum of 1s:

$$= \sum_{k=1}^{\infty} \sum_{n_k \leq n < n_{k+1}} \frac{1}{2^{r(n_k)} r(n_k)},$$

which is nice because it runs over all indices. This is greater than or equal to

$$\sum_{n=1}^{\infty} \frac{1}{2^{r(n)} r(n)}$$

because $r(n_k) \leq r(n)$, so this is equal to

$$\sum_{n=1}^{\infty} \frac{1}{n \log_2 n},$$

which diverges.

Look at how sharp this argument is. □

16.2 Monkey and the Typewriter

You can do this for a much bigger set of events. Instead of just H and T, you can consider the English alphabet (say 40 letters, to deal with punctuation as well). You have the same thing, just with $\log_{40}$.

Then, with enough time, the Monkey will produce War and Peace, and all versions of it. How much time will you need to spend? Exponential time.

A reflection of this in the literature: "Library of Babel" by J. L. Borges. I'll not include it in the final, but it's an interesting library that has all possible books.

NOVEMBER 30, 2007

So a couple of announcements: The final exam is posted, as I promised. Well, I promised to post in on Thursday, but I posted it Friday. You have two weeks to complete it. The midterm solutions are posted: some selected problems where you had difficulties, I guess. In particular the bonus problem, which was quite interesting. In the final: In problem 3, X_n should be X_k.

So the next week is the last week of classes, as you know. I'll be away. So Monday and Wednesday, there will be class. Another faculty will cover for me. But Friday we'll cancel class. So the last day of classes for you will be next Wednesday. There will be no office hours during the week I'm away, except for the TA. There are OHs today and the week after I'm away on Monday. I can not give you any help on the final. I can only clarify the concepts or definitions or the problems themselves. I will not give any hint, so individual work. So think of it as an in-class exam.

Today's topic is "Convergence almost surely", or almost sure convergence.

17 Almost Sure Convergence

One unfortunate fact of probability theory is that we don't have a unique concept of convergence, unlike in real analysis where we have one definition of a sequence of numbers to converge. If you go far enough, in infinite-dimensional spaces,

you'll have different notions of convergence. In operators, there are three notions of convergence.

Similarly in probability theory, there are different notions of convergence. One of which is convergence in probability. So a sequence of random variables $X_n \to X$ in probability if

$$\forall \epsilon > 0, \quad \mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as } n \to \infty.$$

So for every accuracy ϵ, if you're willing to sacrifice this, X_n is close to X (arbirtrarily). So, there are two inaccuracies (there is your tolerance ϵ and then there is this exceptional probability, the piece of probability that you can not control). The problem with this definition is that you do not know anything about the part itself. It may wiggle around. It (the area out of tolerance) can be different for different n.

So, the stronger notion of convergence is the convergence almost surely.

Definition 17.1. *$X_n \to X$ almost surely if $\mathbb{P}(X_n \to X) = 1$.*[52] *This is abbreviated $X_n \to X$ a.s.*

So, for almost all points in the probability space, the sequence converges to X. It's not immediately clear why this notion is stronger. We'll prove that.

A natural way to work with almost sure convergence is to say what it means that X_n does NOT converge almost surely. Recall that a sequence of real numbers $x_n \not\to x$ iff $\exists \epsilon > 0 : |x_n - x| > \epsilon$ infinitely often (for infinite sequence of n's). So there is a subsequence which is "far" from x. Keeping this in mind, we can write the following:

Proposition 17.2. *$X_n \to X$ a.s. iff*[53] *$\forall \epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon \text{ i.o.}) = 0$.*

So the fact that X_n is far from X infinitely often means that X_n does not converge to X. Of course we wrote it this way because we know about infinitely often, and Borel-Cantelli, etc.

The theorem that compares our notions of convergence:

Theorem 17.3 (Convergence a.s. and in probability)**.** *Let X_n and X be random variables.*

1. *$X_n \to X$ a.s. implies $X_n \to X$ in probability.*
2. *If $X_n \to X$ in probability, then there is a subsequence $X_{n_k} \to X$ almost surely.*

Let's first see, before proving this theorem, we will see why the convergence in probability (in an example) does not imply convergence almost surely:

Example 17.4. *You need to control almost every value of this random variable. For almost all $\omega \in \Omega$ fixed, you need convergence. Convergence is probability*

[52] $\mathbb{P}(\omega : X_n(\omega) \to X(\omega)$.
[53] the condition above does not happen

means you have a small exceptional set. We'll make the exceptional set "move", and so there will be a "swipe", a "blip" that "slides acrosss" that we'll notice.

Let $X_n = \mathbf{1}_{A_n}$ for some events A_n. What do we require of the sets? The probabilities of the A_n's must converge to zero. If $\mathbb{P}(A_n) \to 0$, then $X_n \to 0$ in probability. Indeed, the difference between X_n and zero is only on the events A_n. So we will require $\mathbb{P}(A_n) \to 0$.

We want to construct an example that does not converge almost surely to zero. What is the requirement for the sets to not converge almost surely to zero. Let's look at the reforumlation. The indicator must be different from zero infinitely often. This is iff a point is in the set i.o. So, if $\mathbb{P}(A_n \text{ i. o.}) > 0$, then $X_n \not\to 0$ a.s.

For the indicator functions, convergence in probability means the $\mathbb{P}(A_n) \to 0$.

Are there sets for which this is true? Let me give you a hint: Think of independent events. Then we have a criterion for $\mathbb{P}(A_n i.o.) > 0$. Then we have a criterion for this: Borel-Cantelli. You can take adjacent decreasing intervals, and have it "rotate back around." That will work.

For independent sets, $\mathbb{P}(A_n i.o.)$ is zero if the sum of the probabilities converge, and 1 if the sum of probabilities diverge, by Borel-Cantelli[54]*. So, we should pick events A_n such that $\mathbb{P}(A_n) \to 0$ but $\sum_n \mathbb{P}(A_n) = \infty$. Think harmonic series.*

So here's the proof of the theorem:

Proof.

1. Assume $X_n \to X$ a.s. Let $\epsilon > 0$. Then $0 = \mathbb{P}(|X_n - X| > \epsilon \text{ i.o.}) = \mathbb{P}(\limsup\{|X_n - X| > \epsilon\})$. We basically finish by the continuity theorem, which allows us to pull the lim sup in front by only decreasing the value. Thus, $\mathbb{P}(\limsup\{|X_n - X| > \epsilon\}) \geq \limsup \mathbb{P}(|X_n - X| > \epsilon)$. So, we proved that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$. Hence $X_n \to X$ in probability.

2. Assume $X_n \to X$ in probability. We want to find a subsequence on which it converges almost surely. Now we have to somehow think of why our previous example is no longer a counter example to our statement. If you take a subsequence, then you have the probabilities of the events converge to zero. You can take a sequence that decreases fast enough. So, we can find a subseries that converges. Maybe this is a way to prove this. So we pick a subsequence.

   Fix $\epsilon_k \to 0$. We can choose a subsequence (n_k) such that[55]

   $$\mathbb{P}(|X_n - X| > \epsilon_k) < 2^{-k} \text{ for } k = 1, 2, \ldots$$

   Since $\sum_k 2^{-k}$ converges, Borel-Cantelli Lemma says

   $$\mathbb{P}(|X_{n_k} - X| > \epsilon_k \text{ i.o.}) = 0.$$

   Hence, $X_{n_k} \to X$ almost surely.

□

[54] Actually, the corollary to Borel-Cantelli.

[55] we can use below comparison to the terms of any convergent series

It may be a little abstract in the first sight to see this theorem. Think about the example of indicator functions.

Question: Is this like in analysis, when we can come up with a series that converges?

- **Answer:** Well, sort of. These are different notions of convergence.

So the way I understand this theorem best is by characteristic functions. The more general version is the following corollary, which unifies the parts of the theorem:

Corollary 17.5. *$X_n \to X$ in probability iff every subsequence of X_n contains a further subsequence that converges a.s.*

Proof.
- ($\Rightarrow$). The subsequence converges, and by the theorem, it contains a subsequence that converges a.s.
- ($\Leftarrow$). Assume, on the contrary, that $X_n \not\to X$ in probability. Then, $\exists \epsilon > 0$ and a subsequence (n_k) such that[56] $\mathbb{P}(|X_{n_k} - X| > \epsilon) > \epsilon$. Formally, there should be a δ at the end, but you can just take the minimum of them.

  If this is true, then no subsequence of X_{n_k} converges in probability. Therefore, not almost surely (by part 1 of the theorem), a contradiction.

□

This may seem too abstract, but a concrete application of this[57]

Corollary 17.6. *If X_n is a monotone sequence (that is, $\forall \omega$, $X_n(\omega)$ either $\nearrow$ or $\searrow$), then $X_n \to X$ a.s. iff $X_n \to X$ in probability.*

Proof. If $X_n \to X$ in probability, then by part 2 of the theorem, there is a subsequence $X_{n_k} \to X$ a.s. But it's a monotone sequence. The whole sequence is sandwiched in there. By the monotonicity, the full sequence $X_n \to X$ a.s. □

So this is the end of the story about convergence almost surely and convergence in probability.

18 Strong Law of Large Numbers

A word about the final exam. The problems on the final exam are emphasisizng the laws of large numbers, since these are the most important applications of this class. You will need the strong law of large numbers, which you will need to solve the exam. This will get proved later. It's the same is the weak law, but only with almost sure convergence:

[56] such that you can not control X_{n_k} on the big part of the probability space.
[57] well, still kind of abstract

Theorem 18.1. *If $X_1, \ldots, X_n$ are i.i.d. (independent identically distributed) random numbers with finite mean μ. Then $S_n = X_1 + \cdot + X_n$ satisdy*

$$\frac{S_n}{n} \to \mu \text{ a.s.}$$

DECEMBER 3, 2007

Today we are going to have a look at the Strong Law of Large Numbers. Before, we looked at the Weak Law of Large Numbers. Let's compare. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with finite mean μ. Let $S_n := X_1 + \cdots + X_n$. The Weak Law of Large Numbers states that

$$\frac{S_n}{n} \to \mu \text{ in probability.}$$

On the other hand, the Strong Law of Large Numbers states that

$$\frac{S_n}{n} \to \mu \text{ a.s.}$$

But here, since (as we know) almost sure convergence implies convergence in probability: In this sense, we use this notation SLLN (the Strong Law of Large Numbers) is stronger than WLLN (the Weak Law of Large Numbers).

Today, we are going to prove SLLN under the assumption that the fourth moment is finite.

Theorem 18.2 (Strong Law of Large Numbers under the condition that the 4^{th} moment is finite)**.** *Let $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ, and assume $\mathbb{E}X_k^4 < \infty$. Then, $S_n = X_1 + X_2 + \cdots + X_n$ satisfies*

$$\frac{S_n}{n} \to \mu \text{ almost surely.}$$

In the next lecture, we'll drop the condition on fourth moments.

Proof. • Step 1. First, we can assume, without loss of generality, that the mean $\mu = 0$. Indeed, let $X_k' = X_k - \mu$. Then,

$$\frac{\sum_{k=1}^n (X_k - \mu)}{n} = \frac{S_n}{n} - \mu.$$

So assuming that $\mu = 0$, we prove that

$$\frac{S_n}{n} \to \mu = 0 \text{ almost surely.}$$

That is, we prove that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon \text{ i.o.}\right) = 0. \tag{25}$$

- Step 2. We consider the fourth moment of S_n. What is this? Let's make calculation:

$$\mathbb{E}S_n^4 = \mathbb{E}(\sum_{k=1}^{n} X_k)^4 \tag{26}$$

$$= \sum_{1\leq i,j,k,\ell\leq n} \mathbb{E}(X_i X_J X_k X_\ell) \tag{27}$$

We have n^4 terms, but

Claim 18.3. *Only $3n^2 - 2n$ are nonzero.*

Indeed, if i, j, k, ℓ are distinct, then by the independence,

$$\mathbb{E}X_i^3 X_j = \mathbb{E}X_i^3 \mathbb{E}X_j = 0$$

because we assumed that the mean value is 0. Similarly, we have

$$\mathbb{E}X_i^2 X_j = 0.$$

Of course, we have

$$\mathbb{E}X_i X_j X_k X_\ell = 0$$

as long as i, j, k, ℓ are different. By doing this calculation, we see that the only terms that are not zero are of the form:

$$\mathbb{E}X_k^4 \qquad \text{and } \mathbb{E}X_j^2 X_k^2.$$

Let's consider how many of them we have in the formula (27). Observe, we have n terms in $\mathbb{E}X_k^4$. To count the $\mathbb{E}X_j^2 X_k^2$, we choose 2 elements from a n-letter alphabet, so the combination is $\binom{n}{2}$.

Here, we have

$$\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{k=1}^{n} X_k\right)$$

Once we choose a j and a k, the number of ways to choose two j's and two k's is $\binom{4}{2}$ (we are picking one term in each factor in the expression above).

Since

$$\binom{n}{2}\binom{4}{2} = \frac{n(n-1)}{2}\frac{4\times 3}{2\times 1} = 3n(n-1),$$

we have $n + 3n(n-1) = 3n^2 - 2n$ nonzero terms.

- Step 3. Every nonzero term is bounded by $C = \mathbb{E}X_k^4$. This statement has meaning since we assumed that this value is finite.

By Cauchy-Schwarz Theorem,

$$\mathbb{E}X_j^2X_k^2 \leq (\mathbb{E}X_j^4)^{\frac{1}{2}}(\mathbb{E}X_k^4)\frac{1}{2} = C.$$

By using this argument, the term $\mathbb{E}X_j^2X_k^2$ is also bounded by C.

Therefore $\mathbb{E}S_n^4 \leq C \times (3n^2 - 2n) \leq 3Cn^2$.

- Step 4. Now we will use Chebychev's Inequality (we do have positive random variables) to complete the proof. By Chebychev's Inequality, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon\right) &= \mathbb{P}(S_n^4 > \epsilon^4 n^4) \\ &\leq \frac{\mathbb{E}S_n^4}{\epsilon^4 n^4} \text{ by Chebychev} \\ &\leq \frac{3C}{\epsilon^4 n^2} \text{ by Step 3}\end{aligned}$$

The series

$$\sum_{n=1}^{\infty} \frac{3C}{\epsilon^4 n^2}$$

connverges, so by the Borel-Cantelli Lemma,

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \epsilon \text{ i.o.}\right) = 0.$$

□

In the next lecture, we are going to prove this theorem dropping the assumption that $\mathbb{E}X_k^4 < \infty$.

Question: Why do we need the fourth moment, as opposed to say the second moment?

- **Answer:** The series in the proof won't converge.

For the rest of the lecture, I'll try to explain that this mean value $\mathbb{E}|X_k| < \infty$ must hold for this theorem.

Proposition 18.4 (Finite mean is necessary)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with infinite mean: $\mathbb{E}|X_k| = \infty$. Then $S_n = X_1 + X_2 + \cdots + X_n$ satisfies*

$$\mathbb{P}\left(\frac{S_n}{n} \textit{ converges to a finite number}\right) = 0.$$

Proof. An exercise: $\mathbb{E}|X_1| = \infty \Longrightarrow \sum_{n=0}^{\infty} \mathbb{P}(|X_1| > n)$ diverges. Rough picture: If $\mathbb{E}|X_1| < \infty$, $\mathbb{E}|X_1| = \int_0^{\infty} \mathbb{P}(|X_1| > x)\, dx \leq \sum_{n=0}^{\infty} \mathbb{P}(|X_1) > n)$, so we can't use this formula promptly.

Since X_k are identically distributed[58],

$$\mathbb{P}(|X_1| > n) = \mathbb{P}(|X_n| > n) \text{ for } n = 1, 2, \ldots$$

We have this condition:

$$\sum_{n=0}^{\infty} \mathbb{P}(|X_n| > n) = \infty.$$

By Borel-Cantelli Lemma,

$$\mathbb{P}(|X_n| > n \text{ i.o.}) = 1.$$

So, let's define two events:

$$A := \{|X_n| > n \text{ i.o.}\} \qquad (\text{Here, } \mathbb{P}(A) = 1)$$

and

$$B := \{\frac{S_n}{n} \text{ converges to a finite number}\}.$$

We prove that $A \cap B = \emptyset$. Why do we prove this statement? If this is true, then

$$\mathbb{P}(B) \leq \mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A) = 1 - 1 = 0.$$

To prove that $A \cap B = \emptyset$, suppose (for a contradiction) that $A \cap B \not\emptyset$.

Let's consider this value

$$\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{(n+1)S_n - n(S_n + X_{n+1}}{n(n+1)} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}.$$

Then for some event,

$$\frac{S_n}{n(n+1)} \to 0.$$

On the other hand,

$$\left|\frac{X_{n+1}}{n+1}\right| > 1 \text{ i.o.},$$

so when n is large enough,

$$\left|\frac{S_n}{n} - \frac{S_{n+1}}{n+1}\right| > \frac{2}{3} \text{ i.o.},$$

hence $\frac{S_n}{n}$ is not Cauchy. I.e., $\frac{S_n}{n}$ does not converge, which is a contradiction. So the intersection of A and B is actually empty. □

[58] Note the use of the same n in the RHS

In the next next lecture, we prove SLLN without assumptions on fourth moments.

DECEMBER 5, 2007

Today we are going to work on the strong law of large numbers without the assumption on fourth moments.

Theorem 18.5 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \dots$ be independent identically distributed random variables with mean value $\mathbb{E}|X_i| < \infty$ and $\mathbb{E}X_i = \mu$. Then, consider $S_n = X_1 + \cdots + X_n$, which satisfies*

$$\frac{S_n}{n} \to \mu \text{ almost surely.}$$

Proof.

- Step 0. We can assume that $X_i \geq 0$. Indeed, if we divide X_i

  $$X_i = X_i^+ + X_i^-$$

  into a positive part and a negative part, then $\mathbb{E}|X_i^+| < \infty$ and $\mathbb{E}|X_i^|- < \infty$.

- Step 1 (truncation). We define

  $$\overline{X_i} = \begin{cases} X_k & \text{if } X_k \leq k \\ 0 & X_k > k \end{cases}$$

  Then, we claim that $\mathbb{P}(\overline{X_k} \neq X_k \text{ i.o.}) = 0$. First, let's prove this claim.[59] We will show $\mathbb{P}(\overline{X_k} \neq X_k) = \mathbb{P}(X_k > k) = 0$. The series

  $$\sum_{k=1}^{\infty} \mathbb{P}(X_k > k) = \sum_{k=1}^{\infty} \mathbb{P}(X_1 > k) \leq \int_0^{\infty} \mathbb{P}(X_1 > x)\, dx = \mathbb{E}X_1 < +\infty.$$

  By Borel-Cantelli Theorem, the claim is true.

  If we prove that

  $$\frac{\overline{X_i} + \overline{X_2} + \cdots + \overline{X_n}}{n} \to \mu \text{ almost surely,}$$

  then

  $$\frac{X_i + X_2 + \cdots + X_n}{n} \to \mu$$

  almost surely. There is a homework exercise: Show that $\mathbb{E}\frac{\overline{X_1}+\cdots+\overline{X_n}}{n} \to \mu$.

  In the following, let's assume that $X_k \leq k$. (To get a precise proof, we have to interpret these numbers in terms of the truncated function. But for this lecture, let's just assume that this is true for all k).

[59] But, watch out for randomly-falling erasers!!!! ☺

- Step 2. We have a lemma.

  Lemma 18.6.
  $$\sum_{k=1}^{\infty} \frac{\text{var}(X_k)}{k^2} < +\infty.$$

  Proof.
  $$\begin{aligned}
  \text{var}(X_k) &= \mathbb{E}(X_k^2) - (\mathbb{E}X_k)^2 \\
  &\leq \mathbb{E}(X_k^2) \\
  &= \int_0^{\infty} \mathbb{P}(X_k \geq x)\ d(x^2) \\
  &= \int_0^{\infty} 2x\mathbb{P}(X_k \geq x)\ dx \\
  &= \int_0^{\infty} \mathbf{1}_{\{x \leq k\}} 2x\mathbb{P}(X_k \geq x)\ dx.
  \end{aligned}$$

  So let's calculate the summation
  $$\begin{aligned}
  \sum_{k=1}^{\infty} \frac{\text{var}(X_k)}{k^2} &\overset{\substack{\text{since the variables are} \\ \text{identically distributed}}}{=} \sum_{k=1}^{\infty} \int_0^{\infty} \frac{\mathbf{1}_{\{x \leq k\}}}{k^2} 2x\mathbb{P}(X_1 \geq x)\ dx \\
  &\overset{\text{Fubini}}{=} \int_0^{\infty} \sum_{k=1}^{\infty} \frac{\mathbf{1}_{\{x \leq k\}}}{k^2} 2x\mathbb{P}(X_1 \geq x)\ dx \\
  &= \int_0^{\infty} \sum_{k \geq x} \frac{1}{k^2} 2x\mathbb{P}(X_1 \geq x)\ dx
  \end{aligned}$$

  and as an exercise, show $\sum_{k \geq x} \frac{1}{k^2} \leq \int_x^{\infty} \frac{dt}{t^2} = \frac{1}{x}$. Hence
  $$\sum_{k=1}^{\infty} \frac{\text{var}(X_k)}{k^2} \leq \int_0^{\infty} 2\mathbb{P}(X_1 \geq x)\ dx = 2\ EX_1 < +\infty.$$

  □

- Step 3. (Control along a subsequence)

  Let $k(n)$, a subsequence of k, be α^n, where $\alpha > 1$. In this proof, let's assume that α^n is α^n is an integer. Then, we claim that:
  $$\frac{S_{k(n)}}{k(n)} \to \mu \text{ a.s.}$$

  For the proof of this claim, we also use the Borel-Cantelli Lemma. For

each $\epsilon > 0$, we compute

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k(n)}}{k(n)} - \mu\right| > \epsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}(|S_{k(n)} - \mathbb{E}S_{k(n)}| > \epsilon \cdot k(n)) \leq \sum_{n=1}^{\infty} \frac{\text{var}(S_{k(n)}}{(\epsilon k(n))^2}$$

by Chebychev's Inequality. Finally this can be written as

$$\frac{1}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{k(n)^2} \sum_{k=1}^{k(n)} \text{var}(X_k).$$

On the other hand, consider

$$\frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \text{var}(X_k) \sum_{n:k(n)\geq k} \frac{1}{k(n)^2}.$$

Here, we changed the order of summation, and we're not sure that these values are the same. To ensure that they are the same, we need to check for absolute convergence. We we check that the inner sum

$$\sum_{n:k(n)\geq k} \frac{1}{k(n)^2}.$$

This is equal to (by replacing $k(n)$ with α^n)

$$\sum_{n:\alpha^n \geq k} \frac{1}{\alpha^{2n}} \overset{\text{geometric progression}}{\leq} \frac{1}{(1-d^{-2})k^2}.$$

$$\frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \text{var}(X_k) \sum_{n:k(n)\geq k} \frac{1}{k(n)^2} \leq \frac{1}{\epsilon^2(1-d^{-2})} \sum_{k=1}^{\infty} \frac{\text{var}(X_k)}{k^2}$$

By Borel-Cantelli Lemma,

$$\mathbb{P}\left(\left|\frac{S_{k(n)}}{k(n)} - \mu\right| > \epsilon \text{ i.o.}\right) = 0$$

thus

$$\frac{S_{k(n)}}{k(n)} \to \mu \text{ almost surely.}$$

We have to rearrange terms to be careful.

- Step 4. (Filling gaps in the sequence).

  So, we sandwich. Want to know about all the ks. We want to know about all of the S_k's, so we estimate them

  $$S_{k(n)} \leq S_k \leq S_{k(n+1)}$$

Then, we have the following formula:

$$\frac{1}{\alpha} \cdot \frac{S_{k(n)}}{k(n)} = \frac{S_{k(n)}}{k(n+1)}$$

because we defined $k(n) = \alpha^n$. And

$$\frac{1}{\alpha} \cdot \frac{S_{k(n)}}{k(n)} = \frac{S_{k(n)}}{k(n+1)} \leq \frac{S_k}{k} \leq \frac{S_{k(n+1)}}{k(n)} \leq \alpha \cdot \frac{S_{k(n+1)}}{k(n+1)}$$

Thus

$$\frac{1}{\alpha} \cdot \frac{S_{k(n)}}{k(n)} \to \frac{\mu}{\alpha} \text{ almost surely.}$$

Hence,

$$\frac{\mu}{\alpha} \leq \liminf \frac{S_k}{k} \leq \limsup \frac{S_k}{k} \leq \alpha\mu$$

with probability one.

Since $\alpha > 1$ is arbitrary, let $\alpha \to 1$. Then, the limit

$$\lim \frac{S_k}{k}$$

exists, and is equal to μ almost surely.

□

If you have some question about this kind of thing $\overline{X_k}$ or $\overline{X}$, I have some memos. I also have some memo about the bound of integration. We can have discussion after.

Lecture on Friday was cancelled.

JANUARY 7, 2008

We are not televised this quarter. Welcome back, to the survivors of the past quarter, and the past storm. So this time, we are offering two quarters of probability theory. There is A and B, and no C. So what I thought to do is to cover the other major theorem of probability theory, the Central Limit Theorem (or rather Theorems). The second part will be the theory of martingales. Will leave out Brownian motion and Markov chains, which would normally be covered in 235C. There no instructor for 235C.

The Central Limit Theorem will have prerequisite characteristic function. So, we'll need the Fourier transform, so we'll need to know all basic operations with complex numbers. We will not rely on the whole undergraduate complex analysis class, if you have not taken one. But you will need to know how to manipulate with things like Euler's formula

$$e^{it} = \cos t + i \sin t.$$

We will need the first two or three weeks of undergraduate complex analysis.

The textbook is another painful problem for us. I don't want to push you to buy another book. We'll keep the current textbook. I will interpolate between the text by Gut and the text by Durrett. In Durrett, the lectures will mostly follow chapters 2 and 4, which is Central Limit Theorems and Martingales. There are similar chapters in Gut. You will not need to buy the Durrett text. It will be on reserve in Shields Library. If I assign homework based on some textbook, it will be from Gut. My lectures will stay between Durrett and Gut.

The second reason you do not want to buy the book is you will find a link to these notes. Find the link on my webpage.

The assessment is a bit different. 30% for HW, 30% for MT, and 40% for final. There will be no late homework at all. Instead, we'll drop one lowest homework. I ask you to put the HW on the table **before** class, or submit to my office door **sufficiently-before** class, because I will give to the TA directly after class. Please work together, but write individually.

My office hours are Monday 3:30-4:30. This is just for this class. If you are having troubles with this office hours, you can come at the same time on Friday, the priority will be for the calculus class. On the webpage, you'll find all of the information just spoken. The MT and final will be take-home again.

Let's do a little review and preview of the course. Recall random variables and basic quantities associated with random variables. Let X be a random variable[60]. There are two measures that tell us much about a random variable. We'll denote the *mean* $\mathbb{E}X = \mu$, and the *variance* $\operatorname{var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \sigma^2$. What variance measures is the distance from X to its mean. The square is to ensure positivity.

The mean scales just fine. If you multiply X by a, the mean is $a\mu$. For variance, you'll pull out a square: a^2. This is not as nice, so sometimes the variance is replaced by the *standard deviation* of X, which is just $\sqrt{\operatorname{var}(X)} = \sigma$. So, we have these two numbers μ and σ.

We prefer to work with random variables with $\mu = 0$ and $\sigma = 1$. This is called standard normal, and we say that *normalization* is the process of taking a random variable X and converting it to this form. Take X and consider instead

$$\frac{X - \mu}{\sigma}.$$

This is a random variable with mean 0 and variance 1. Still remember this?

Most of the time in probability theory, we work with sums of independent random variables instead of just one random variable. So we have X_1, X_2, and so on. Let these be independent, identically distributed random variables with mean μ and variance σ^2. This is a typically object that we study. Actually, we consider their sum $S_n = X_1 + \cdots + X_n$. All of these random variables, they are copies of some random variable X, and we "have access" to their values. For instance, we can record the results of coin flips. By linearity of expectation,

$$\mathbb{E}S_n = n\mu,$$

[60] r.v., for random variable, not my initials!

and the variance is (in general) **not** linear. The variance **is** linear when the random variables are independent, so

$$\text{var}(S_n) = n\sigma^2,$$

so

$$\text{st. dev}(S_n) = \sqrt{n} \cdot \sigma.$$

Now this innocent computation is very powerful, actually. It shows that the mean of the sum, the typical value, is linear in n, but the distance to the sum is much smaller (being $\sqrt{n}$). The amount it deviates about its center shrinks. Most of the time, it stays within a small interval $\sqrt{n} \cdot \sigma$. The bigger the n, the smaller this interval around its mean. This is called a *"Concentration phenomenon"*. This can be stated (it's half the proof of) the Weak Law of Large Numbers.

We apply Chebychev's Inequality, that tells us what's the probability that a random variable is within some quantity of its mean:

$$\mathbb{P}(|S_n - n\mu| > t \cdot \sqrt{n}\sigma) \leq \frac{1}{t^2}.$$

Now, for the Law of Large Numbers, we need something like ϵ, so let T be s.t. $t \cdot \sqrt{n}\sigma = \epsilon n$. Then, $t = \frac{\epsilon\sqrt{n}}{\sigma}$, so then we rewrite Chebychev's inequality as

$$\mathbb{P}(|S_n - n\mu| > \epsilon n) \leq \frac{\sigma^2}{\epsilon^2 n}$$

Maybe we divide by n,

$$\mathbb{P}(|\frac{S_n}{n} - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2 n}$$

So as n goes to ∞, this quantity gets small. This is called a *deviation inequality*, because it measures how a random variable S_n/n deviates from its mean. This implies the Weak Law of Large Numbers. This actually is the WLLN. That is, for every ϵ,

$$\mathbb{P}(|\frac{S_n}{n} - \mu| > \epsilon) \to 0$$

as $n \to \infty$. So $\frac{S_n}{n} \to \mu$ in probability.

So I just wanted to show what is the power of the fact given in our simple concentration at the beginning. So, this was a little review of what we did. Now, we did it in the last lectures under a finite variance condition. In the last lectures, the variance condition was dropped. But most of the time, you're fine with this.

Now, what are we going to do next? Central Limit Theorems. Here's a preview of Central Limit Theorems.

19 Central Limit Theorems

We already know that S_n has mean $n\mu$ and variance $n\sigma^2$. This alone implies WLLN. This alone implies concentration. We only used the nature of S_n (that

it is the sum of independent random variables) once. What's most remarkable (and this is what we'll show) is that the Central Limit Theorems (CLTs) allow us not only that S_n concentrates around its mean, but that it's very close to a standard normal distribution. So in some sense, $S_n \approx N(N\mu, n\sigma^2)$, a Gaussian (or normal) random variable with mean $n\mu$ and variance $n\sigma^2$. So, the keyword here is "normal". If you've seen this for the first time, this is counter-intuitive in some sense. If you start with any random variables, once you sum them up, you will get the same distribution. In physics, this is called "universality". The large-scale behavior should be the same.

More precisely, what do we mean? We need first, to normalize. So, we make S_n into a random variable with zero mean and variance one. So,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \to N(0,1)$$

converges to the standard normal variable as $n \to \infty$ ("in distribution"). We'll talk about why this is not one of the standard convergences.

Apart from just curiosity of this phenomenon, this is very useful in applications to deviation inequalities. What does this new information about convergence to the *normal* random variables say. Let's try to see what is the consequence of this.

Now we know that

$$\mathbb{P}(|\frac{S_n - n\mu}{\sqrt{n}\sigma}| > t) \approx \mathbb{P}(|g| > t),$$

where g is $N(0,1)$. So, this right hand side can just be computed. We have a density for g. So the right side is

$$2 \cdot \frac{1}{2\pi} \int_t^\infty e^{-x^2/2}\, dx.$$

If you ignore all of the constants in front, then this is

$$\approx e^{-t^2/2},$$

which is Proposition 6.9. Let t be such that $t\sqrt{n}\sigma = \epsilon n$. So $t = \frac{\epsilon\sqrt{n}}{\sigma}$, as before. So what we get is

$$\mathbb{P}(|S_n - n\mu| > \epsilon n) \sim e^{-\epsilon^2 n/2\sigma^2}.$$

Maybe we'll also divide by n on both sides to turn it into the form of WLLN:

$$\mathbb{P}(|\frac{S_n}{n} - \mu| > \epsilon) \sim e^{-\epsilon^2 n/2\sigma^2},$$

and this is our new deviation inequality. So let's compare. What's the difference of this deviation inequality to the previous one? Let's ignore everything except n. What we see in the previous statement is linear in n. The one here is much stronger since it is **exponential** in n, much stronger than inverse-linear.

This should not be surprising given the CLT, because the Gaussian decays at an exponential rate. This is just one application of the CLT. There are many more applications.

We will prove the CLT for Bernoulli random variables, actually for the flip of a coin. It will be useful if you get your hands on a very specific case. So we'll do this central limit theorem for Bernoulli random variables first. It is called the De Moivre-Laplace Central Limit Theorem. This is a section that will take us about a lecture.

19.1 CLT for independent Bernoulli random variables X_n

To be specific, we'll pick $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = \frac{1}{2}$. Therefore, the sum $S_n = X_1 + \cdots + X_n$ is a Binomial random variable with parameters $(n, \frac{1}{2})$, the number of heads in n coin tosses.

It was no surprise that De Moivre was interested in Bernoulli random variables. He was a gambler and would sell knowledge about gambling. So, it was important for him. So this is our situation and how can we estimate the sum of the Bernoulli random variables? What is the probability that S_n is equal to some number, say k? What is $\mathbb{P}(S_n = k)$? There are $\binom{n}{k}$ was to choose the position of the heads. The probability of that specific arrangement is $\frac{1}{2^n}$. So,

$$\mathbb{P}(S_n = k) = \binom{n}{k} 2^{-n}, \qquad k = 0, 1, \ldots, n.$$

The sum of all probabilities is one:

$$\sum_{k=0}^{n} \mathbb{P}(S_n = k) = 1.$$

So, the corollary, which doesn't have to do with probability, but rather combinatorics is:

Corollary 19.1. $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$.

So, what is the smallest binomial? For $k = 0$ or n. Now, what we are trying to prove in the CLT is something like "the Binomials form a bell-shaped curve." So the biggest binomial is the middle binomial $\binom{n}{n/2}$. This agrees, in principle, with the LLN (this is not a mathematical fact), that S_n should be concentrated around its mean $\frac{n}{2}$. This is philosophical. So we'll prove this next time. We'll do an asymptotic analysis of binomials.

JANUARY 9, 2008

I forgot to mention last time that the TA will also have office hours. Tuesday 1-2. This info is also on my webpage.

We have started De Moivre-Laplace Central Limit Theorem. So in general, when you see the Central Limit Theorem, there is a random variable that is

approximated by some normal random variable. We'll start with the Binomial distribution. The setting is as follows:

Let $X_1, X_2, \dots$ be independent Bernoulli random variables. For simplicity, we will only cover the case where the probabilities are equal, namely $\mathbb{P}(X_n = 0) = \mathbb{P}(X_n = 1) = \frac{1}{2}$. This is, for instance, the coin toss. We consider their sum $S_n = X_1 + \cdots + X_n$. This shows you the number of heads. This is a Binomial random variable with parameters $(n, \frac{1}{2})$.

Before we even state the theorem, I'd like to cover it. Our goal will be to find a useful approximation of the distribution of S_n. We know the mean and the variance. We know that $\mathbb{E}S_n = \frac{n}{2}$ by linearity. Similarly, the variance $\operatorname{var}(X_n) = \frac{1}{4}$, so $\operatorname{var}(S_n) = \frac{n}{4}$. Here, we use independence. So $\operatorname{st.dev}(S_n) = \frac{\sqrt{n}}{2}$. So the mean is linear in n, but the concentration winds tightly. Our goal is to do something stronger than this. We want to approximate the whole distribution of S_n.

How do we compute the distribution of S_n? Let's do it in a straightforward way, without any approximation. What's the probability of k heads? $\mathbb{P}(S_n = k) = \binom{n}{k} \cdot 2^{-n}$, where $k = 0, 1, \dots, n$. This is a distribution. Sometimes you're satisfied with this formula. But, in practice, we want something more. We usually ask about ranges of successes. We don't ask, out of 100 experiments, what are the chance over 54 successes? We want some range (like 50-60) successes. Summing up binomials is not an easy task. So, we want a useful asymptotic formula for the binomials $\binom{n}{k}$. That's the first step. Once we do this, it will become evident how to sum.

The binomials have some kind of pattern to them. The middle binomial is the largest. Then they settle down from the middle in some sort of "smooth" way. The curve they make is a normal curve (that's the content of the theorem). Here, we have a discrete situation, but in the limit we have a curve. Hopefully our asymptotics will discover a bell-shaped curve.

Let $k = tn$ for some $t \in (0, 1)$. We think of k as a proportion of n. Why do we want to do this? We think of t as about $\frac{1}{2}$ when we are "in the middle". What do we know about the binomial? Not much. We know

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

and that's about it. Factorials are not fun to add. We use Stirling's Approximation: $n! \sim \sqrt{2\pi n}(n/e)^n$. So $n!$ is about n^n up to a correction. Whenever we say that $a_n \sim b_n$ (are asymptotically equivalent) means that $a_n/b_n \to 1$ as $n \to \infty$.

We know other formulas for more preciseness. We'll need

$$n! = \sqrt{2\pi n}(n/e)^n e^{\lambda n},$$

where

$$\frac{1}{12n+1} \leq \lambda_n \leq \frac{1}{12n}.$$

You can find this on Wikipedia.

We'll use the Stirling Approximation. Let's work term-by-term. Then,

$$\binom{n}{k} \sim \sqrt{\frac{2\pi n}{2\pi k \cdot 2\pi(n-k)}} \cdot \frac{e^{-n}}{e^{-k}e^{-(n-k)}} \cdot \frac{n^n}{k^k(n-k)^{n-k}}$$

As a side computation, we note that $\frac{n}{k(n-k)} = \frac{1}{t(1-t)n}$. We'll use this for the first piece above. The second piece cancels out completely. For the third term, we ma write n^n as $n^k n^{n-k}$. So, the third term $\frac{n^n}{k^k(n-k)^{n-k}}$ will be $\frac{n^k n^{n-k}}{k^k(n-k)^{n-k}} = \frac{1}{t^k(1-t)^{n-k}} = \frac{1}{[t^t(1-t)^{1-t}]^n}$.

So, the bottom line is

$$\binom{n}{k} \sim \frac{1}{\sqrt{2\pi nt(1-t)}} \cdot 1 \cdot \frac{1}{[t^t(1-t)^{1-t}]^n}.$$

We'll take logarithm of this.

$$\binom{n}{k} \sim \frac{1}{\sqrt{2\pi nt(1-t)}} \cdot \exp(nH(t)), \tag{28}$$

where $H(t) = -t\log t - (1-t)\log(1-t)$. On purpose, I put the minus sign in front so to take care of the negative values of logarithm. This function H is called the Entropy function. It appears in physics and in information theory. It's a measure of the complexity of a source that shoots out letters. It's a very remarkable function, defined on $[0,1]$. The function is symmetric about $\frac{1}{2}$.

[Function H is graphed.]

It is a positive function. The maximum is attained at $\frac{1}{2}$, with value $\ln 2$.

Once we note that H is explicit (depends only on t), we can do something. We can estimate the middle binomial $\binom{n}{n/2}$ right away:

$$\binom{n}{n/2} \sim \sqrt{\frac{2}{\pi n}} \cdot 2^n.$$

So, the corollary to this is

Corollary 19.2. $\mathbb{P}(S_n = \frac{n}{2}) \sim \sqrt{\frac{2}{\pi n}}$

because we already have the 2^n cancels. So, it is unlikely that you have exactly half heads. But it's not that bad actually, because this does not decay exponentially. The middle binomial takes a big chunk of the sum (up to 1).

This approximation to the binomial given in (28) is actually very useful in other places.

We want the decay from the middle binomial: That is, what is $\binom{n}{k} \sim$? for $k = tn$ and $t = \frac{1}{2} + o(1)$. We know what happens to H at $\frac{1}{2}$. We can use Taylor's expansion (make this an exercise) around $t_0 = \frac{1}{2}$ to show that $H(t) \sim \ln 2 - 2(t-\frac{1}{2})^2$ for $t = \frac{1}{2} + o(1)$.

Question: Are you using ln and log differently?

- **Answer:** We are not distinguishing: All logarithms today are natural.

Then, for such t, we can write the asymptotics very easily. We use the formula (28) very easily.

$$\begin{aligned}\binom{n}{tn} &\sim \sqrt{\frac{2}{\pi n}} \cdot \exp(n \ln 2 - 2(t - \frac{1}{2})^2) \\ &= \sqrt{\frac{2}{\pi n}} \exp(-2n(t - \frac{1}{2})^2) \cdot 2^n.\end{aligned}$$

What does the formula here teach us? Once we're off the middle binomial, there is an exponential decay. Most of the mass is concentrated at the middle binomial. So there is an exponential decay off $t_0 = \frac{1}{2}$.

Now, what is our t here? The Central Limit Theorems talk about normalized random variables. So, our t will have this form: We will use this for $\binom{n}{k}$, where k has the form

$$k = \frac{n}{2} + x\frac{\sqrt{n}}{2}$$

that is, k is the mean plus some multiple of the standard deviation. This will be our tn. Then,

$$t = \frac{1}{2} + \frac{x}{2\sqrt{n}}.$$

Thus $(t - \frac{1}{2})^2 = (\frac{x}{2\sqrt{n}})^2 = \frac{x^2}{4n}$. Multiplying by $2n$, we get $2n(t - \frac{1}{2})^2 = 2n(\frac{x}{2\sqrt{n}})^2 = \frac{x^2}{2}$. So

$$\binom{n}{k} \sim \sqrt{\frac{2}{\pi n}} \cdot e^{-x^2/2} \cdot 2^n.$$

Our k was a little off the mean, and now we see the decay as an exponential, as a Bell-shaped curve. So, we just proved the Local Limit Theorem.

Theorem 19.3 (Local Limit Theorem)**.** *Let S_n be Binomial $(n, \frac{1}{2})$. Then,*

$$\mathbb{P}\left(\frac{S_n - n/2}{\sqrt{n}/2} = x\right) \sim \sqrt{\frac{2}{\pi n}} \cdot e^{-x^2/2}.$$

There's only one step to the true Central Limit Theorem. Here, we're talking about exact numbers of successes. The true theorem will talk about ranges. Now we go toward the true limit theorem. So we need to consider the range of the values

$$\mathbb{P}(a \leq \frac{S_n - n/2}{\sqrt{n}/2} \leq b).$$

What's that?

$$p := \mathbb{P}(a \leq \frac{S_n - n/2}{\sqrt{n}/2} \leq b) = \sum_{a \leq x_k \leq b} \mathbb{P}(\frac{S_n - n/2}{\sqrt{n}/2} = x_k),$$

where $x_k = \frac{k-n/2}{\sqrt{n}/2}$. This is an arithmetic progression. Its step (the mesh) is

$$\Delta x = x_k - x_{k-1} = \frac{2}{\sqrt{n}}.$$

When $p \sim \sum_{a \le x_k \le b} \sqrt{\frac{2}{\pi n}} e^{-x_k^2/2}$. by the Local Limit Theorem. We have a small mesh, we can replace this by an integral.

$$p \sim \sum_{a \le x_k \le b} \Delta x \cdot \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \sim \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

So we proved the De Moivre-Laplace Central Limit Theorem.

Theorem 19.4 (De Moivre-Laplace Central Limit Theorem)**.** *Let S_n be Binomial random variable with parameters $(n, \frac{1}{2})$. Then, for every $a < b$, we have*

$$\mathbb{P}\left(a \le \frac{S_n - n/2}{\sqrt{n}/2} \le b\right) \to \mathbb{P}(a \le g \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx,$$

where g is $N(0,1)$.

So there are three exercises:

1. Convergence is uniform in a and b.
2. The CLT holds for arbitrary Bernoulli random variables, that is for arbitrary p rather than $\frac{1}{2}$, but it's not uniform anymore. (It will be Poisson.)
3. Do the rigorous analysis of "$\sim$". How does $\sim$ play with our integral? You want to be extra-careful.

All of these are covered in Shiryaev (Probability).

Finally, I'll show you two slides. The first is an experiment for Binomial distribution. You see minimal number of trials is 10, with probability $p = \frac{1}{5}$. Here, you already see the CLT manifest itself. The second is a Bell-shaped curve. Note $\mathbb{P}(a \le g \le b)$ is almost 1 if $a < 0$ and $b > 0$. If $a = -2$ and $b = 2$, then $\mathbb{P} = 0.96$.

JANUARY 11, 2008

The HW for next week is already posted. Due Wednesday.

20 Convergence in Distribution

We're starting a topic that's called the convergence in distribution. It roughly corresponds to section 2.2 in Durrett's book. We just finished this De Moivre-Laplace CLT. Let's look back and try to see... It's a theorem about convergence.

The Binomial distribution, properly normalized, converges to the normal distribution. But in what sense is this convergence? We studied at least two notions of convergence.

Is this convergence almost sure convergence? Why is this not a.s. convergence? S_n is a binomial random variable. The CLT says that $\frac{S_n - n/2}{\sqrt{n}/2} \to g$, where g is $N(0,1)$. Almost sure convergence would say that this convergence happens for all $\omega \in \Omega$. We can define Ω to be different in each case. In the left, it can be defined to be a discrete space. So, the convergence in the CLT is not almost sure convergence. For the same reason, it's not the convergence in probability. S_n and g may (and have to) be defined on different probability spaces.

The convergence in CLT is the convergence in *distribution*. The alternate name (which Durrett follows) is *weak convergence*. It is the weakest notion of convergence of random variables[61].

Definition 20.1. *Let $X_1, X_2, \ldots$ be random variables with distribution functions[62] $F_1, F_2, \ldots$, and let X be a random variable with distribution function F. We say that*

$$X_n \to X \text{ in distribution as } n \to \infty$$

(or sometimes write $X_n \xrightarrow{d.} X$) if

$$F_n(x) \to F(x)$$

for all x that are points of continuity of F.

This "points of continuity" for F is trivial for Gaussians.

What is the advantage of this notion? It is defined only using the distribution functions, not the values of the random variables themselves. So this convergence only depends on the distributions of $X, X_1, X_2, \ldots$, not on their actual values (unlike in a.s. and in probability convergences).

We have the discrete probability space. If we shuffle around the values of X, it won't change convergence in distribution. Because it relies on the distribution functions, sometimes people write

$$F_n \to F \text{ weakly.}$$

Or even more, people would write[63]

$$\mathbb{P}_n \to \mathbb{P} \text{ weakly.}$$

Sometimes $\mathcal{L}$ is used for the distribution, so this may get used as well. The literature is very diverse.

There is a deep connection between weak convergence in Banach spaces and the weak convergence here. You can consider the Banach space X of all

[61] weakest that we will study.
[62] Recall distribution functions $F_X(x) = \mathbb{P}(X \le x)$.
[63] Recall $\mathbb{P}_n(A) = \mathbb{P}(X \in A)$

measures μ_n on $(\Omega, \mathcal{F})$. They are not required to be probability measures (or even positive). We need this to be able to consider them as vectors. Then the linear functionals f on X:

$$f(\mu_n) = \int_\Omega f \, d\mu_n$$

are functions on Ω. (There's some notion of topology here.)

This is certainly an action of linear functionals. Then $\mu_n \to \mu$ weak means $f(\mu_n) \to f(\mu) \ \forall \ f$. In particular, if $\Omega = \mathbb{R}$, and $f = \mathbf{1}_{(-\infty, x]}$,

$$f(\mu_n) = \mu_n(-\infty, x]$$

and its limit will need to be

$$f(\mu) = \mu(-\infty, x].$$

In our situation, this matches $\mathbb{P}_n \to \mathbb{P}$ weakly is the same as to say $\mathbb{P}_n(-\infty, x] \to \mathbb{P}(-\infty, x]$. From this we read directly $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$.

So the name of weak convergence comes from Hilbert and Banach space theory. This is not a rigorous argument, but just an indication of why we have the name.

In a more familiar form, and equivalent definition of the weak convergence is: $X_n \to X$ is distribution if[64] $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ for all x such that $\mathbb{P}(X = x) = 0$.

Also equivalently, when you have a one-sided inequality, you can get a two-sided inequality: $\mathbb{P}(a \leq X_n \leq b) \to \mathbb{P}(a \leq X \leq b)$ for all $a \leq b$ such that $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$.

Examples:

1. One example is of course the De Moivre-Laplace CLT. It says (in our situation), the convergence

   $$\frac{S_n - n/2}{\sqrt{n}/2} \to g \text{ in distribution.}$$

   In words, we say that "the Binomial distribution $(n, \frac{1}{2})$ properly normalized converges weakly to the standard normal distribution."

2. Even before we clarify the implications between the difference convergences, we want to check the convergence of distribution for the Law of Large Numbers.

   In the SLLN,

   $$\frac{S_n}{n} \to \frac{1}{2} \text{ almost surely (thus in probability as well)}$$

[64] well, if and only if

Though we will have a theorem later, let's just directly check that the convergence in distribution holds.

Convergence in probability is the statement

$$\mathbb{P}(|\frac{S_n}{n} - \frac{1}{2}| > \epsilon) \to \epsilon) \to 0 \text{ as } n \to \infty.$$

We're interested in

$$\mathbb{P}(\frac{S_n}{n} \leq x).$$

What does it converge to? $x = \frac{1}{2}$ will be a singularity point.

$$F_n(x) = \mathbb{P}(\frac{S_n}{n} \leq x) = \begin{cases} 0, & x < \frac{1}{2} \\ \frac{1}{2}, & x = \frac{1}{2} \\ 1, & x > \frac{1}{2}. \end{cases}$$

This carries over to when X_i is a symmetric random variable.

When we say that it should converge to $\frac{1}{2}$ is distribution, we think of it as a constant function. So

$$F(x) = \mathbb{P}(\frac{1}{2} \leq x) = \begin{cases} 0, & x < \frac{1}{2} \\ 1, & x \geq \frac{1}{2}. \end{cases}$$

So, let's plot these two functions $\lim F_n$ and F. Indeed, F_n converges to F except at $x = \frac{1}{2}$, which is a point of discontinuity of F.

We have to care about the points of discontinuity. Each F_n was a distribution function, but the limit of them is not a distribution function (because it is not right-continuous).

I may also mention one exam exercise.

Lemma 20.2 (Uniqueness)**.** *If $X_n \to X$ in distribution, then the distribution of X is uniquely defined.*

In other words, the limit is unique. I'll leave it as an exercise.

JANUARY 14, 2008

One announcement about the HW. Regarding Proposition 6.9. There is a a problem in the statement. The correct form of this proposition is:

Proposition 20.3.

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) e^{-x^2/2} \leq \int_x^\infty e^{-y^2/2}\, dy \leq \frac{1}{x} e^{-x^2/2}$$

for all $x > 0$.

Actually, one only uses this for x large. It only makes sense for $x > 1$.

In the exercise, x is large. In the exercise, n is large. I'll put this correction online.

Today, we'll clarify the connections of the different types of convergence. We know at least three main types of convergence of random variables.

Recall that $X_n \to X$:

- almost surely if $\mathbb{P}(X_n \to X) = 1$
- in probability if $\forall \epsilon > 0, \mathbb{P}(|X_n - X| > \epsilon) \to 0$
- in distribution if $F_n(x) \to F(x)$ for all points of continuity of F. Equivalently, $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ whenever $\mathbb{P}(X = x) = 0$.

We've seen almost sure convergence in SLLN. We've seen convergence in probability in WLLN. We've seen convergence in distribution in the CLT. So, now we'd like to clarify what implies what. Almost sure convergence implies convergence in probability. The converse does not hold, though there is some statement about subsequences. We'll show that convergence in probability implies convergence in distribution.

Proposition 20.4. *$X_n \to X$ in probability $\implies$ $X_n \to X$ in distribution.*

We'll do it directly. We know that $\mathbb{P}(|X_n - X| > \epsilon) \to 0$.

Proof. Assume that we know that $X_n \leq x$. What can we say about X itself? Then either $X \leq x+\epsilon$ or $|X_n - X| > \epsilon$. The probability of the event $|X_n - X| > \epsilon$. But we know that this unlikely, or an exceptional event. So, we may say that

$$\mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

We know that $\mathbb{P}(X_n - X| > \epsilon) \to 0$ as $n \to \infty$. So, let's take a $\limsup_{n\to\infty}$ on both sides. Then,

$$\limsup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \epsilon)$$

since the second term doesn't even depend on n. Now, allow $\epsilon \to 0$. Hence,

$$\limsup_n \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x),$$

which is half of what we need to prove. Note that on the right sides, we used the continuity.

For the other half, we need a formula like this:

$$\mathbb{P}(X_n \leq x) \geq \mathbb{P}(X \leq x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon).$$

Let's call this an exercise. Move the negative term to the other side and argue by cases.

Then we take a limit as $n \to \infty$. Taking a lim inf this time, the second term goes to zero again.

$$\liminf_n \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X \leq x - \epsilon).$$

Let $\epsilon \to 0$. Now, we can't use the right continuity, but we do have continuity of probability. So, we have

$$\liminf_n \mathbb{P}(X_n \leq x) \geq \mathbb{P}(X < x),$$

but since $\mathbb{P}(X = x) = 0$, we don't have to worry about this value. □

Convergence in distribution can not imply convergence in probability. This can't make sense, because probabilities $\mathbb{P}$ can be defined on different outcome spaces. Nevertheless, we can take random variables converging in distribution and put them into a common probability space where they converge almost surely (actually, point-wise).

Theorem 20.5 (Skorokhod's Representation Theorem)**.** *Suppose $X_n \to X$ in distribution. Then there exist random variables Y_n distributed identically with X_n and a random variable Y distributed identically with X, and such that*

$$Y_n \to Y \text{ everywhere}^{65} \text{ (thus a.s.)}$$

Skorokhod was a Ukrainian mathematician. He is now at Michigan State University. Recall that there is a canonical representation of a random variable X on the probability space $\Omega = [0,1]$, $\mathcal{F} = \{\text{Borel sets}\}$, $\mathbb{P} = \text{uniform (Lebesgue) measure}$. How do we do that? Let $F(x)$ be the distribution function of X. We just invert the distribution function. Take an $x \in \Omega = [0,1]$ and take[66] $Y(x) := F^{-1}(x)$. Then Y has the same distribution as X. Y is distributed identically with X.

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(x \in [0,1] : F^{-1}(x) \leq y) \\ &= \mathbb{P}(x \in [0,1] : x \leq F(y)) \\ &= F(y) = \mathbb{P}(X \leq y).\end{aligned}$$

So, we'll prove Theorem 20.5:

Proof. Consider the canonical representation of X_n and X. F_n is the distribution of X_n and F is the distribution of X. First, we'll do a little bit of "wrong proof". Assume that $X_n \to X$ in distribution.

Then $F_n(x) \to F(x)$ for almost all x. We apply the inverse on this. Then $F_n^{-1}(x) \to F^{-1}(x)$ for almost all x. Then $Y_n(x) \to Y(x)$ for almost all x.

Redefine $Y_n(x) = 0, Y(x) = 0$ for all x where this convergence fails. This will not change the distribution function, since this changes only countably many points. But, it makes these random variables converge everywhere.

So, why is this proof wrong? The problem is in taking preimages. The implication we implicitly used is not justified. We'll try to justify it now. We have the preimages defined with the supremum

$$F^{-1}(x) = \sup\{y : F(y) < x\}.$$

[66] In general, we have to decide what we do about points of discontinuity. It's obvious what to do with points of discontinuity. When the function is flat, we have to decide to take something. For this, we decided the left endpoint. So, this is $F^{-1}(x) = \sup\{y : F(y) < x\}$

So assume that we have a flat region. Consider the "flat regions" defined as follows: for every x, we look at the intervals with endpoints a_x and b_x:

$$a_x = F^{-1}(x) = \sup(y : F(y) < x)$$
$$b_x = \inf(y : F(y) > x).$$

Exceptional sets: the set of all flat regions is

$$\Omega_0 = \{x : (a_x, b_x) \neq \emptyset\}.$$

So we record all the points where we have a flat region. As an exercise, Ω_0 is countable. Hence $\mathbb{P}(\Omega_0) = 0$. Recall that this notation $\mathbb{P}$ here is the Lebesgue/canonical measure.

We know that $F_n(x) \to F(x)$ for all x that are points of continuity of F. We want to show $F_n^{-1}(x) \to F^{-1}(x)$ for all $x \notin \Omega_0$. Let's leave as an exercise these two things:

1. $\liminf F_n^{-1}(x) \geq F^{-1}(x)$ for $x \notin \Omega_0$.
2. $\limsup F_n^{-1}(x) \leq F^{-1}(x)$ for $x \notin \Omega_0$.

Let me sketch the first fact. It suffices to show that an infinite tail is $> y$. So, it suffices to show that $\forall y < F^{-1}(x), y < F_n^{-1}(x)$ for sufficiently large n. Now we can truly invert things: So, $F_n(y) < x$. By continuity, this implies (exercise) that $F(y) < x$. By inverting, this implies $y < F^{-1}(x)$. We'll leave part (2) as an exercise. □

JANUARY 16, 2008

I will basically discuss two results. The first really important result today is the Continuous Mapping Theorem. Let me remind you that $X_n \to X$ in distribution if the distribution functions $F_n(x) \to F(x)$ point-wise at the points of continuity of F. We can't expect more than that because of the classic example: $X_n = \frac{1}{n}$ and $X = 0$.

Theorem 20.6 (Continuous Mapping Theorem)**.** *Let X_n, X be random variables.*

(i) Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then $X_n \to X$ in distribution implies

$$f(X_n) \to f(X) \text{ in distribution}^{67}.$$

(ii) Moreover[68]*, let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function and consider the set*[69] *$D_f = \{x \in \mathbb{R} : f \text{ is discontinous at } x\}$. If $\mathbb{P}(X \in D_f) = 0$, then $f(X_n) \to f(X)$ in distribution.*

[68] the harder part
[69] measurable, exercise.

What this says is that f might have discontinuities, but the random variable "does not feel" these discontinuities. The most elegant proof of this that I know uses the Skorokhod Representation Theorem. Let me briefly remind you: One may have $X_n \to X$ in distribution, but it is unreasonable to expect convergence almost everywhere. However, we can find random variables Y_n with same distributions as X_n such that here we have convergence everywhere. So, there exist Y_n, Y such that

$$Y_n(\omega) \to Y(\omega) \quad \forall \omega \in \Omega$$

where Y_n has identical distribution to X_n and Y has identical distribution to X.

Proof. (i) Convergence in distribution only depends on information on distribution functions. $f(Y_n) \to f(Y)$ for all $\omega \in \Omega$, by continuity of f. We have $f(Y_n) \to f(Y)$ in distribution. But remember that $f(Y_n)$ has the same distribution as $f(X_n)$. So, $f(X_n) \to f(X)$ in distribution.

(ii) If f is not continuous, how do we make that first jump? You have convergence not everywhere, but on a set of full measure.

Suppose f has some discontinuities. Then

$$Y_n(\omega) \to Y(\omega) \forall \omega$$

implies

$$f(Y_n) \to f(Y) \forall \omega$$

fails where? It fails on discontinuties. But

$$\mathbb{P}(\omega : f \text{ is discont. at } Y(\omega)) = 0$$

(Note, by Skohorkhod, we can write $\mathbb{P}(Y \in D_f) = 0$.) These are the only points where we can not jump in the implication. On the complement, we can make this step. So we can conclude

$$f(Y_n) \to f(Y)$$

almost everywhere. Then we can proceed as before.

□

It is always a good idea to see that all assumptions are necessary. So please check that if the condition $\mathbb{P}(X \in D_f) = 0$ does not hold, find a counterexample.

Is the convergence in the theorem any stronger? Suppose that $X_n \to X$ in probability. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuous. Now, is it true that $f(X_n) \to f(X)$ in probability? I think it should work, basically with no tricks. Let's talk about this after class. It's dangerous to talk on-the-fly and then get something wrong.

The second important result gives you basically three ways to define convergence in distribution. One definition was already given to you. You can view the next theorem as two more equivalent definitions. So let me write it down.

Theorem 20.7 (Characterization of convergence in distribution)**.** *The following three statements are equivalent.*

(i) $X_n \to X$ *in distribution.*

(ii) If f *is a continuous and bounded*[70] *function,* $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$.[71]

(iii) $\mathbb{P}(X_n \to A) \to \mathbb{P}(X \in A)$ *for any Borel set* A *such that* $\mathbb{P}(X \in \partial A) = 0$.

Recall, $\partial A = \overline{A} \backslash \text{interior}(A) = \{$limits of sequences of points in A that are also limits of sequences of points

Proof. As is customary, we'll show each condition implies the other.

- Let's show (i) $\Rightarrow$ (ii). Suppose that $X_n \to X$ in distribution. This is not very hard, because we will use a result of measure theory. We'll use the Skorokhod Representation Theorem. We have $Y_n \to Y$. This is step number 1.

  Step number 2. As in the previous theorem, $f(Y_n) \to f(Y)$ everywhere, since f is continuous. To show (ii), it's enough to show it for the Y's. We need to show
  $$\mathbb{E}f(Y_n) \to \mathbb{E}f(Y).$$
  Since f is bounded, we can apply the Dominated Convergence Theorem to obtain
  $$\mathbb{E}f(Y_n) \to \mathbb{E}f(Y).$$
  The only thing that we used here is that the $|f(Y_n)|$ are bounded by something which is integrable:
  $$\max_{x \in \mathbb{R}} |f(x)|.$$
  Therefore, we have immediately that
  $$\mathbb{E}f(X_n) \to \mathbb{E}f(X).$$
  We relied on a nice result from measure theory, which made this easy.

- Let's proceed. (i) $\Rightarrow$ (iii).

  Let's try to mimick the proof we just presented. If we examine the left hand side,
  $$\mathbb{P}(X_n \in A) = \mathbb{E}f(X_n)$$
  where $f = \mathbf{1}_A$. If f is continuous, we'd be done. I will, nonetheless, keep this choice of f. I will show
  $$f(Y_n) \to f(Y)$$

[70] If f is unbounded, then we can see that the sequence can be unbounded.
[71] This must hold for **every** bounded continuous f.

almost everywhere, as in the proof of the Continuous Mapping Theorem. The convergence does not occur on a null set. Then, we can sprint towards the end of the result as before, using Dominated Convergence Theorem, since f is an indicator (thus bounded by 1).

Therefore, we immediately integrate them by DCT, and

$$\mathbb{E}f(Y_n) \to \mathbb{E}f(Y).$$

Using the same argument as before, $\mathbb{E}f(Y_n) = \mathbb{E}f(X_n)$ and $\mathbb{E}f(X) = \mathbb{E}f(Y)$. Thus,

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X).$$

Recall that f is an indicator, so

$$\mathbb{P}f(X_n \in A) \to \mathbb{E}f(X \in A).$$

- Now, if you're bored, the next part is just one line. (iii) $\Rightarrow$ (i).

  Why? Choose $A = (-\infty, x]$. Now, $\partial A = \{x\}$ is a singleton. What we have immediately is

  $$\mathbb{P}(X_n \in (-\infty, x]) \to \mathbb{P}(X \in (-\infty, x])$$

  for all x such that $\mathbb{P}(X = x) = 0$.

  The last condition says that x is a point of continuity for the distribution function of X.

- Let us prove (ii) $\Rightarrow$ (i).

  Let's lie again in the most horrible way. Let's consider a discontinuous function and dream that it is continuous. Let's consider the indicator $f = \mathbf{1}_{(-\infty,x]}$.

  Suppose that we had the property (ii) for this discontinuous function. Then, it would immediately imply

  $$F_n(x) = \mathbb{P}(X_n \in (-\infty, x]) \to F(x) = \mathbb{P}(X \in (-\infty, x]).$$

  We will approximate f be a new function f_ϵ, which is **almost** our indicator (in the interval $[x, x + \epsilon]$, f_ϵ is a straight line from 1 to 0). Then

  $$\mathbb{P}(X \leq x) = \mathbb{E}f(X) \leq \mathbb{E}f_\epsilon(X),$$

  just by looking at the graph, this is obvious.

  Moreover, you also have the inequality

  $$\mathbb{E}f_\epsilon(X) \leq \mathbb{P}(X \in (-\infty, x + \epsilon])$$

  Then, by (ii),

  $$\mathbb{E}f_\epsilon(X_n) \to \mathbb{E}f_\epsilon(X),$$

since $\mathbb{E}$ is continuous.

Then,

$$\limsup \mathbb{P}(X_n \leq x) \leq \mathbb{E}f_\epsilon(X) \leq \mathbb{P}(X \leq x + \epsilon)$$

Then, let $\epsilon \to 0$.

All of the proofs here are easy, and there are actually different orders that you can prove these in. All of the individual proofs are maybe three lines each. □

JANUARY 18, 2008

A sequence of random variables $X_n \to X$ in distribution $\Leftrightarrow \mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for every bounded and continuous f. Sometimes, some literature defines this as the convergence in distribution. This is sometimes more useful than the standard definition. Let's see why.

As an example, we can take $f(x) = |x|^p$ to be a polynomial (I'll take absolute value). So, if $X_n \to X$ in distribution, then

$$\mathbb{E}|X_n|^p \to \mathbb{E}|X|^p \text{ for all } 0 < p < \infty.$$

these are called the *absolute moments*. In particular, if you take the p^{th} root, then you get the L_p norm, so

$$\|X_n\|_{L_p} \to \|X\|_{L_p}.$$

The f here is not bounded. We can probably fix this by the truncation.

Above in the equivalence, we don't get either direction by taking out the boundedness property.

20.1 Helly's Selection Theorem

This completes our picture of convergence. As we know, almost sure convergence implies convergence in probability, which in turn implies convergence in distribution. If all are on the same probability space, then Skorokhod's implies full circle. Helly's Selection Theorem, tells us that it's almost true that for every sequence of random variables that there is some subsequence that converges in distribution (so in turn in all of the other forms). This is a statement that's analogous to compactness.

Theorem 20.8 (Helly's Selection Theorem)**.** *For every sequence of distribution functions F_n, there exists a subsequence F_{n_k} and a bounded, right-continuous, non-decreasing function F such that*

$$F_{n_k}(x) \to F(x) \text{ at all continuity points } x \text{ of } F.$$

Why do we not say that F is a distribution function? It may not be the case that the limits at $\pm\infty$ are correct. That is, $F(x) \not\to 0$ as $x \to -\infty$ and $F(x) \not\to 1$ as $x \to \infty$. Here is an example:

$X_n = n$ with probability 1. The point-wise limit of the F_n's is zero, and 0 is not a distribution function of any random variable. Of course, the same happens for any subsequence: $F_{n_k} \to 0$ for $\forall n_k$. In particular, X_n does not convergence in distribution (to anything). The only possible candidate was zero. Of course, we could have already seen this because there's just a big mass at n.

We'll prove the theorem first, and then we'll try to get rid of this obstacle. We can give a condition that will ensure it converges to a distribution function. What follows is [Durrett 2.4, Billingsley 2.5, Gut]. This is a proof of Helly. It is elementary in that it does not rely on any other theorems, but it is very non-trivial.

Proof of Theorem 20.8. We want a statement about compactness. For every fixed x, $0 \leq F_n(x) \leq 1$. By Bolsano-Weierstrass, there is a subsequence n_k for which $F_{n_k}(x)$ converges. We can repeat this for countably-many x. This is a Cantor's diagonal argument.

By Cantor's Diagonal Method, we can find a subsequence F_{n_k} such that

$$F_{n_k}(q) \text{ converges for every } q \in \mathbb{Q}.$$

We call this limit F_∞. That is,

$$F_{n_k}(q) \to F_\infty(q), \forall q \in \mathbb{Q}.$$

This makes sense: we did what we could. Do what's true for an x, and repeat for as many as possible. The problem is that we do not know to fill the gaps. It's not right-continuous. It's not non-decreasing.

So, what we do is this: Now define

$$F(x) := \inf\{F_\infty(q) : q \in Q, q > x\}.$$

This function F "truncates" away the bumps in the middle of the function that made it non-decreasing. I did many things at once. First, I defined F on all of $\mathbb{R}$. Second, I removed the not non-decreasing property. This is obvious. Third, F is right-continuous. We can check this:

$$\lim_{x_n \searrow x} F(x_n) = \inf_{y > x} F(y),$$

since F is non-decreasing. Then this is just

$$= \inf\{F_\infty(q) : q > x\} = F(x).$$

The only thing to patch up is the convergence. By the definition of F, there exists a rational $q > x$ such that

$$F_\infty(q) < F(x) + \epsilon. \tag{29}$$

(Think of this q as a "witness" to the infimum.)

Then, by the definition of $F_\infty(q)$,

$$F_{n_k}(q) \to F_\infty(q).$$

Since $q > x$, $F_{n_k}(x) \leq F_{n_k}(q)$. This along with (29) imply

$$F_{n_k}(x) \leq F(x) + \epsilon$$

for sufficiently large k.

Now we need a lower bound for $F_{n_k}(x)$. Now we want to do things to the left, and we do not have a control on what is happening (no witnesses). We know only one thing that can help us to the left, and that's right-continuity. This says that if we take an x and jump a little to the left, the value can not dramatically drop.

Let x be a point of continuity of F. That is, by the left continuity of F, there exists $r < x$ such that

$$F(r) > F(x) - \epsilon.$$

Choose $r' \in \mathbb{Q}$ such that $r < r' < x$. We argue as before,

$$F_{n_k}(r) \to F_\infty(r').$$

Now, we want to compare this to x. $F_\infty(r') \geq F(r) \geq F(x) - \epsilon$ since $F(r)$ is the infimum of all the values (definition of F). Also, $F_{n_k}(x) \geq F_{n_k}(r')$ because F_{n_k} is non-decreasing. Thus, we have proved the lower bound

$$F_{n_k}(x) \geq F(r) - \epsilon \text{ for sufficiently large } k.$$

By combining our inequalities,

$$F(x) - \epsilon \leq F_{n_k}(x) \leq F(x) + \epsilon.$$

Then $\limsup F_{n_k}(x) \leq F(x) + \epsilon$ and $\liminf F_{n_k} \geq F(x) - \epsilon$. Since $\epsilon > 0$ is arbitrary, we can tighten this. So, we conclude

$$\lim F_{n_k}(x) = F(x).$$

□

I'll ask one quick question here. How do we know that this statement is not just empty? How do we know that there even exists a good F? There are only countably many discontinuities (proved as before).

JANUARY 25, 2008

21 Characteristic Functions

We started the big chapter on characteristic functions, which is a new way of studying random variables using Fourier analysis. The *characteristic function* of a random variable X is a function $\varphi : \mathbb{R} \to \mathbb{C}$

$$\varphi_X(t) = \varphi(t) = \mathbb{E}e^{iXt}, \quad t \in \mathbb{R}.$$

How is this connected to Fourier analysis? If X has density f, then

$$\varphi(t) = \int_{-\infty}^{\infty} e^{ixt} f(x)\, dx = \hat{f}(t). \tag{30}$$

So for these functions, the characteristic function is the Fourier transform of the density function. Even when X doesn't have a density, we can heuristically think of this like a Fourier transform. We replace $f(x)\, dx$ by $d\mathbb{P}(x)$.

Why do we need this? Why is this approach to characteristic functions important? It's because of the following fundamental properties:

1. There is a one-to-one correspondence between distributions of random variables and characteristic functions of random variables[72]. So, no information is lost studying a distribution this way.

2. If X and Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$. This is nice[73] The convolution goes into product in Fourier transforms. This property is true even when the density isn't defined.

3. Pointwise convergence of characteristic functions implies convergence in distribution of random variables. The converse is almost true[74].

These will be our strategies for central limit theorems.

21.1 Properties

Let's start with simple properties first:

Property 1 (Boundedness). $\varphi(0) = 1$ and $|\varphi(t)| \leq 1$ for all t.

Proof. $\varphi(0) = \mathbb{E}e^{iX0} = \mathbb{E}1 = 1$. And $|\varphi(t)| = |\mathbb{E}e^{iXt}| \leq \mathbb{E}|e^{iXt}| = 1$. □

Boundedness is important when you try to use Dominated Convergence.

Property 2. $\varphi(-t) = \overline{\varphi(t)}$, where if $x = a + ib$, $\overline{x} = a - ib$.

Proof. Recall Euler's Formula: $e^{i\theta} = \cos\theta + i\sin\theta$. $\varphi(-t) = \mathbb{E}e^{-iXt} = \mathbb{E}\cos(Xt) - \mathbb{E}i\sin(Xt) = \mathbb{E}(\overline{\cos(Xt) + i\sin(Xt)}) = \overline{\mathbb{E}e^{iXt}} = \overline{\varphi(t)}$. □

[72] Not between the random variables and the characteristic functions. Why? The integral depends only on the distribution

[73] If you try to study X and Y through their densities, you get convolution, which is not an easy operation at all.

[74] This is important in central limit theorems. Note the conclusion.

Corollary 21.1. *$\varphi_{-X}(t) = \overline{\varphi_X(t)}$. In particular, if X is symmetric, then $\varphi(t)$ is real*[75].

Property 3. If X, Y are independent random variables, then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

Proof.

$$\begin{aligned}\mathbb{E}e^{i(X+Y)t} &= \mathbb{E}(e^{iXt}e^{iYt}) \\ &= \mathbb{E}e^{iXt} \cdot \mathbb{E}e^{iYt},\end{aligned}$$

by independence. □

Let's do some examples.

1. The first non-trivial random variable is Bernoulli. I want to make a symmetric Bernoulli. $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$. The expectation is

$$\varphi(t) = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos t.$$

2. Normal distribution $N(0, 1)$. It has density $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. We use (30).

$$\varphi(t) = \hat{f}(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{ixt}e^{-x^2/2}\,dx.$$

Now, we now how to integrate f, so we want to make something that looks like it by completing the square.

$$\frac{x^2}{2} - ixt = \frac{1}{2}(x - it)^2 + \frac{t^2}{2}.$$

So

$$\begin{aligned}\varphi(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-it)^2}e^{-t^2/2}\,dx \\ &= e^{-t^2/2} \cdot \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-it)^2}\,dx \\ &= e^{-t^2/2} \cdot \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\,dy \\ &= e^{-t^2/2}\end{aligned}$$

Justify $y = x + it$ by a symmetry or Cauchy integral argument.

[75] Actually, the converse is true as well.

3. Uniform distribution on $[a, b]$. Has density $f(x) = \frac{1}{b-a} \cdot \mathbf{1}_{[a,b]}$.

$$\varphi(t) = \int_a^b \frac{e^{ixt}}{b-a} dx = \frac{1}{it(b-a)} \int_{iat}^{ibt} e^y \, dy = \frac{e^{ibt} - e^{iat}}{it(b-a)}.$$

In the case $a = -1$ and $b = 1$, we get

$$\varphi(t) = \frac{e^{it} - e^{-it}}{2it} = \frac{\sin t}{t}.$$

This is sometimes called the "sinc(t)" function.

I will try to give you a feeling next of the Fourier analysis. This won't be needed if you're not familiar, but it's helpful.

22 Heuristics using Fourier Analysis

One of the main principles of the Fourier transform is Parseval's identity. Consider the Hilbert space $L^2(\mathbb{R})$ of functions with the inner product

$$\langle f, g \rangle \int f \overline{g}.$$

Then, the L^2 norm will be $\|f\|_{L^2}^2 = \int |f|^2$. Parseval's identity says that the Fourier tranform is a unitary map. Due to the normalization of probability, we have a corrective constant: $U : f \mapsto \frac{1}{\sqrt{2\pi}} \hat{f}$ is unitary. So $\|f\|_{L^2} = \|Uf\|_{L^2}$. Thus,

$$\int |f|^2 = \frac{1}{2\pi} \int |\hat{f}|^2$$

We will use $U^*U = \text{Id}$. $U^{-1} = U^*$.

$$(U^* g)(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} g(t) \, dt.$$

Therefore

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} \hat{f}(t) \, dt$$

So if you know the Fourier transform of a function, then you will know the function itself. This is called the *Fourier Inversion Formula*. This is because unitaries (rotations) are easy to invert. This is essentially the reason why we have property 1.

JANUARY 28, 2008

Suppose μ is a probability measure. Let $\varphi(t) = \int e^{itx} \mu(dx)$ be the corresponding characteristic function. The characteristic function uniquely determines the measure μ.

The inversion formula says

Theorem 22.1 (Inversion Formula)**.** *If $a < b$, then*

$$\lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\ dt = \mu(a,b) + \frac{1}{2}\mu(\{a,b\}).$$

Before proving the inversion formula, a couple of remarks:

Remark 22.2. *a) The integrand is*

$$\frac{e^{-ita} - e^{-itb}}{it}\varphi(t).$$

This is just the integral

$$\left(\int_a^b e^{-itx}\ dx\right)\varphi(t)$$

so

$$\left|\frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\right| \le \int_a^b |e^{-itx}|\ dx|\varphi(t)| \le |b-a|.$$

b) If $\mu = \delta_0$ is a point mass at 0, then $\varphi(t) \equiv 1$. In this case, if $b = 1$ and $a = -1$, the integrand is

$$\frac{2\sin t}{t},$$

and this is not integrable. The integral does not converge absolutely.

So let's prove the inversion formula:

Proof. Let

$$I_T = \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\ dt$$

which, when plugging in is

$$\int_{-T}^{T}\int \frac{e^{-ita} - e^{-itb}}{it}e^{itx}\mu(dx)\ dt.$$

Since the integrand is bounded, μ is a probability measure, and $[-T,T]$ is a finite interval, and $\cos(-u) = \cos u$ and $\sin(-u) = -\sin u$, we have

$$\begin{aligned} I_T &= \int\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it}e^{itx}\ dt\ \mu(dx) \\ &= \int\left[\int_{-T}^{T}\frac{\sin t(x-a)}{t}\ dt - \int_{-T}^{T}\frac{\sin t(x-b)}{t}\ dt\right]\ \mu(dx). \end{aligned}$$

There's a cosine part that goes away, because $\cos(-x) = \cos x$.

Let's introduce a function

$$R(\theta, T) = \int_{-T}^{T} \frac{\sin \theta t}{t} dt.$$

Then

$$I_T = \int (R(x-a, T) - R(x-b, T))\ \mu(dx) \tag{31}$$

Let $S(T) = \int_0^T \frac{\sin x}{x} dx$
Then for $\theta > 0$ changing variables $t = \frac{x}{\theta}$ gives

$$R(\theta, T) = 2 \int_0^{T\theta} \frac{\sin x}{x} dx = 2S(T\theta).$$

If $\theta < 0$,

$$R(\theta, T) = -R(|\theta|, T) = -2S(T\theta).$$

Thus, for any $\theta \in \mathbb{R}$, one has the formula

$$R(\theta, T) = 2\,\mathrm{sgn}(\theta) S(T|\theta|).$$

As $T \to \infty$, $S(T) \to \frac{\pi}{2}$. (see Durrett, Appendix exercise 6.6.) So $R(\theta, T) \to \pi\,\mathrm{sgn}(\theta)$, and

$$R(x-a, T) - R(x-b, T) \to \begin{cases} 2\pi & a < x < b \\ \pi & x = a \text{ or } x = b \\ 0 & x < a \text{ or } x > b \end{cases}$$

and $|R(\theta, T)| \leq 2 \sup_y S(y) < \infty$. Using the Bounded Convergence Theorem with (31), we get

$$\frac{1}{2\pi} I_T \to \mu(a, b) + \frac{1}{2}\mu(\{a, b\}),$$

which is the inversion formula. □

Let's do a little practice with characteristic functions. This is Durrett, exercise 3.4. If $X_1 \sim N(0, \sigma_1^2)$ and $X_2 \sim N(0, \sigma_2^2)$ and X_1 and X_2 are independent, then $X_1 + X_2 \sim N(0, \sigma_1^2 + \sigma_2^2)$.

Proof. $X_1 \stackrel{d.}{=} \sigma_1 Z$ and $X_2 \stackrel{d.}{=} \sigma_2 Z$ where $Z \sim N(0, 1)$. Let me remind you

$$\varphi_Z(t) = e^{-t^2/2}.$$

Then we have the general formula that

$$\begin{aligned} \varphi_{a+bX}(t) &= \mathbb{E}(e^{(a+bX)ti}) \\ &= e^{ati} \varphi_X(bt) \end{aligned}$$

So

$$\begin{aligned}\varphi_{X_1+X_2}(t) &= \varphi_{X_1}(t)\varphi_{X_2}(t) \\ &= e^{-\frac{1}{2}(\sigma_1 t)^2} \cdot e^{-\frac{1}{2}(\sigma_2 t)^2} \\ &= e^{-\frac{1}{1}2(\sigma_1^2+\sigma_2^2)t^2} \\ &= \text{characteristic function for } N(0, \sigma_1^2 + \sigma_2^2).\end{aligned}$$

By the inversion formula, we're done. □

Another fact is that φ is integrable only if the underlying measure is nice:

Theorem 22.3. *If $\int |\varphi(t)|\ dt < \infty$, then μ has bounded continuous density given by*

$$f(y) = \frac{1}{2\pi} \int e^{-ity} \varphi(t)\ dt.$$

Proof of this is deferred. Let's look at an example of how to use this. Recall Durrett, example 3.5. "triangular distribution." The graph of the density is

$$f(x) = 1 - |x|, x \in (-1, 1)$$

The characteristic function is $2(1 - \cos t)/t^2$. The (second) inversion formula from Theorem 22.3 gives

$$(1 - |y|)^+ = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-isy} \frac{2(1 - \cos s)}{s^2}\ ds.$$

This gives us a clever way of figuring out the density function. It actually turns out that $\frac{1-\cos t}{\pi t^2}$ is the density of a random variable. So, we can actually use this formula to figure out its characteristic function as $(1 - |y|)^+$.

JANUARY 30, 2008

22.1 Inversion Formula

We are going through the inversion formula, which tell you characteristic functions determine the distribution. We'll state it in a little bit easier way than was stated last time. The inversion formula involves an integral, but it involves a limit of integrals centered at zero.

The *principal value* of the integral is

$$\text{p. v.} \int_{-\infty}^{\infty} = \lim_{T \to \infty} \int_{-T}^{T}.$$

Theorem 22.4 (Inversion Formula)**.** *Let X be a random variable with characteristic function φ. If $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$, then*

$$\mathbb{P}(a < X \leq b) = \frac{1}{2\pi} \text{p. v.} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\ dt.$$

The formula itself may not look too pretty with the difference of exponentials, but one thing to remember is the representation: This kernel is

$$\frac{e^{-ita} - e^{-itb}}{it} = \int_a^b e^{-itx}\ dx,$$

which may sound even more complicated, but let's leave the following as an exercise: Try to interpret this using Fourier Analysis. We are taking an inverse Fourier transform (with the minus sign on the exponent).

One important corollary that explain the meaning of the inversion formula is:

Corollary 22.5. *The characteristic function determines the distribution of X uniquely.*

So there is a one-to-one correspondence between distributions and characteristic functions. Why is this true? We want to know the probabilities of X on intervals. The values on $\mathbb{P}(a < X \leq b)$ will determine the distribution functions. The only problem is the condition $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$.

Proof. First, we note that

$$\mathbb{P}(a < X \leq b) = F(b) - F(a).$$

The condition means that F is continuous at a and at b (there is no jump). So, the inversion formula determines $F(b) - F(a)$ for all points of continuity a, b of F. We can send a to $-\infty$. By sending $a \to -\infty$, this determines $F(b)$ for all points b of continuity of F. By the right continuity of F, we can recover the values at the discontinuity points. □

So, we can study random variables through their characteristic functions, and this will be our approach to the Central Limit Theorem.

Now, we're going to simplify the Inversion Formula a little bit. We usually apply it when X has density.

Theorem 22.6 (Inversion Forumla for densities)**.** *Let X be a random variable with characteristic function φ, and assume*

$$\int_{-\infty}^{\infty} |\varphi(t)|\ dt < \infty.$$

This alone means that X has continuous density, and we can recover the density. Specifically,

$$f(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \varphi(t)\ dt.$$

Remark 22.7. *Recall that $\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\ dx = \hat{f}(t)$. If we recall the geometry of the Fourier transform, if we recall the linear operator $U : f \mapsto \frac{1}{\sqrt{2\pi}} \hat{f}$*

is a unitary operator in L^2. In other words, U is like a rotation. To invert, we rotate back. Then

$$U^{-1} = U^*,$$

and $U^*\varphi$ can be easily computed: $U^*\varphi = \frac{1}{\sqrt{2\pi}}\overline{\widehat{\varphi}}$. So

$$(U^*\varphi)(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\varphi(t)\ dt.$$

This is a remark based on Fourier Analysis as to why this formula is true, but we'll actually prove it.

Proof. First, we'll try to understand why the density is continuous.

1. Every function f of this form is continuous. Indeed, let $y_n \to y$. Then, $e^{-ity_n}\varphi(t) \to e^{-ity}\varphi(t)$. So, the integrands will converge for every t. And they are bounded by something integrable (by hypothesis). Specifically, $|e^{-ity_n}\varphi(t)| = |\varphi(t)|$, which is an integrable function. By the Dominated Convergence Theorem, the integrals converge: $f(y_n) \to f(y)$.

2. We want to show that X has density. To start, we'll show that X has no point masses. That is, $\mathbb{P}(X = x) = 0$ for all x. This is of course a necessary condition (though not sufficient[76]).

   It suffices to prove, for any two points a and b such that $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$, and $a < x < b$, that $\mathbb{P}(a < X \leq b) \leq C(a, b) \to 0$ as $a \to x$, $b \to x$.

   We use Inversion Formula and the inequality

   $$\left|\frac{e^{-ita} - e^{-itb}}{it}\right| = \left|\int_a^b e^{-ita}\ dt\right| \leq \int_a^b 1\ dt \leq |b - a|.$$

   This gives

   $$\mathbb{P}(a < X \leq b) \leq \frac{1}{2\pi}\,\text{p. v.}\int_{-\infty}^{\infty} |b - a||\varphi(t)|\ dt \leq |b - a|\int_{-\infty}^{\infty} |\varphi(t)|\ dt.$$

3. Formula for the density: Let $a < b$ be arbitrary. Thus (since X has no point masses),

   $$\begin{aligned}\mathbb{P}(a < B \leq b) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\varphi(t)\ dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\int_a^b e^{-ity}\ dy\right)\varphi(t)\ dt \\ &= \int_a^b \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\varphi(t)\ dt\right)\ dy\end{aligned}$$

[76] If F is continuous but not differentiable (like the Cantor distribution), this has no point masses, but no density function.

and so we're done as $f(y)$, which is in parentheses, is the density of X. (Note the application of Fubini's theorem above.)

4. How do we use the principal value? Justify going from p.v. to the true integral $\int_{-\infty}^{\infty}$.

   $f_T(y) := \frac{1}{2\pi}\int_{-T}^{T} e^{-ity}\varphi(t)\,dt$. We know that $f_T(y) \to f(y)$ as $T \to \infty$ for all y, by the Dominated Convergence Theorem.

   Moreover, $|f_T(y)| \le \frac{1}{2\pi}\int_{-T}^{T} |e^{-ity}\varphi(t)|\,dt \le \frac{1}{2\pi}\int_{-\infty}^{\infty} |\varphi(t)|\,dt < \infty$. By the Dominated Convergence Theorem,

$$\lim_T \int_a^b f_T(y)\,dy \to \int_a^b f(y)\,dy.$$

□

The intuitive idea is to recognize the kernel as an integral itself. This is Fourier inversion formula for densities. If the Fourier transform is in L^1 and bounded, then the Fourier inversion formula holds. It's a nice general result in analysis.

FEBRUARY 1, 2008

22.2 Continuity Theorem

Here we go with one of the main results of characteristic function theory.

Theorem 22.8 (Continuity Theorem)**.** *Let X_n, X be random variables with characteristic function φ_n, φ. Then $X_n \to X$ in distribution if and only if $\varphi_n(t) \to \varphi(t)$ for every t.*

This is a full equivalence, but you must ensure that φ is the characteristic function of a random variable. The reverse direction is the important direction.

Proof of necessity. Let $X_n \to X$ in distribution. Then $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for any continuous and bounded function f. Hence $\mathbb{E}e^{itX_n} \to \mathbb{E}e^{itX}$ because the function e^{itx} is bounded. It is not real-valued, but you can split it into real and imaginary parts. □

Here is another property of characteristic functions:

Property 4. Characteristic functions are continuous.

Proof. We want to show that if $t_n \to t$, then $\mathbb{E}e^{it_nX} \to \mathbb{E}e^{itX}$. This is true by the Dominated Convergence Theorem, because $|e^{it_nX}| \le 1$ and $e^{it_nX(\omega)} \to e^{itX(\omega)}$ for every $\omega \in \Omega$. □

As an exercise, show that characteristic functions are uniformly continuous (on $\mathbb{R}$).

We will want to show sufficiency for Theorem 22.8. I do not know any trivial argument. We will first show the tightness of the sequence. For the proof of sufficiency, we will first show that: $\varphi_n(t) \to \varphi(t)$ for all t implies that (X_n) is tight. Recall, tightness is: $\forall \epsilon > 0, \exists M : \mathbb{P}(|X_n| \geq M) \leq \epsilon$ for all n. That is, mass can not escape.

We would want to use the inversion formula, but this will not work. We have point-wise convergence, and we can not use something like dominated convergence.

Now and in the sequel, the main theme is to read as much as possible from the characteristic function. How can we read tightness information from the characteristic function? Most of the information we want to know is contained in the interval around zero in the characteristic function.

Recall the characteristic function takes value 1 at 0.

Here is the idea:

Lemma 22.9. *For every*[77] $u > 0$,

$$\frac{1}{2u}\int_{-u}^{u}(1-\varphi(t))\,dt \geq \frac{1}{2}\mathbb{P}\left(|X| \geq \frac{2}{u}\right).$$

Proof. By the definition of the characteristic function,

$$\frac{1}{2u}\int_{-u}^{u}(1-\mathbb{E}e^{itX})\,dt = \mathbb{E}\left[\frac{1}{2u}\int_{-u}^{u}(1-e^{itX})\,dt\right] = \mathbb{E}\left(1-\frac{\sin(uX)}{uX}\right).$$

The sinc function will be eventually below $\frac{1}{2}$. Thus, the quantity above is

$$\geq \mathbb{E}(1-\frac{1}{2})\mathbf{1}_{\{|uX|\geq 2\}} = \frac{1}{2}\mathbb{P}(|uX| \geq 2),$$

which was what we needed to prove. □

The left hand side is actually real by the proof.

Proof of tightness. Our limiting characteristic function φ is one specific function. We choose u so that the left hand side is a small constant. We apply the lemma to the limiting random variable. The φ_n converge to φ pointwise, so we can apply a limiting argument.

Since $\varphi(0) = 1$ and φ is continuous at 0, given an $\epsilon > 0$, we can choose $u > 0$ such that

$$\frac{1}{2u}\int_{-u}^{u}(1-\varphi(t))\,dt < \epsilon.$$

Since $\varphi_n(t) \to \varphi(t)$ pointwise, by the Dominated Convergence Theorem, $\exists n_0$ such that

$$\frac{1}{2u}\int_{-u}^{u}(1-\varphi_n(t))\,dt < 2\epsilon \text{ for all } n \geq n_0.$$

[77] u will be very small, and we'll study intervals $(-u, u)$.

By the Lemma,

$$\mathbb{P}(|X_n| \geq \frac{2}{u}) \leq 4\epsilon) \text{ for all } n \geq n_0.$$

We have proved that $(X_n)_{n \geq n_0}$ is tight[78]. Therefore, $(X_n)_{n \geq 1}$ is tight. □

Proof of sufficiency in Theorem 22.8. We have already proved that the pointwise convergence of the $\varphi_n \to \varphi$ implies tightness of (X_n). Then Helly's theorem tells us that convergence is distribution holds on a subsequence.

For every subsequence of X_n, we can find a further subsequence that converges. We do not know that the analogy (of topology) from $\mathbb{R}$ applies perfectly, and we do not know that every subsequential limit is the same.

By Helly's Selection Theorem, there is a subsequence

$$X_{n_k} \to X' \text{ in distribution, for some random variable } X'.$$

We claim that $X = X'$. Indeed, by the necessity part of the theorem, $\varphi_{n_k}(t) \to \varphi_{X'}(t)$ for every t. But by the assumption, $\varphi_{n_k}(t) \to \varphi_X(t) = \varphi(t)$. Hence $\varphi_{X'}(t) = \varphi_X(t)\ \forall t$. By the Inversion Formula, X has the same distribution as X'.

We now know that every subsequence of X_n has a further subsequence that converges to X in distribution. We claim that $X_n \to X$ in distribution. (We know that this holds for the reals: if it didn't hold, we'd have a subsequence that does not converge, and the property above would not hold for this subsequence.)

Let f be a bounded continuous function. Then every subsequence of $a_n = \mathbb{E}f(X_n)$ has a further subsequence that converges to $a = \mathbb{E}f(X)$. Then $a_n \to a$. This is an exercise in real numbers.

Therefore $X_n \to X$ in distribution. □

This is a pretty ingenious proof. First, we used some local property of characteristic function, then Helly's theorem, and from a subsequential limit and empty hands, we get a full convergence statement. An alternative proof (which is more traditional) is in the textbook of Khoshevisan. It's a proof using the full proof of Fourier analysis and without using the topological tricks.

An exercise: If we do not know if φ is a characteristic function, then there is still a theorem: Suppose $\varphi_n(t) \to g(t)$ and g is continuous at 0. Then g is a characteristic function of some random variable X and

$$X_n \to X \text{ in distribution.}$$

In the proof, we only used continuity of $\varphi(t)$ and $\varphi(0) = 1$, but this second fact can be built from convergence.

FEBRUARY 4, 2008

It is possible that class will be cancelled on Friday.

[78] We can take care of the first n_0 by simply throwing them in and increasing M.

We are trying to study random variables through their characteristic functions. Characterisitic functions are expectations of exponentials. One way to study functions is through Taylor series. We can use the linearity of expectation.

22.3 Taylor expansions of Characteristic Functions

Suppose $\varphi(t) = \mathbb{E}e^{itX}$. Recall

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Then

$$\varphi(t) = \sum_{k=0}^{\infty} \mathbb{E}\frac{(itX)^k}{k!} = \sum_{k=0}^{\infty} \frac{(it)^k}{k!}\mathbb{E}(X^k).$$

In the right hand side, all the information about X that is needed is contained in its moments. If you know the moments of X, then you know the characteristic function. Intuitively, the moments of X should determine its distribution (if the series converge, etc.). By viewing this is a formal series, we can get the moments back from its Taylor expansion:

$$\varphi^{(k)}(0) = i^k\mathbb{E}(X^k). \tag{32}$$

So, the derivatives of the characteristic function at 0 give moments of a random variable. That looks very powerful, actually. The left hand side is simply local information about the characteristic function. Then you get all of the moments, and from all of the moments, you can plug back in and get the characteristic function at all t. You may have to worry about convergence. All of this was based on formal series. We have been careless about converges.

We will concern ourselves with the first two terms of the Taylor expansion. Usually, this is already a fairly good approximation of the exponential.

Lemma 22.10.

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \leq \frac{2|x|^n}{n!}.$$

When we truncate the power series, the biggest contribution is given by the last term.

The next lemma usually gives a much better approximation.

Lemma 22.11.

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \leq \frac{|x|^{n+1}}{(n+1)!}$$

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

This will not be true without the i in e^{ix} on the left hand side. The proof is not intuitive.

Proof. We'll integrate something like a Gamma function:

$$\int_0^x (x-s)^n e^{is}\ ds$$

If you integrate this, we will hopefully see the term in the sum in the statement of the lemma. We integrate by parts[79]: We will raise the exponent n to $n+1$. So $u = e^{is}$ and $dv = (x-s)^n\ ds$. Then $v = -\frac{(x-s)^{n+1}}{n+1}$. The integral becomes

$$\begin{aligned} uv|_0^x - \int_0^x v\ du &= -\frac{(x-s)^{n+1}}{n+1} e^{is}|_0^x + i\int_0^x \frac{(x-s)^{n+1}}{n+1} e^{is}\ ds \\ &= \frac{x^{n+1}}{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1} e^{is}\ ds \end{aligned}$$

When $n = 0$ (and then we'll use this formula inductively), we have

$$\int_0^x (x-s)^0 e^{is}\ ds = \int_0^x e^{is}\ ds = \frac{1}{i} e^{is}|_0^x = \frac{e^{ix}-1}{i}$$

by direct computation. By substituting into the previous formula,

$$\frac{e^{ix}-1}{i} = x + i\int_0^x (x-s) e^{is}\ ds.$$

By rearranging terms,

$$e^{ix} = 1 + ix + i^2 \int_0^x (x-s) e^{is}\ ds.$$

We just recycle this formula again. When $n = 1$,

$$\int_0^x (x-s) e^{is}\ ds = \frac{x^2}{2} + \frac{i}{2}\int_0^x (x-s)^2 e^{is}\ ds.$$

When we plug this in for what we have for one-term integration, we get

$$e^{ix} = 1 + ix + \frac{i^2 x^2}{2} + \frac{i^3}{2}\int_0^x (x-s)^2 e^{is}\ ds.$$

Here we get a three-term Taylor series. In the term with the integral, we see the error.

Inductively using this formula, we get

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\int_0^x (x-s)^n\ e^{is}\ ds.$$

Now, we estimate the error term $\frac{i^{n+1}}{n}\int_0^x (x-s)^n\ e^{is}\ ds$.

$$|\text{Error}| = \frac{1}{n!}\left|\int_0^x (x-s)^n e^{is}\ ds\right| \le \frac{1}{n!}\int_0^x (x-s)^n\ ds = \frac{1}{n!}\cdot\frac{x^{n+1}}{n+1} = \frac{x^{n+1}}{n+1}.$$

□

[79] Here, x is a constant and s is a variable.

When you apply these two lemmas for the characteristic function, then you get:

Corollary 22.12.

$$\left|\varphi(t) - \sum_{k=0}^{n} \frac{(it)^k}{k!}\mathbb{E}(X^k)\right| \leq \mathbb{E}\min\left\{\frac{t^{n+1}}{(n+1)!}|X|^{n+1}, \frac{2t^n}{n!}|X|^n\right\}$$

Note, by having the expectation on the right hand side on the outside of the minimum, the right hand side always exists.

This may sound too complicated, so let's do first term and second term approximation. So, for $n = 1$, we have this corollary:

Corollary 22.13 (Linearization)**.** *If* $\mathbb{E}|X| < \infty$*, then*

$$\varphi(t) = 1 + it\mathbb{E}X + o(t).$$

Proof. Use the Taylor expansion, recognizing 1 and $it\mathbb{E}X$ as the first two terms in the Taylor expansion.

$$\begin{aligned}|\varphi(t) - 1 - it\mathbb{E}X| &\leq \mathbb{E}\min\left(\frac{t^2}{2}|X|^2, 2t|X|\right)\\ &= t\cdot\mathbb{E}\min\left(\frac{t}{2}|X|^2, 2|X|\right)\\ &= t\cdot\mathbb{E}f(t,X),\end{aligned}$$

where $f(t,X) = \min\left(\frac{t}{2}|X|^2, 2|X|\right)$. Then $f(t,X) \leq 2|X|$. On the other hand, $f(t,X) \to 0$ as $t \to 0$.

By the Dominated Convergence Theorem, $\mathbb{E}f(t,X) \to 0$ as $t \to 0$. □

Corollary 22.14. *If* $\mathbb{E}|X| < \infty$*, then* $\varphi'(0) = i\mathbb{E}X$*.*

For $n = 2$, we get

Corollary 22.15 (Second order approximation)**.** *If* $\mathbb{E}|X|^2 < \infty$*, then*

$$\varphi(t) = 1 + it\mathbb{E}X - \frac{t^2}{2}\mathbb{E}(X^2) + o(t^2).$$

This is very useful. We will use this corollary in the proof of the Central Limit Theorem.

Similarly, one can read off the the second derivative information here. The proof of this is the same as the previous. If we keep going, then we have

Corollary 22.16. *If* $\mathbb{E}|X|^n < \infty$*, then* $\varphi^{(n)}(0) = i^n\mathbb{E}(X^n)$*.*

So $\frac{\varphi^{(n)}(0)}{i^n}$ is always real.

Heuristically, if you know the moments, you know the derivatives at 0, and then you know all of φ. The truth is not quite this. You can ask a general question. If you only know the moments, then do you know the distribution of X? Sometimes "yes" and sometimes "no". This problem is usually called the *moment problem*.

Question 22.17 (Moment Problem)**.** *Do the moments of X determine the distribution of X?*

In general, no. If you do not care how moments grow, then your Taylor series will diverge. If you have moments with bounded growth, then the answer is yes. If the moments grow moderately, then yes. It is clear why the growths of the moments should be necessary here (since you have Taylor series). If the moments grow like e^k, then you're okay. If the moments grow super-exponentially, then you have a problem.

Theorem 22.18. *If $\mu_k = (\mathbb{E}X^k)^{1/k} = O(k)$ for $k \in 2\mathbb{Z}$, $k \to \infty$, then there exist only one distribution of X with these moments.*

This is one solution to the moment problem. If the moments grow like k^k, then we are okay.

FEBRUARY 6, 2008

In a couple of hours, I'll post the first midterm. You will be able to use any material we've covered so far. Please do not use any theorems you've learned in other courses. Feel free to use any of the complex analysis you've learned. You can use any standard tables, or maybe computerized tables.

On Friday there will be no class (no office hours as well).

22.4 Central Limit Theorem for i.i.d. r.v.'s

So, this is where all our work in characteristic functions were for. Here's our result.

Theorem 22.19 (Lindeberg-Lévy)**.** *Let (X_n) be a sequence of independent and identically distributed random variables with mean μ and variance σ^2. Then $S_n = X_1 + \cdots + X_n$ satisfies the "Central Limit Theorem":*

$$\frac{S_n - \mu n}{\sigma\sqrt{n}} \to N \text{ in distribution as } n \to \infty,$$

where N is a standard normal random variable.

Of course the means and standard deviations on both sides must agree, which is why we standardized on the left. This is probably the simplest form of the Central Limit Theorem, established in the 20's. We definitely need variance, since N has variance.

The conclusion, equivalently, is: for $a \leq b$,

$$\mathbb{P}\left(a < \frac{S_n - \mu n}{\sigma\sqrt{n}} \leq b\right) \to \mathbb{P}(a < N \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

If you're interested in approximating a random variable over some interval, this gives an approximation.

Note, the De Moivre-Laplace CLT is a partial case (for Bernoulli X_n).

The proof, given what we know about characteristic functions, is easy. We'll take the characteristic function of N and the characteristic function of the left hand side. We'll check that the product converges.

Proof. First, we'll simplify our life. Without loss of generality, we can assume that $\mu = 0$ and $\sigma = 1$. For this, it's usual: we normalize the variables beforehand (and not when we're ready to apply the theorem). So, consider $X_n' = \frac{X_n - \mu}{\sigma}$.

We want to show that $\frac{S_n}{\sqrt{n}} \to N$ in distribution. We take the characteristic functions of the left hand side, and we have a sum of independent random variables, so we get a product. As we know, if $\mathbb{E}(X^2) < \infty$ (i.e., X has a second moment), then

$$\varphi(t) = 1 + it\mathbb{E}X - \frac{t^2}{2}\mathbb{E}(X^2) + o(t^2) \text{ as } t \to 0.$$

We need to know, however, a statement for all t. Actually, the $\sqrt{n}$ will help us. Namely, if we use this for X_n, then

$$\varphi_{X_k}(t) = 1 - \frac{t^2}{2} + o(t^2) \text{ as } t \to 0,$$

since $\mathbb{E}X = 0$. We only know this for t going to zero, but we need to know it for all t values.

What we are truly interested in is

$$\varphi_{X_k/\sqrt{n}}(t) = \varphi_{X_k}(t/\sqrt{n}).$$

This is great, because we are evaluating φ_{X_k} at only very small values of input (as n grows). So this will be

$$= 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

as $n \to \infty$ for any fixed t. The little o should say $o(t^2/n)$, but we've fixed t.

Thus, the characteristic function is a little parabola. We know this local behavior. Then, by independence,

$$\varphi_{S_n/\sqrt{n}}(t) = \prod_{k=1}^{n} \varphi_{X_k/\sqrt{n}}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n.$$

If there were no error term here (we need to justify this), then this would go to

$$\sim \left(1 - \frac{t^2}{2n}\right)^n \to e^{-t^2/2} = \varphi_N(t) \text{ as } n \to \infty.$$

Therefore, $\frac{S_n}{\sqrt{n}} \to N$ in distribution, by the Continuity Theorem. □

This is very interesting, because it's a local proof. We used the asymptotic as $t \to 0$, and since we were dividing by $\sqrt{n}$, we only used local (to 0) values of φ.

We need to justify $\sim$. If you believe that complex analysis is the same as real analysis, then you'd take logarithm of both sides. We use

$$\log(1+z) = z + o(1) \text{ as } z \to 0.$$

We'll leave that method/option as an exercise. We will not do it this way. We really need complex analysis (it's legitimate proof), since the error term is complex-valued.

A second way to justify this is with a lemma, which we will need later. The problem was that we had a product of n terms. If we had a sum, then we'd have already been done.

Lemma 22.20. *Let $z_1, \dots, z_n \in C$ and $w_1, \dots, w_n \in \mathbb{C}$ be of modulus at most 1. Then*

$$|z_1 \cdots z_n - w_1 \cdots w_n| \leq \sum_{k=1}^{n} |z_k - w_k|.$$

In the above work, we clearly have the correction terms being of modulus at most 1.

Proof. Because there is no constant in front, the most plausible thing to do is induction on n.

$$\begin{aligned} z_1 \cdots z_n - w_1 \cdots w_n &= z_1 \cdots z_n - w_1 z_2 \cdots z_n + w_1 z_2 \cdots z_n - w_1 \cdots w_n \\ &= (z_1 - w_1) z_2 \cdots z_n + w_1 (z_2 \cdots z_n - w_2 \cdots w_n) \end{aligned}$$

By applying absolute value everywhere, by the triangle inequality,

$$|z_1 \cdots z_n - w_1 \cdots w_n| \leq |z_1 - w_1| + |z_2 \cdots z_n - w_2 \cdots w_n|$$

and we are done by induction. □

We use this lemma for $z_k = 1 - \frac{t^2}{2n}$ and $w_k = 1 - \frac{t^2}{2n} + o(\frac{1}{n})$. This writing is a little bit informal here, since there error terms in w_k and different for different k.

Hence, $|\prod z_k - \prod w_k| \leq \sum_{k=1}^{n} o(\frac{1}{n}) = o(1)$ as $n \to \infty$.

Thus, $\sim$ in the proof of the theorem means

$$\left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n = \left(1 - \frac{t^2}{2n}\right)^n + o(1).$$

To use the Binomial Theorem, there are probably too many terms. There are exponentially many terms, but this was a good suggestion.

A couple of remarks:

1. Finite variance is needed.[80] A properly normalized sum of Cauchy random variables is a Cauchy random variable (just done as a HW exercise). So, the CLT does not hold for Cauchy random variables. The reason is that Cauchy random variables have heavy tails.

   For every $1 < p < 2$, there exists random variables (X_n) independent and identically distributed, with finite p^{th} moments, but with infinite 2^{nd} moment, for which CLT fails. In other words, you need the second moment. These are called *stable random variables* or *stable laws*. Intuitively, they are like Gaussians, except the exponent 2 is replaced by p: $e^{x^p/2}$.

2. CLT holds for Poisson random variables. In fact, Poisson random variables have all moments.

   Here is a warning. It looks like it should not be true, since a Poisson random variable can be estimated by a sum of Bernoulli random variables. If you add Bernoulli's with mean $\frac{1}{n}$, you get a Poisson: If X_k are Bernoulli with mean λ/n, their sum $S_n = X_1 + \cdots + X_n$ has mean λ. Then $S_n \to \text{Poisson}(\lambda)$. In particular, the CLT fails. The CLT fails because the mean $\mu = \lambda/n$ is not a constant.[81]

3. Finally, if you recall the Laws of Large Numbers (both Weak and Strong)[82], then they hold even if one only has pair-wise independence. Recall this is because $\mathbb{E}S_n^2 = \mathbb{E}(X_1 + \cdots + X_n)^2$ only keep pairwise terms, which are all zero.

   The Laws of Large Numbers hold for pairwise independent random variables. In this scenario, the CLT does not hold. We consider an example:

   Example 22.21. *Consider independent random variables (ξ_n) with*

   $$\mathbb{P}(\xi_k = -1) = \mathbb{P}(\xi_k = 1) = \frac{1}{2}.$$

   Of course, these satisfy CLT, but from these, we will cook up an example that does not. Consider

   $$\xi_1(1+\xi_2)(1+\xi_3)\cdot\cdots\cdot(1+\xi_{n+1}) = S_{2^n}.$$

   Why did we name it S_{2^n}? When we expand, it is the sum of 2^n terms of the form of different products of ξ_k's. All of these terms are pairwise independent (exercise: for any two terms, there are shared factors and not shared factors. We condition on the shared factors.).

   However, S_{2^n} is far from normal. Most of the time, it takes the value zero. Specifically

   $$S_{2^n} = \begin{cases} 0, & \text{prob.} 1 - 2^{-n} \\ 2^n, & \text{prob.} 2^{-n-1} \\ -2^n, & \text{prob.} 2^{-n-1} \end{cases}$$

[80] The variance is needed for any sort of Central Limit Theorem to hold at all.

[81] You could also say that CLT fails because they are not identically distributed, but we will see a form of CLT there too. So the main reason is that the mean is not constant.

[82] Say, only to discuss the Weak Law.

So, it's a three-valued random variable, which is far from normal.

So, in the theorem, everything is needed: mean, variance, and independence.

FEBRUARY 11, 2008

In problem six, $S_{\nu_n} = X_1 + \cdots + X_{\nu_n}$. In problem four, change $[-1, 1]$ to $[0, 1]$.

Today we will do the central limit theorem for non-identical distributed random variables. In some cases, this is the most general central limit theorem that you will ever need to use.

The set up is that you have (X_n) is a sequence of independent, but not necessarily identically distributed random variables. One example where this is needed: suppose you do have (Z_n) is a sequence of independent identically distributed random variables, but you look at them with weights

$$S_n = \sum_{k=1}^{n} a_k Z_k, \text{ where } a_k \in \mathbb{R}.$$

Then the $X_k = a_k Z_k$ are independent, but not identically distributed. Even in this example, it is interesting to see what condition you need on the a_k's. Obviously if they are all the same, then you have a central limit theorem. Again obviously if they are all zero except one of them, then you don't have a central limit theorem.

For the central limit theorem, we obviously need some "uniformity" of X_ns. One condition that we have for uniformity is tightness. This bounds all the range of X_ns within some interval, except some outlyers. This will not work, since tightness has nothing to do with expectations or moments or anything else. But you saw from homework that this won't work: little epsilons can escape very quickly. So, instead of tightness, which is the condition of examining tails:

$$\mathbb{P}(|X_n| > M) < \epsilon \text{ for all } n$$

we replace the $\mathbb{P}$ with $\mathbb{E}$. So, we will look at expectations. We will look at something like

$$\mathbb{E}|X_n| \cdot \mathbf{1}_{\{|X_n|>M\}} =: \mathbb{E}(|X_n| : |X_n| > M),$$

denoted as on the right hand side for convenience of notation.

Proposition 22.22. *For any random variable X,*

$$\mathbb{E}(|X| : |X| > M) \to 0 \text{ as } M \to \infty.$$

Proof. The event $\{|X| > M\} \to \emptyset$ as $M \to \infty$. Thus, $\mathbf{1}_{\{|X|>M\}} \to 0$ pointwise as $M \to \infty$.

Then $|X| \cdot \mathbf{1}_{\{|X|>M\}} \to 0$ as $M \to \infty$ and is bounded by $|X|$. By the Dominated Convergence Theorem, $\mathbb{E}|X| \cdot \mathbf{1}_{\{|X|>M\}} \to 0$ as $M \to \infty$. □

So we know this one fact that for a single random variable. If this holds uniformly for a sequence of random variables, then we will have a central limit theorem.

This is the remarkable condition, called *Lindeberg's Condition*.

Definition 22.23 (Lindeberg's Condition)**.** *Consider independent random variables X_n with means 0 and variances σ_n^2. Let*

$$s_n^2 = \sum_{k=1}^{n} \sigma_k^2 = \sum_{k=1}^{n} \mathbb{E}(X_k^2).$$

We say that the sequence (X_n) satisfies Lindeberg's condition if, $\forall \epsilon > 0$,

$$\sum_{k=1}^{n} \mathbb{E}(X_k^2 : |X_k| > \epsilon s_n) = o(s_n^2) \text{ as } n \to \infty.$$

We will prove the theorem, and then unwrap this rather technical definition.

Example 22.24. *Suppose X_ns are i.i.d. with variance σ^2. This is the easiest example for which to understand this. Then $s_n^2 = \sigma^2 n$. Then the Lindeberg's condition is verified:*

$$n \cdot \mathbb{E}(X_1^2 : |X_1| > \epsilon\sigma\sqrt{n}) = o(n),$$

which is equivalent to saying that

$$\mathbb{E}(X_1^2 : |X_1| > \epsilon\sigma\sqrt{n}) \to 0 \text{ as } n \to \infty.$$

So, this is true by Proposition 22.22.

Think of all variances as being common, and being 1. Then, the sum in the definition is $o(1)$. So, the average will (collectively) go to zero. We'll work with the condition next time.

Theorem 22.25 (Lindeberg-Feller's Central Limit Theorem)**.** *Let X_n be random variables satisfying Lindeberg's condition. Then, the Central Limit Theorem holds: The partial sums $S_n = X_1 + \cdots + X_n$ satisfy*

$$\frac{S_n}{s_n} \to N \text{ as } n \to \infty \text{ in distribution,}$$

where N is the standard normal random variable.

The independence for the random variables is stated within the definition. The Lindeberg condition is a necessary condition for the CLT to hold. Levy proved that it is a necessary fact, so it is the weakest condition that must hold.

Remark 22.26. *We can change random variables X_k as we increase n. More formally, we will look at random variables (X_{nk}), $k = 1, \dots, n$ and $n = 1, 2, \dots$. That is,*

$$\begin{aligned} S_1 &= X_{11} \\ S_2 &= X_{21} + X_{22} \\ S_3 &= X_{31} + X_{32} + X_{33} \end{aligned}$$

We say that the (X_{nk}) form a triangular array.

Proof of Theorem 22.25 for Triangular Arrays. We can simplify by assuming that $s_n = 1$. Why can we do this? We do this by considering $X'_{nk} = X_{nk}/s_n$. How does this affect Lindeberg's condition?

The proof is similar to how we proved the CLT before. We will examine characteristic functions and Taylor series. Use Corollary 22.12 for a random variable X. Recall $\varphi(t) = \mathbb{E}e^{itx}$. Then by Taylor's expansion,

$$|\varphi(t) - (1 + it\mathbb{E}X - \frac{t^2}{2}\mathbb{E}X^2)| \le \mathbb{E}\min(|tX|^3, |tX|^2).$$

In our case, for X_ks,

$$|\varphi_k(t) - (1 + 0 - \frac{1}{2}t^2\sigma_k^2)| \le \mathbb{E}\min(|tX_k|^3, |tX_k|^2). \tag{33}$$

Now comes a crucial difference with the other proof. Why do we need two ways of bounding? One must be good for different situations. We will use $|tX_k|^3$ when $|X_k$ is small, and we will use $|tX_k|^2$ for large values of $|X_k|$.

So the right hand side in (33) is

$$\le \mathbb{E}(|tX_k|^3 : |X_k| < \epsilon) + \mathbb{E}(|tX_k|^2 : |X_k| > \epsilon).$$

Now the problem is that we need to avoid integration, because we might not have third moment. A very useful trick in analysis is: $|tX_k|^3 = |tX_k| \cdot |tX_k|^2 \le \epsilon|t|^3X_k^2$. So the right hand side of (33) obeys

$$\text{RHS} \le \epsilon|t|^3\sigma_k^2 + t^2\mathbb{E}(X_k^2 : |X_k| > \epsilon).$$

Now we take a sum on both sides.

$$\begin{aligned} \sum |\varphi_k(t) - (1 - \frac{1}{2}t^2\sigma_k^2)| &\le \epsilon|t|^3 \sum \sigma_k^2 + t^2 \sum \mathbb{E}(X_k^2 : |X_k| > \epsilon) \\ &\le \epsilon|t|^3 \sum \cdot 1 + t^2 \sum \mathbb{E}(X_k^2 : |X_k| > \epsilon). \end{aligned}$$

By the Lindeberg condition,

$$\limsup_n \sum |\varphi_k(t) - (1 - \frac{1}{2}t^2\sigma_k^2)| \le \epsilon|t|^3.$$

Since ϵ is arbitrary, the limit exists and is zero:

$$\sum_{k=1}^{n} |\varphi_k(t) - (1 - \frac{1}{2}t^2\sigma_k^2)| \to 0 \text{ as } n \to \infty$$

for every t.

Now, why is this important? Not only do we know an individual estimate, but we have a collective bound, that the sum goes to zero. Why is the sum important? Remember that the error in the product is bounded by the difference of the sum of errors. In particular,

$$|\prod \varphi_k - \prod (1 - \frac{1}{2}t^2\sigma_k^2)| \to 0.$$

So we use this pivotal fact to bound errors in products. We see the light at the end of the channel and have proved the CLT today! □

This theorem is one of the most technical things we do.

FEBRUARY 13, 2008

We're halfway through the proof of the Lindeberg Central Limit Theorem.

Recall the *Lindeberg Condition*: (X_n) independent random variables with mean 0 and variances σ_n^2, and define

$$s_n^2 = \sum_{k=1}^{n} \sigma_k^2 = \sum_{k=1}^{n} \mathbb{E}X_k^2.$$

The collection (X_n) *satisfy Lindeberg's Condition* if

$$\sum_{k=1}^{n} \mathbb{E}(X_k^2 : |X_k| > \epsilon s_n) = o(s_n^2) \text{ as } n \to \infty.$$

This condition always holds for $n = 1$ by the fact from last time. So, this is is interesting for $n \to \infty$. When chopping the random variables, do you have a good effect overall, a uniform condition?

Then, the Central Limit Theorem stated that X_n satisfying the Lindeberg Condition implies that $\frac{S_n}{s_n} \to N(0, 1)$. We started to prove this.

1. We showed that even between

$$S_n = X_1 + \cdots + X_n$$
$$S_{n+1} = X_1 + \cdots + X_{n+1}$$

we can permute the $X_1, \ldots, X_n$. So technically, $X_k \to X_{nk}$. This was the triangular array. So, we could, without loss of generality, assume that $s_n = 1$. At every step, we just normalize by the variance. That was the first step.

2. We tried to mimick the proof of the previous Central Limit Theorem. We took characteristic functions and wrote out a two-term approximation for them. Each characteristic function looked like

$$\varphi_k(t) \sim 1 + 0 + i^2 t^2 X^2/2 = 1 - \frac{1}{2} t^2 \sigma_k^2 \text{ as } t \to 0.$$

Not only did we know this, but we knew that the sum of errors is fine:

$$\sum_{k=1}^{n} |\varphi_k(t) - 1 + \frac{1}{2} t^2 \sigma_k^2| \to 0 \text{ as } n \to \infty, \text{ for every } t. \tag{34}$$

By the Lemma 22.10, this sum (on the LHS in (34)) is almost what we need: it almost follows that

$$\left| \prod_{k=1}^{n} \varphi_k(t) - \prod_{k=1}^{n} (1 - \frac{1}{2} t^2 \sigma_k^2) \right| = o(1) \text{ as } n \to \infty.$$

Let us call this our claim. But why almost, and not exactly? The Lemma requires that the modulus is 1. This is clear for the elements in the first term, since $|\varphi_k(t)| \leq 1$. However, t may be large, and to complete the proof of our claim, we need to check that $|\frac{1}{2} t^2 \sigma_k^2| \leq 1$, which will follow for large n if we can prove that

$$\max_{1 \leq k \leq n} \sigma_k \to 0 \text{ as } n \to \infty. \tag{35}$$

Here, this is a slight abuse of notation, because we mean to have

$$\max_{1 \leq k \leq n} \sigma_{nk} \to 0 \text{ as } n \to \infty.$$

If we can prove (35), then no matter how large t gets, we still have an estimate on the modulus. This almost follows from the Lindeberg assumption. Let's separate out small and large parts in considering

$$\begin{aligned} \sigma_k^2 = \mathbb{E} X_k^2 &= \mathbb{E}(X_k^2 : |X_k| \leq \epsilon) + \mathbb{E}(X_k^2 : |X_k| > \epsilon) \\ &\leq \epsilon^2 + \mathbb{E}(X_k^2 : |X_k| > \epsilon). \end{aligned}$$

By Lindeberg's Condition, and because $\epsilon > 0$ is arbitrary, (35) follows. This is the first time we're using the Lindeberg condition. This is the usual two-step limit argument. Thus, our claim is proved.

This is great, because we can get away from characteristic functions and use asymptotic expressions. The next claim (Claim 2) is that the asymptotic expression is close to what we need. Claim:

$$\left| \prod_{k=1}^{n} (1 - \frac{1}{2} t^2 \sigma_k^2) - \prod_{k=1}^{n} e^{-t^2 \sigma_k^2/2} \right| = o(1) \text{ as } n \to \infty.$$

Thus, $1 - x \sim e^{-x}$. This is almost true. What else can we do other than to apply Lemma 22.10 again? We get that the LHS in this claim is bounded:

$$\text{LHS} \leq \sum_{k=1}^{n} |e^{-t^2\sigma_k^2/2} - 1 + \frac{1}{2}t^2\sigma_k^2|.$$

We know that the $e^{-t^2\sigma_k^2/2}$ are individually small (close to $1 - x$). But, how small? So, let's take a number $x \in \mathbb{R}$, and let's look at the quality of the approximation:

$$|e^x - 1 - x| = \left|\sum_{k=2}^{\infty} \frac{x^k}{k!}\right| \leq x^2 \sum_{k=2}^{\infty} \frac{|x|^{k-2}}{k!} \leq \sum_{k=2}^{\infty} \frac{|x|^{k-2}}{(k-2)!} = x^2 e^{|x|}.$$

This is bad for large x, but we will only use this for small x, for $x = t^2\sigma_k^2/2$. We are aiming at bounding every term in this sum:

$$|e^{-t^2\sigma_k^2/2} - 1 + \frac{1}{2}t^2\sigma_k^2| \leq t^2\sigma_k^4 e^{t^2\sigma_k^2/2}$$

Since $\sigma_k^2/2 \leq 1$ for large n, and t is a constant as well, the LHS in Claim 2 is bounded:

$$\text{LHS} \leq \sum_{k=1}^{n} t^4 e^{t^2} \sigma_k^4 \leq t^4 e^{t^2} \left(\max_{1\leq k\leq n} \sigma_k^2\right) \cdot \sum_{k=1}^{n} \sigma_k^2$$

By (35), $\left(\max_{1\leq k\leq n} \sigma_k^2\right) \to 0$ and $\sum_{k=1}^{n} \sigma_k^2 = s_n = 1$, so the LHS $\to 0$ as $n \to \infty$. This prove Claim 2. The claims together imply

$$\prod_{k=1}^{n} \varphi_k(t) = \prod_{k=1}^{n} e^{-t^2\sigma_k^2/2} + o(1) = e^{-t^2/2} + o(1).$$

The left hand side is $\varphi_{S_n}(t)$. The right hand side is $\varphi_N(t) + o(1)$. Thus,

$$\varphi_{S_n}(t) \to \varphi_N(t) + o(1) \text{ as } n \to \infty.$$

By the Continuity Theorem, the CLT is proved. □

Don't worry if you don't understand this the first time you see it. The key is that we are using the Continuity Theorem and keeping track of all errors in Taylor estimation.

Remark 22.27. *If the Central Limit Theorem holds, it is the case that Lindeberg Condition "almost" holds. You can look up what the specific situation is. In your mind, you can almost believe that the Lindeberg Condition is almost logically equivalent to the Central Limit Theorem condition.*

In practice, the Lindeberg condition is difficult, so we propose a new condition:

Definition 22.28 (Lyapunov's Condition)**.** *A sequence (X_n) of independent random variables with mean 0 and variances (σ_n^2) is said to satisfy Lyapunov's Condition if $\exists\ p > 2$ such that $\mathbb{E}|X_n|^p < \infty$, and*

$$\left(\sum_{k=1}^{n} \mathbb{E}|X_k|^p\right)^{1/p} = o\left(\sum_{k=1}^{n} \mathbb{E}|X_k|^2)\right)^{1/2} = o(s_n) \text{ as } n \to \infty.$$

Note that one always has $(\sum |a_k|^p)^{1/p} \leq (\sum |a_k|^q)^{1/q}$ for any real numbers (a_k) and $p \geq q$. This is basically a consequence of the Hölder Inequality. This says that the p norm is smaller than the q norm. The way I remember the direction of the inequality is by thinking for the case $p = \infty, q = 1$.

So, one always has this inequality, but you may have the equality case: $(a_k) = (1, 0, 0, 0, \ldots)$ for sparse sequences. How can we expect the Central Limit Theorem to hold when we have basically all zeroes? Heuristically, if (a_k) is not sparse (if a_k is "spread") then

$$\left(\sum |a_k|^p\right)^{1/p} = o\left(\sum |a_k|^q\right)^{1/q}.$$

This sort of guarantees that all the random variables are "basically the same."

It is very simple to prove that in this case, the Central Limit Theorem holds:

Proposition 22.29. *Lyapunov's Condition implies Lindeberg's Condition.*

Proof. Let's assume Lyapunov's Condition. We are trying to check the Lindeberg Condition. If we are in the situation that $|X_k| > \epsilon s_n$, then

$$X_k^2 \leq \frac{|X_k|^p}{(\epsilon s_n)^{p-2}}$$

is estimated using a higher moment. (The easy way to check the correctness of the above is to cross-multiply.)

To check the Lindeberg Condition,

$$\frac{1}{s_n^2} \sum_{k=1}^{n} \mathbb{E}(X_k^2 : |X_k| > \epsilon s_n) \leq \frac{1}{\epsilon^{p-2} s_n^p} \sum_{k=1}^{n} \mathbb{E}|X_k|^p,$$

and ϵ is fixed, so this goes to 0, by Lyapunov's Condition (except we haven't taken pth root). □

FEBRUARY 15, 2008

I graded you exam. Your exam will sit in my office, since I have yet to enter grades into the system. You can have a look at your midterm. You can make a copy of your midterm. When I contemplate your final grade, sometimes it's useful to look at your midterm.

There was actually one interesting mistake that I didn't think of. You had to define for yourself that $\nu_n \to \infty$ in probability. Let's make an analogous statement for reals. When you think a sequence a_n of reals converges to $a \in \mathbb{R}$, we know this is $\forall \epsilon$, $|a_n - a| < \epsilon$ for large n. So, it may be false at first, but it's a sure event eventually. Now, in probability, $\nu_n \to \nu$ in probability if $\forall \epsilon$, $\mathbb{P}(|\nu_n - \nu| < \epsilon) \to 1$.

Now what about ∞? What's a neighborhood of infinity? It's (b, ∞), for $b \in \mathbb{N}$. So, the statement in reals is $a_n \to \infty$ in probability if $\forall M$, $a_n > M$ for large n. In the probability setting, then $\nu_n \to \infty$ if $\forall \mu, \mathbb{P}(\nu_n > M) \to 1$. Many of you thought, for some reason, to say $\frac{\nu_n}{n} \to 1$ a.s., or something. What if our sequence goes to n^2, for example? This just says that ν_n goes to ∞ at a linear rate.

Today, we'll do the replacement method of Lyapunov. I'll mostly follow the new textbook by Khohnevisan. Over the next couple lectures, we'll try to understand quantitative versions of the Central Limit Theorem. The CLT states that normalized sums converge to the Gaussian. In practice, you are not satisfied with this. How fast does the error go to zero? Just give me some numbers. I want an error ϵ between the distributions, then how many samples do I have to take? There are lots of quantitative versions of the Central Limit Theorem, and they are all about error bounds.

We will do a couple of them. One of this is this replacement method by Lyapunov. It is an alternative proof of the Central Limit theorem, and it will give you quantitative data.

The idea is simple: we will estimate the distribution of the sum $S_n = X_1 + \cdots + X_n$ by replacing its increments X_k one at a time by independent normal random variables Z_k. I want to replace X_n by Z_n and keep track of the error. Then I want to replace X_{n-1} by Z_{n-1} and add to the error. But then in the end, I'll have a sum of normal random variables. I only have to keep track of the errors. How far do we go from S_n by replacing one of the X_i with Z_i? This is basically the idea. It doesn't go through the theory of characteristic functions.

Theorem 22.30. *Let (X_n) be independent random variables with mean 0, and finite third moments*[83]*. Let $S_n = X_1 + \cdots + X_n$. Its variance is s_n^2. Then the following holds*[84]*: Then, for $f \in C^3(\mathbb{R})$,*

$$|\mathbb{E}f(S_n) - \mathbb{E}f(N)| \leq CM_f \sum_{k=1}^{n} \mathbb{E}|X_k|^3,$$

where N is distributed $N(0, s_n^2)$, and $M_f = \sup_z |f'''(z)|$.

This is a non-limiting result: it's just a genuine inequality. This is provided that M_f is finite.

[83] This is a weak assumption: it looks a little bit like Lypanov's condition. You can actually replace this with $2 + \epsilon$ moments, but this will be more convenient for proof.

[84] How do we estimate the distance? There are several ways to do this, and one way is to say that $X_n \to X$. Recall, this is equivalent to $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded and continuous f. This is a sequence of *numbers*. So, we'll go with the right hand side.

Remark 22.31. *The best case to understand this is for i.i.d. random variables Z_k with mean 0 and variance σ^2. Let $S_n = \frac{Z_1+\cdots+Z_n}{\sigma\sqrt{n}}$. This is a normalized sum, and $s_n = \sigma\sqrt{n}$. Then $X_k = \frac{Z_k}{\sigma\sqrt{n}}$. Then, in the theorem, the RHS is n identical terms, the third moment of X_k (all the same). We have*

$$RHS \leq CM_f \cdot n\frac{\mathbb{E}|Z_1|^3}{\sigma^3 n^{3/2}} = O\left(\frac{1}{\sqrt{n}}\right).$$

So this error size $\frac{1}{\sqrt{n}}$ is remarkable. Then, an exercise is to derive from this the Central Limit Theorem. The little problem is that you only have f is only three-times differentiable, but we need some smoothing trick here.

This is a non-asymptotic version of the Central Limit Theorem. It's just an inequality, and it tells you an error. One thinks (at first) the bound is not so good, because it's a third moment. After normalization, all of this will become very small.

We will replace random variables one at a time, and keep track of errors. Before the proof, we'll need some conventional notation from analysis. People in analysis will denote $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ for a random variable X. The reason for this is that it is a norm in L^p. One of the consequences of the Hölder inequality is a monotonicity property:

$$\|X\|_p \leq \|X\|_q \text{ for } p \leq q.$$

One way to remember the direction is to use $p = 1$ and $q = \infty$. Then, we compare expectation of X and maximum of X. This is all we need for the proof.

We know the function somewhere, and we estimate the function somewhere else by using Taylor's expansion.

Proof. Using Taylor's expansion,

$$f(x_0 + x) = f(x_0) + xf'(x_0) + \frac{x^2}{2}f''(x_0) + r(x),$$

where $|r(x)| \leq \frac{M_f}{6}|x|^3$. Use this for $x_0 = S_{n-1}$ and $x = X_n$. What we get is

$$f(S_n) = f(S_{n-1}) + X_n f'(S_{n-1}) + \frac{X_n^2}{2}f''(S_{n-1}) + r(X_n),$$

where

$$|r(X_n)| \leq \frac{M_f}{6}|X_n|^3.$$

This holds true for any fixed realization of random variables. Now we take the expectation. The linear term will disappear because of mean zero: $\mathbb{E}X_n = 0$, and $\mathbb{E}X_n^2 = \sigma_n^2$. By independence[85], what we get is

$$\mathbb{E}f(S_n) = \mathbb{E}f(S_{n-1}) + \frac{\sigma_n^2}{2}\mathbb{E}f''(S_{n-1}) + R,$$

[85] S_{n-1} and X_n are independent: This is used for the quadratic term.

where $|R| \leq \frac{M_f}{6}(\|X_n\|_3)^3$. This holds true for every distribution[86] of X_n. In particular, it will hold true if we replace X_n by Z_n, which is $N(0, \sigma_n^2)$, and we note Z_n has third moments. Then,

$$\mathbb{E}(S_{n-1} + Z_n) = \mathbb{E}f(S_{n-1}) + \frac{\sigma_n^2}{2}\mathbb{E}f''(S_{n-1}) + R',$$

with

$$|R'| \leq \frac{M_f}{6}\|Z_n\|_3^3. \tag{36}$$

The RHS did not change (essentially, except for the R term), because it only dealt with S_{n-1}, and not X_n. So, what is the effect before and after replacement? Let's compare:

$$|\mathbb{E}f(S_n) - \mathbb{E}f(S_{n-1} + Z_n)| = |R - R'| \leq |R| + |R'|. \tag{37}$$

If in (36), Z_n is replaced by X_n, we'd be done. So, we claim that basically we're done:

Claim 22.32. $\|Z_n\|_3 \leq C\|X_n\|_3$.

Proof of the Claim. First, we divide by σ_n. Then Z_n/σ_n is $N(0,1)$. Set $C = \|Z_n/\sigma_n\|_3$. Hence,

$$\|Z_n\|_3 \leq C\sigma_n = C\|X_n\|_2 \leq C\|X_n\|_3,$$

by Hölder's inequality. □

With the claim proved, the RHS in (37) is $\leq C'|R| = C''M_f\|X_n\|_3^3$.
Here's what we proved:

$$|\mathbb{E}f(S_n) - \mathbb{E}f(S_{n-1} + Z_n)| \leq C''M_f\|X_n\|_3^3.$$

We keep doing this iteratively, or formally, we apply induction. Then,

$$\mathbb{E}f(S_n) - \mathbb{E}f(\sum_{k=1}^{n} Z_k)| \leq C''M_f \sum_{k=1}^{n} \|X_k\|_3^3.$$

Replace one at a time and use the triangle inequality.

Because Z_k is $N(0, \sigma_k^2)$ independent, we get

$$\sum_{k=1}^{n} Z_k$$

is $N(0, \sum_{k=1}^{n} \sigma_k^2) = N(0, s_n^2)$, which finishes the proof of the theorem. □

[86] Now, comes the replacement idea.

This is a nice idea: replace one at a time with normals, and use Taylor's expansion.

The exercise again is to get the familiar form of the CLT from this statement. What is the difference of the probabilities between events on S_n and events on N?. I would suggest you use functions f that are indicators, but f is discontinuous, so if you smooth it a little bit. You only lose a little bit in the energy of the Gaussian. So, instead of the expectations, it would be good to get

$$\mathbb{P}(a \leq S_n \leq b) - \mathbb{P}(a < N \leq b)| \leq ?$$

FEBRUARY 20, 2008

Today, I will give you (without proof) one theorem that is slightly stronger than the replacement method of Lyapunov. It has a long proof, but it's simple to believe it.

22.5 Berry-Esseen Theorem

This corresponds to the Replacement Method of Lyapunov for $f = \mathbf{1}_{(-\infty,a]}$. The way we estimated distance between distributions was to compute

$$|\mathbb{E}f(S_n) - \mathbb{E}f(N)|. \tag{38}$$

For the CLT to hold, we needed this to $\to 0$. The Replacement Method gave us an upper bound on (38). We would like to apply this to indicator functions, but we had the requirement that $f \in C^3(\mathbb{R})$. This theorem corrects this problem:

Theorem 22.33 (Berry-Esseen Theorem)**.** *Let (X_k) be independent random variables with mean zero and finite third moments. Consider $S_n = X_1 + \cdots + X_n$. Define*

$$s_n := \operatorname{var}(S_n) = \sum_{k=1}^{n} \sigma_k^2 = \sum_{k=1}^{n} \mathbb{E}X_k^2.$$

Then,

$$\sup_x \left| \mathbb{P}\left(\frac{S_n}{s_n} \leq x\right) - \mathbb{P}(N \leq x) \right| \leq \frac{C \sum_{k=1}^{n} \mathbb{E}|X_k|^3}{s_n^3},$$

and C is an absolute constant.

So, if you could apply the Replacement Method of Lyapunov for indicator functions, then this is what you get.

Example 22.34. *The basic example for which we want to understand this is for (X_k) are i.i.d., with $\mathbb{E}X_k^2 = \sigma^2$ and $\mathbb{E}|X_k|^3 = \gamma^3$. Then, $s_n = \sigma\sqrt{n}$. Thus,*

$$\sup_x \left| \mathbb{P}\left(\frac{S_n}{s_n} \leq x\right) - \mathbb{P}(N \leq x) \right| \leq \frac{C\gamma^3 n}{\sigma^3 n^{3/2}} = \frac{C\gamma^3}{\sigma^3} \cdot \frac{1}{\sqrt{n}}.$$

So the error in the CLT is $O\left(\frac{1}{\sqrt{n}}\right)$.

So, right from here we get the usual Central Limit Theorem, and we get a rate of convergence. The proof of the Theorem can be found, for example, in [Gut].

As we increase x, the error should reflect this, and the error should decrease.

Remark 22.35. *The RHS in the example above can be improved to*

$$\frac{C\gamma^3}{\sigma^3} \cdot \frac{1}{(1+|x|^3)\sqrt{n}}$$

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

In the example above,

$$\left| \mathbb{P}\left(\frac{S_n}{s_n} \leq x \right) - \mathbb{P}(N \leq x) \right| \leq \frac{C\gamma^3}{\sigma^3} \cdot \frac{1}{(1+|x|^3)\sqrt{n}}$$

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

This can be found in [Petrov: Sums of independent random variables].

Remark 22.36. *People invested a whole lot of time into finding the best known constant C. The best known is $C \leq 0.79$, according to [Gut].*

Remark 22.37. *In the Replacement Method of Lyapunov, and also in the Berry-Esseen Theorem, the third moments can be replaced by any $(2+\delta)$-moment, for any $\delta > 0$.*

For the CLT itself, no moment higher than 2 is necessary. But for more useful/quantative results, you'll need a stictly higher moment (see [Gut]).

22.6 Large Deviation Inequalities

What is so good about having a random variable close to the normal random variable, in practice? What's so spectacular about them? It settles down very quickly. Normal random variables have light tails. Namely,

$$\mathbb{P}(|N| > t) \sim e^{-t^2/2}, \quad t \in \mathbb{R}.$$

For example, if $t = 3$, then $1 - \mathbb{P}(|N| > t) = 99.7\%$.

So, in the spirit of the LLN, the sum of variables is near the mean. But, the CLT suggests the convergence to N very very quickly.

The tails being light is useful for estimating deviations of S_n from its mean. We can expect that

$$\mathbb{P}(|S_n/s_n| > t) \lesssim e^{-t^2/2}, \quad t \in \mathbb{R}. \tag{39}$$

This is our ideal. Maybe there is an extra numerical constant in front. We do not know this:

However, the Berry-Esseen Central Limit Theorem is much weaker, because of the error $O(1/\sqrt{n})$. We can only expect that

$$\mathbb{P}(|S_n/s_n| > t) \lesssim e^{-t^2/2} + \frac{c}{\sqrt{n}}.$$

This is too bad. We had a small tail before, but now after applying the CLT, we might have a large tail. A tail whose mass is even larger than the error.

So here comes the Theory of Large Deviations. We don't care about normal approximation. We will no longer compare $\frac{S_n}{s_n}$ to N. We want to estimate the tails of $\frac{S_n}{s_n}$ directly. Our ideal is (39). To prove this through the CLT won't work. For (39) to be true in general, one will need higher moments, because the tail is so good: It is necessary that $\frac{S_n}{s_n}$ has all moments. Is it clear why?

$$\mathbb{E}|X|^p = \int_0^p |t|^p \ \mathbb{P}(|X| > t) \ dt.$$

Theorem 22.38. *Let (X_k) be independent random variables with mean zero, and $|X_k| \leq a_k$. (In particular, we have all moments.) Let $a^2 = \sum_{k=1}^n a_k^2$. Consider their sum $S_n = X_1 + \cdots + X_n$. Then,*

$$\mathbb{P}(|S_n| > t) \leq 2e^{-t^2/2a^2}, t > 0.$$

The presence of a^2 on the right hand side is only for normalization.

Example 22.39. *Think about $a = 1$. $S_n = \sum_{k=1}^n a_k Z_k$, with Z_k being a $-1, 1$ Bernoulli random variable. So, S_n is a weighted sum of Bernoulli's. Then* $\mathrm{var}(S_n) = a = 1$. *Here, the consequence of this theorem is that*

$$\mathbb{P}(|S_n| > t) \leq 2e^{-t^2/2}, t > 0.$$

The RHS doesn't even have an n in it! The CLT will fail for a finite number of random variables, yet this deviation inequality will hold.

In practice, we draw intuition from the CLT to see what is the proposed limit. But we base our rigorous conclusions on deviation inequalities.

So, we will prove this theorem:

Proof. The proof uses the *moment generating function* M of a random variable X, which is defined (see [Billingsley] or [Gut]) to be

$$M(\lambda) = \mathbb{E}e^{\lambda X}, \lambda \in \mathbb{R}.$$

We will use this fact: If X has mean zero and $|X| \leq 1$, then

$$M(\lambda) \leq e^{\lambda^2/2}.$$

This will be an exercise in the HW. You might use Taylor's Expansion, up to three terms.

How do we use this? We start with Chebychev's Inequality: $\mathbb{P}(|X| > t) \leq \frac{\mathbb{E}|X|}{t}$. This is too weak, because of the polynomial decay, instead of super-exponential decay. So, instead, multiply both sides by λ, and then exponentiate. We take $\lambda > 0$. Then use Chebychev's inequality:

$$\mathbb{P}(X > t) = \mathbb{P}(e^{\lambda X} > e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}e^{\lambda X}. \tag{40}$$

This is already better, because we have an exponential tail. We will optimize the choice of λ to be proportional to t later.

We will use this for S_n. Since S_n is a sum, e^{S_n} will factor into a product. The expectation will factor. So in the end, we'll end up with moment generating functions of individual random variables.

Using the fact,

$$\mathbb{E}e^{\lambda X_k} = \mathbb{E}e^{(\lambda a_k)(X_k/a_k)} \leq \mathbb{E}e^{\lambda^2 a_k^2/2}.$$

Now we estimate the moment generating function of the sum:

$$\mathbb{E}e^{\lambda S_n} = \mathbb{E}\prod_{k=1}^{n} e^{\lambda X_k} \overset{indp}{=} \prod_{k=1}^{n} \mathbb{E}e^{\lambda X_k} \leq \prod_{k=1}^{n} e^{\lambda^2 a_k^2/2} = e^{\lambda^2 a^2/2}.$$

By (40),

$$\mathbb{P}(S_n > t) \leq e^{-\lambda t + \lambda^2 a^2/2}.$$

Now λ was a parameter, and now we optimize. We optimize λ to make $\lambda t - \lambda^2 a^2/2$ maximal. So we differentiate to get

$$t - \lambda a^2/0,$$

so $\lambda = t/a^2$. So the maximal value is $\frac{t^2}{a^2} - \frac{t^2}{a^2}\frac{a^2}{2} = \frac{t^2}{2a^2}$.

So, the conclusion is

$$\mathbb{P}(S_n > t) \leq e^{-t^2/2a^2},$$

which is precisely what we want.

We repeat the argument for $-S_n$ to complete the proof. □

FEBRUARY 22, 2008

One remark to finish last time's large deviation inequalities. In the literature, you'll find large deviation inequalities under the names

- Berstein Inequality. This is more general. Check out Wikipedia.
- Chernoff Inequality. This is for Bernoulli random variables
- Prokhorov-Bennett Inequality. The most general. It covers both normal and Poisson limits.

Recall what we did, reflected very much the Berstein Inequality. There are inequalities that interpolate between Poisson and normal random variables. There is a very large body of literature on large deviation limits. I thought I'd just give you some names to help you in reading.

We will start going towards Central Limit Theorem in higher dimensions.

23 Limit Theorems in $\mathbb{R}^d$

They are important for two reasons: Geometry and Multivariate. Ninety percent of the theory is appropriate modifications of the one-dimensional case. It will be useful to remind you the 1-dim case. Occasionally, there will be some interesting facts.

23.1 Review of Theory of Random Vectors

We will consider now a random vector X with values in $\mathbb{R}^d$ (rather than $\mathbb{R}^1$). The distribution of X is given by the values $\mathbb{P}(X \in A)$, where $A \subset \mathbb{R}^d$ is Borel. The distribution function defines the distribution of X. The distribution $F : \mathbb{R}^d \to [0, 1]$ is defined as

$$F(x) = \mathbb{P}(X \leq x).$$

It's enough to know values on a basis. In higher dimension, we need to interpret what it means for one vector to be smaller than another. Here, we mean coordinate-wise:

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d).$$

The random variable X has density $f : \mathbb{R}^n \to \mathbb{R}_+$ if

$$F(x) = \int_{-\infty}^{x} f(y)\ dy.$$

This was the one-dimensional case. So, now we know, from our coordinate-wise partial order, that this should mean (in d dimensions)

$$\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(y)\ dy_d \cdots dy_1.$$

Let $X, X_1, X_2, \ldots$ be random vectors of the same dimension. We say that a sequence $X_n \to X$ in distribution if

$$F_n(x) \to F(x) \text{ for all points of continuity } x \text{ of } F.$$

23.2 Analogous results

There is the same basic theory for random vectors:

Theorem 23.1 (Continuous Mapping Theorem)**.** *(a) If f is a continuous function and $X_n \to X$ in distribution, then $f(X_n) \to f(X)$ in distribution.*

(b) Even if f is not continuous, but

$$\mathbb{P}(X \in D_f) = 0$$

where D_f is the set of discontinuities of f, then

$$f(X_n) \to f(X) \text{ in distribution.}$$

It holds without any change for random vectors, and the proof is the same. Helly's Selection Theorem holds also. The proof is similar.

23.3 Characteristic functions

The interesting part happens when we look at characteristic functions. The *characteristic function* of a random vector X is $\varphi : \mathbb{R}^d \to \mathbb{R}$ given by the rule

$$\varphi(t) = \mathbb{E}e^{i\langle t, X\rangle}$$

where $t \cdot X = \langle t, X\rangle = \sum_{k=1}^{d} t_k X_k$.

A variant of the Inversion Formula holds, and in particular, it implies the Uniqueness Theorem.

Theorem 23.2 (Uniqueness Theorem)**.** *The characteristic function function determines the distribution of a random variable uniquely.*

Also, a variant of the Convergence Theorem holds:

Theorem 23.3 (Convergence Theorem)**.** *$X_n \to X$ in distribution iff $\varphi_n(t) \to \varphi(t)$ for every $t \in \mathbb{R}^d$.*

In one direction, this is easy. If $X_n \to X$ in distribution, then then the Continuous Mapping Theorem will give us what we need. We'll have convergence almost surely (after applying the Skorohod Representation Theorem). In the opposite direction, it was not easy, as you recall. We show the sequence is tight, and then used Helly's Selection Theorem.

Thus, there are no surprises so far. Everything is like the one-dimensional case. The surprise will come now. The following has no analogue in one dimension.

23.4 Cramer-Wold Device

This is a way to reduce higher-dimensional probabilistic claims to one dimension. It is a way to reduce it right away from d to 1.

So here is one remarkable theorem that is not quite the Cramer-Wold Device, but it's nice, and has all the flavor of it:

Theorem 23.4. *The distribution of a random vector X in $\mathbb{R}^d$ is uniquely determined by $\mathbb{P}(X \in H)$, where H is a half-space.*

Equivalently, a positive measure μ on $\mathbb{R}^d$ is uniquely determined by its values $\mu(H)$ on halfspaces H. This is a remarkable result! There is no elementary proof of this known. Modulo what we already know, the proof is simple, though.

Proof. Define a linear functional $h_t : \mathbb{R}^d \to \mathbb{R}$ by $h_t(x) = t \cdot x, x \in \mathbb{R}^d$. Then all halfspaces have the form[87]

$$H_{t,\alpha} = \{x : h_t(x) \leq \alpha\} \text{ for some } t \in \mathbb{R}^d, \alpha \in \mathbb{R}.$$

Then $\mathbb{P}(X \in H_{t,\alpha}) = \mathbb{P}(h_t(X) \leq \alpha)$ is known. The RHS is much better, because we have a random variable (on the LHS, we have a random vector).

As a consequence: To know the probability $\mathbb{P}(X \in H_{t,\alpha}$ of all half-spaces is to know the distribution of the random variables $h_t(X)$ for every t. So, we know the characteristic functions

$$\varphi_{h_t(x)}(s) = \mathbb{E}e^{ish_t(X)} = \mathbb{E}e^{ist\cdot X} = \varphi_X(st) = \varphi_X((st_1, st_2, \ldots, st_d)).$$

So we know know the RHS (since we know the RHS). Using this for $s = 1$, I now know the characteristic function $\varphi_X(t)$ for every t. By the Uniqueness Theorem, we are done. We know the distribution of X. □

Every "we know" is clear: it means uniquely determined.

So this is the Cramer-Wold Device. We introduced a set to reduce to one dimension.

Theorem 23.5 (Cramer-Wold Device)**.** *Let (X_n) be a sequence of random vectors in $\mathbb{R}^d$. Then $X_n \to X$ in distribution iff the random variables*

$$t \cdot X_n \to t \cdot X \text{ in distribution for every } t \in \mathbb{R}^d.$$

This is the device. If we want to prove a higher-dimensional result, we can reduce to one-dimensional results.

Proof. We will leave necessity as an exercise. It should be very close in spirit to the proof of the first Cramer-Wold Device proof. So, it should follow by the continuity theorem.

[87] The equations of the form $t \cdot x = \alpha$ are hyperplanes.

For sufficiency: We will just check the characteristic functions. By the Convergence Theorem,

$$\varphi_{t\cdot X_n}(s) \to \varphi_{t\cdot X}(s) \text{ for every } s \in \mathbb{R}.$$

That is,

$$\mathbb{E}e^{ist\cdot X_n} \to \mathbb{E}e^{ist\cdot X}.$$

For $s = 1$, this means

$$\varphi_{X_n}(t) \to \varphi_X(t).$$

Since t was arbitrary, another application of the Convergence Theorem shows that $X_n \to X$ in distribution. □

FEBRUARY 25, 2008

23.5 Normal Distribution and CLT in $\mathbb{R}^d$

Today we'll do the Central Limit Theorem is higher dimensions.

A random vector $X = (X_1, \ldots, X_d)$ in $\mathbb{R}^d$ *has standard normal distribution* if its components X_k are independent standard normal random variables. We also call it the *Gaussian in* $\mathbb{R}^d$. We can compute the density, since we know the density of each coordinate, and because of the independence. Thus, the density

$$f_X(x) = \prod_{k=1}^{d} f_{X_k}(x_k) = \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} = \frac{1}{(2\pi)^{d/2}} e^{-\sum x_k^2/2} = \frac{1}{(2\pi)^{d/2}} e^{-|x|^2/2},$$

where $|x|$ denotes the Euclidean norm of $x \in \mathbb{R}^d$. Similarly, the characteristic function[88] is

$$\begin{aligned}
\varphi_X(t) &= \mathbb{E}e^{i\langle t, X\rangle} \\
&= \mathbb{E}e^{i\sum_{k=1}^{d} t_k X_k} \\
&= \mathbb{E}\prod_{k=1}^{d} e^{it_k X_k} \\
&= \prod_{k=1}^{d} \mathbb{E}e^{it_k X_k}, \text{ by independence} \\
&= \prod_{k=1}^{d} \varphi_{X_k}(t_k) \\
&= \prod_{k=1}^{d} e^{-t_k^2/2} \\
&= e^{-|t|^2/2}.
\end{aligned}$$

[88] Remember here that t is a vector.

for $t \in \mathbb{R}^d$.

Two remarkable things happened here. The density and the characteristic function only depended on the Euclidean norm, not the direction. We can call this a *rotation invariance*.

Definition 23.6. *A measure μ in $\mathbb{R}^d$ is rotationally invariant if for every orthogonal transformation U of $\mathbb{R}^d$, and every Borel set $B \subseteq \mathbb{R}^d$, if*

$$\mu(B) = \mu(UB).$$

Recall that a matrix U is *orthogonal* if $U'U = I$.

Corollary 23.7. *The standard normal distribution in $\mathbb{R}^d$ is rotationally invariant.*

The density doesn't depend on the direction.

Proof. We need to check that $\mathbb{P}(X \in UB) = \mathbb{P}(X \in B)$ is true for every B and U. This follows from

$$\mathbb{P}(X \in UB) = \int_{UB} f_X(x)\, dx.$$

We make the change of variable which changes UB to B. The determinant is 1, thus the integral is

$$\int_B f_X(Uy)\, dy.$$

But the density is rotationally invariant (that is $|Uy| = |y|$), so this is

$$\int_B f_X(y)\, dy,$$

and this is $\mathbb{P}(X \in B)$. □

So, the standard normal distribution is rotationally invariant. Thus the density is a body of revolution.

This has an interesting geometric consequence that would be hard to construct otherwise. It immediately follows from the rotational invariance:

Corollary 23.8. *There exists a rotationally invariant measure on the unit Euclidean sphere S^{d-1}.*

We have to adjust the definition of rotational invariance for spheres in the obvious way. It would be hard to get this result using Lesbesgue measure because of the rectangular shape. We already have a rotationally invariant measure on $\mathbb{R}^d$. Now we have to contract it.

Proof. The proof is by contraction of the standard normal distribution in $\mathbb{R}^d$ onto S^{d-1}.

Let X be a standard normal vector in $\mathbb{R}^d$. we consider $Z = \frac{X}{|X|}$. Then Z is a random vector with values in S^{d-1} and Z is rotationally invariant[89]: $\mathbb{P}(Z \in UB) = \mathbb{P}(\frac{X}{|X|} \in UB)$ (exercise, using the Law of Large Numbers). □

Now comes a little bit different part. If we want a Central Limit Theorem in $\mathbb{R}^d$, we need to normalize by the variance. As soon as we go from random variables to random vectors, there's no concept as variance anymore, since different components have different variances. We need to normalize in a different way, so we need the concept of *covariance*.

23.5.1 Covariance Matrix of a Random Vector

Let X be a random vector in $\mathbb{R}^d$. Its *covariance matrix* Σ is the $d \times d$ matrix with entries

$$\sigma_{ij} = \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j).$$

If $d = 1$, then Σ's only entry is the variance.

Proposition 23.9. *Σ is a symmetric and positive semidefinite[90] matrix.*

Proof. Symmetry is obvious. For the positive semidefiniteness, we can assume, without loss of generality, that the expectations $\mathbb{E}X_i$ are all zero, by translating our random variables, the coordinates of the random vector. Let $x \in \mathbb{R}^d$ be arbitrary. Then,

$$\langle x, \Sigma x \rangle = \sum_{i,j} \sigma_{ij} x_i x_j = \mathbb{E}\left(\sum_{i,j} X_i X_j x_i x_j \right) = \mathbb{E}\left(\sum_i x_i X_i \right)^2 \geq 0.$$

□

Example 23.10. *The basic example is when the components are independent. Let $X = (X_1, \ldots, X_d)$ be a random vector with independent components. Then*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_d^2 \end{bmatrix},$$

where σ_i^2 is the variance of X_i. So for independent random variables, we don't need a matrix.

When we studied random vectors, our study was to bring back problems to variance 1. We want to do the same thing here. We will invert the matrix.

[89] This is shorthand for the distribution of Z is rotationally invariant.

[90] A matrix Σ is *positive semidefinite* if $\langle x, \Sigma x \rangle \geq 0$ for all x.

Proposition 23.11. *Let X be a random vector in $\mathbb{R}^d$ with mean 0 and covariance matrix I. Then,*

1. *If A is a $d \times d$ matrix*[91]*, then the random vector $Y = AX$ has mean 0 and covariance matrix*
$$\Sigma = AA'.$$

2. *If Σ is a covariance matrix of some random vector, then*
$$\Sigma = AA'$$
for some $d \times d$ matrix A.

From the form $\Sigma = AA'$, it is clear that A is symmetric and positive semidefinite.

Proof. 1. $\sigma_{ij} = \mathbb{E}Y_iY_j = \mathbb{E}\langle A_i, X\rangle\langle A_j, X\rangle$, where A_i is the i^{th} row of the matrix A. This is equal to $\mathbb{E}(\sum_k A_{ik}X_k)(\sum_\ell A_{j\ell}X_\ell)$. We multiply all of them out. Of the d^2 terms, the off-diagonal ones are all zero. The only terms that survive are when $k = \ell$. So this is now equal to $\mathbb{E}\sum_k A_{ik}A_{jk}X_k^2 = \sum_k A_{ik}A_{jk}$, since the covariance matrix is the identity. And this is equal to $(AA')_{ij}$.

2. Then Σ is a symmetric and positive semidefinite matrix, by Proposition 23.9. Use the Polar[92] decomposition[93], that $\Sigma = U'DU$, where U is orthogonal and $D \geq 0$ is diagonal. Let $A = U\sqrt{D}$.
Then $\Sigma = (U'\sqrt{D})(\sqrt{D}U) = AA'$.

□

FEBRUARY 27, 2008

We are trying to prove the CLT in higher dimension. The first the thing we do is to describe what is the Normal distribution in $\mathbb{R}^d$.

23.6 Normal distribution in $\mathbb{R}^d$

We know what the standard normal is. It's a random vector whose components are independent standard normals. A normal is the linear image of a standard normal:

Definition 23.12. *A random vector Y in $\mathbb{R}^d$ has centered normal distribution if $Y = AX$, where A is some non-singular $d \times d$ matrix, and X is a standard normal random vector.*

[91] A linear transformation of this vector
[92] or Singular Value Decomposition
[93] For a positive semidefinite matrix M, the choice of the right basis shows that M stretches in certain directions, never rotates.

We always look at mean zero. This is the significance of the word "centered." A compeletly general normal distribution may have a non-zero translation (i.e., a non-zero mean) as well.

Once we have this, then, the covariance matrix is easy: $\Sigma = AA'$. This is by Proposition 23.11. Its characteristic ufnciton is

$$\varphi_Y(t) = \mathbb{E}e^{i\langle t, AX\rangle}.$$

We want to move A from the right side of $\langle\cdot,\cdot\rangle$ to the left side. Thus, this is equal (by using adjoint) to

$$= \mathbb{E}e^{i\langle A't, X\rangle} = \varphi_X(A't) = e^{-|A't|^2/2}.$$

Now, how do we compute $|A't|^2$?

$$|A't|^2 = \langle A't, A't\rangle = \langle t, AA't\rangle = \langle t, \Sigma t\rangle,$$

hence the characteristic function is

$$\varphi_Y(t) = e^{-\langle t, \Sigma t\rangle/2}.$$

Apart from being just a computation, this has an important consequence: the characteristic function is determined by Σ only. And the characteristic function determines the distribution uniquely. So, by the Uniqueness Theorem, we have the important corollary

Corollary 23.13. *A centered normal distribution in $\mathbb{R}^d$ is uniquely determined by its covariance matrix.*

Of course, it doesn't happen for other distributions. There are different distributions with identity covariance matrix. This is the analogue of the case in $\mathbb{R}^1$: when we know σ (and assume $\mu = 0$), we know exactly the Gaussian described.

Not only does this determine (as the corollary promises) the distribution of the centered normal, it can also be defined in this way.

Now, we need to compute the density, which will give us a third alternative definition of the centered normal. Recall the 1-dimensional case. The density of a centered normal random variable $N(0, \sigma^2)$ with variance σ^2 was

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-|x|^2/2\sigma^2}.$$

To check whether or not σ belongs under the radical or not, we can do a change of variables, with X being $N(0,1)$ and $Y = \sigma \cdot X$ being $N(0, \sigma^2)$. Then we need the lemma

Lemma 23.14 (Change of variable)**.** *Let X be a random vector in $\mathbb{R}^d$ with density $f(x)$. Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a one-to-one continuously differentiable map.*

Let $T = g^{-1}$, also one-to-one and continuously differentiable. Then, the density of the random variable $g(X)$ is

$$f(T(x)) \cdot |\det J(x)|,$$

where $J(x)$ is the Jacobian matrix of T.

What is the Jacobian matrix? Recall: write $T(x) = (T_1(x), \ldots, T_d(x))$, where $T_k : \mathbb{R}^d \to \mathbb{R}$. Then the Jacobian matrix is

$$J(x) = \begin{bmatrix} \frac{\partial T_1}{\partial x_1} & \cdots & \frac{\partial T_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_d}{\partial x_1} & \cdots & \frac{\partial T_d}{\partial x_d} \end{bmatrix}.$$

How do we prove this? With higher dimensional calculus:

Proof. For a Borel set A in $\mathbb{R}^d$,

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(X \in T(A)).$$

We know the density of X, so this is just the integral of the density, namely

$$\int_{T(A)} f(x)\, dx.$$

Now we change variables in the standard way in calculus. We take $x = Ty$. Then, this is equal to

$$\int_A f(Ty) |\det J(y)|\, dy.$$

The proof is complete, because the integrand in the previous expression must be the density for Y. □

In our case, we want to compute the density of a normal random variable. So, in our case, the vector $Y = AX$ is a linear image of the standard normal random vector X, and A is a matrix given by the covariance: $\Sigma = AA'$. So, by the Lemma, our g is a linear map, with $g = A$ and $T = A^{-1}$. Hence[94], $J(x) = A^{-1}$. Therefore, the density of Y is

$$f(A^{-1}x) |\det A^{-1}|, \tag{41}$$

where $f(x) = \frac{1}{(2\pi)^{d/2}} e^{-|x|^2/2}$ is the density of the standard normal.

So we need to compute $\det A^{-1}$. Our ultimate goal is to express everything in terms of Σ. Well, $\det(A^{-1}) = (\det A)^{-1}$. By the same multiplicativity property, $\det \Sigma = (\det A)(\det A') = (\det A)^2$. So

$$\det(A^{-1}) = (\det \Sigma)^{-1/2}.$$

[94] The Jacobian of a matrix is the matrix itself.

There is another A (namely the A^{-1} in the argument of f) to take care of in (41). What is $|A^{-1}x|^2$?

$$|A^{-1}x|^2 = \langle A^{-1}x, A^{-1}x\rangle = \langle x, (A^{-1})'A^{-1}x\rangle = \langle x, \Sigma^{-1}x\rangle.$$

So now we have everything in the formula (41). Putting this all together, we obtain

Proposition 23.15. *The density of the centered normal distribution in $\mathbb{R}^d$ with covariance matrix Σ is*

$$f(x) = \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} e^{-\langle x,\Sigma^{-1}x\rangle/2}.$$

All of the constant in front is the appropriate normalization to make the integral of f on all of $\mathbb{R}^d$ equal to 1. In $\langle x, \Sigma^{-1}x\rangle$, we are "distorting" the standard normal: Σ makes the stretches perhaps different in different directions.

In this Proposition, even in this, we implicitly use the Corollary. Here, we applied the Corollary just to state this.

Now, the main result we are heading to is the Central Limit Theorem in higher dimensions.

23.7 Central Limit Theorem in $\mathbb{R}^d$

Recall the one-dimensional Central Limit Theorem: for i.i.d. random variables X_k with mean 0 and variance σ^2, consider the sum $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n}{\sigma\sqrt{n}} \to N(0,1) \text{ in distribution.}$$

Now, in higher dimensions, we can not have Σ in the denominator, since it is a matrix. Thus, we must state into some kind of reasonable form, and so we'll provide the analogue for $\mathbb{R}^1$ as

$$\frac{S_n}{\sqrt{n}} \to N(0,\sigma^2) \text{ in distribution.}$$

So, this rewording can be generalized, since we know what will take the role of σ here.

Theorem 23.16 (Central Limit Theorem in $\mathbb{R}^d$)**.** *Let (X_n) be independent and identically distributed centered random vectors in $\mathbb{R}^d$ with covariance matrix Σ. Then the sums $S_n = X_1 + \cdots + X_n$ satisfy*

$$\frac{S_n}{\sqrt{n}} \to N \text{ in distribution,}$$

N is a centered normal random vector in $\mathbb{R}^d$ with covariance matrix Σ.

The proof of this is simple modulo the Cramer-Wold Device.

Proof. It suffices to show that if we hit the left and right side by any fixed vector t, then we have proved the result. Thus, we wish to show that for all $t \in \mathbb{R}^d$, one has

$$\langle t, S_n/\sqrt{n}\rangle \to \langle t, N\rangle$$

in distribution. Note, these are scalars.

We can write the LHS as a sum of i.i.d random variables:

$$\langle t, \frac{S_n}{\sqrt{n}}\rangle = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \langle t, X_k\rangle. \tag{42}$$

We use the one-dimensional Central Limit Theorem for this sum. We first need to check the variance (since we've only normalized by $\sqrt{n}$ and not $\sigma\sqrt{n}$. So we check

$$\mathbb{E}\langle t, X_k\rangle^2 = \mathbb{E}\left(\sum_{j=1}^{d} t_j X_{kj}\right)^2 = \mathbb{E}\sum_{i,j} t_i t_j X_{ki} X_{kj},$$

When we bring the expectation inside here[95]

$$\sum_{t,j} t_i t_j \sigma_{ij},$$

where $\Sigma = (\sigma_{ij})$. This is

$$= \langle t, \Sigma t\rangle.$$

For a similar reason, the variance $\mathbb{E}\langle t, N\rangle^2$ is also $\langle t, \Sigma t\rangle$. Thus the variances agree. Then, by the one-dimensional CLT and (42),

$$\langle t, \frac{S_n}{\sqrt{n}}\rangle \to \langle t, N\rangle \text{ in distribution.}$$

We are done by the Cramer-Wold Device. □

This completes the big part of the course. In what remains, we'll start on conditional expectation and the theory of martingales, going as far as we can.

FEBRUARY 29, 2008

I've updated the home work 7, problem number 5. Today we are starting a big section on conditional expectation.

24 Condition Expectation

I will roughly follow [Durrett, 4.1]. When you talk about conditional expectation in probability theory, it gets very high-level. If you haven't seen it before, it might be difficult to get used to.

[95] The random vectors are independent, but the coordinates inside each are not necessarily.

So, what is the problem? We know how to define conditional probability: for events A and G,

$$\mathbb{P}(A|G) = \frac{\mathbb{P}(A \cap G)}{\mathbb{P}(G)}.$$

This may be interpreted as the probability that A occurs given the information that some other event G occurs: If we know that G occurs, we want to stay only in G. But, G itself is not a probability space. The division by $\mathbb{P}(G)$ is for normalization.

Using this, the best idea to define conditional expectation of a random variable is as follows: The *conditional expectation* of a random variable X given an event G is[96]

$$\mathbb{E}(X|G) = \frac{1}{\mathbb{P}(G)}\mathbb{E}(X\mathbf{1}_G) = \frac{1}{\mathbb{P}(G)}\int_G X \, d\mathbb{P} = \frac{1}{\mathbb{P}(G)}\int_G X(\omega) \, d\mathbb{P}(\omega).$$

We normalize the same way since we want to stay in G.

This is fine. Everything is okay, except when $\mathbb{P}(G) = 0$. This is not defined when $\mathbb{P}(G) = 0$. Who cares about these events that are very small?

For example, suppose X and Y are random variables with densities. There might be some correlation between them. If we do not know Y, the best guess about X is $\mathbb{E}X$. If we know the value of Y, say $Y = y$, our best guess about X given the information $Y = y$ would be

$$\mathbb{E}(X|Y = y).$$

But, because Y has density, the probability of this event $Y = y$ is zero. So, this conditional expectation is not defined. So, to treat this case in particular, we will need a rich theory of conditional expectation. This will also affect our theory of conditional probability. The theory will be based on this miracle theorem in Lesbesgue mesaure, which we will not prove called the Radon-Nikodym Theorem.

This develops the motivation. We will give a definition of expectation given a σ-algebra. This particular case $\mathbb{E}(X|G)$ will occur when the σ-algebra has just two non-trivial[97] events: just G and G^c.

24.1 Conditional expectation given a σ-algebra

We will consider probability spaces $(\Omega, \mathcal{F}), \mathbb{P})$ with different $\mathcal{F}$. The idea of varying a σ-algebra is very natural here: As soon as you want to take expectation with respect to G, you are down to the σ-algebra generated by G.

A random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is called *$\mathcal{F}$-measurable*.

Definition 24.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose X is an $\mathcal{F}$-measurable random variable. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra*[98]*. The conditional expectation of X given $\mathcal{G}$, denoted $\mathbb{E}(X|\mathcal{G})$ is a random variable Y such that:*

[96] So far, this should conceptually make sense what we are trying to ask here.

[97] The trivial events are $\emptyset$ and the whole probability space.

[98] $\mathcal{G}$ is a coarser σ-algebra. For example, $\mathcal{G}$ may be generated by two events.

(i) Y *is* $\mathcal{G}$*-measurable.*

(ii) For $G \in \mathcal{G}$*,*

$$\int_G X \, d\mathbb{P} = \int_G Y \, d\mathbb{P}.$$

We will prove existence and uniqueness later.

Example 24.2. *Let* G *be an event such that* $\mathbb{P}(G) > 0$*, and* $\mathcal{G} = \sigma(\{G\}) = \{\emptyset, G, G^c, \Omega\}$*. Then,*

$$\mathbb{E}(X|\mathcal{G})(\omega) = \left\{ \begin{array}{ll} \mathbb{E}(X|G), & \text{if } \omega \in G \\ \mathbb{E}(X|G^c), & \text{if } \omega \in G^c. \end{array} \right\} = Y(\omega)$$

The random variable defined above is a conditional expectation:

Proof. Y is $\mathcal{G}$-measurable, because $\{\omega : Y(\omega) \in B\}$ is in $\mathcal{G}$ for every Borel set B in $\mathbb{R}$. Every pre-image we will attempt to take pre-image of will either contain the real number $\mathbb{E}(X|G)$ or it wont, and will either contain the number $\mathbb{E}(X|G^c)$ or it wont.

How about (ii)? We need to show that for arbitrary events, the integrals agree. We need to check that

$$\int_G X \, d\mathbb{P} = \int_G Y \, d\mathbb{P}$$

and

$$\int_{G^c} X \, d\mathbb{P} = \int_{G^c} Y \, d\mathbb{P}.$$

Let's check out the first of these. The RHS

$$\int_G Y \, d\mathbb{P} = \int_G Y(\omega) \, d\mathbb{P}(\omega) = \int_G \mathbb{E}(X|G) \, d\mathbb{P} = \mathbb{P}(G) \cdot \mathbb{E}(X|G),$$

since $\mathbb{E}(X|G)$ is a constant. By the definition of conditional expectation, this is

$$\int_G X \, d\mathbb{P}.$$

□

The interpretation of this is: Suppose that for an outcome ω, we know whether G occurs or not, for every G, but we do not know the outcome ω itself. Our only access to the nature of the experiment is by observing whether events G in the σ-algebra $\mathcal{G}$ occur or not[99]. Then $\mathbb{E}(X|\mathcal{G})(\omega)$ is our best guess of X given this information[100].

[99] In the example of a plane, we might know whether planes were delayed or not, but not the exact weather condition.

[100] Example: We have a history of planes. We want to guess a random variable X. Suppose there is a storm. Then, we want to take an average of the history of planes only in a storm.

24.2 Existence and Uniqueness of Conditional Expectation

We prove that the conditional expectation exists and is unique. This follows from a very general and very beautiful theorem in measure theory, the Radon-Nikodym Theorem.

Definition 24.3. *Let μ, ν be two measures on $(\Omega, \mathcal{F})$. We say that ν is absolutely continuous with respect to μ, and write $\nu \ll \mu$, if*

$$\mu(A) = 0 \text{ implies } \nu(A) = 0 \text{ for } A \in \mathcal{F}.$$

Example 24.4.

- *μ is the Lebesgue measure on $\mathbb{R}$, and ν is the standard normal distribution on $\mathbb{R}$:*

$$\nu(A) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2}\, dx.$$

- *Generally, $\nu \ll \mu$ if*

$$\nu(A) = \int_A f\, d\mu$$

where f is μ-integrable function (density).

A measure μ is *σ-finite* if there is a decomposition of Ω into a countable number of sets:

$$\Omega = \bigcup_{k=1}^{\infty} \Omega_k$$

such that $\mu(\Omega_k) < \infty$ for all k. One example is Lesbesgue measure.

The Radon-Nikodym theorem says that this general example is the only way these arise:

Theorem 24.5 (Radon-Nikodym)**.** *Let μ, ν be σ-finite measures on $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, then there exists an $\mathcal{F}$-measurable function f such that*

$$\nu(A) = \int_A f\, d\mu \text{ for } A \in \mathcal{F}. \tag{43}$$

The function f is usually called the *Radon-Nikodym derivative* of ν with respect to μ, and is denoted $f = \frac{d\mu}{d\nu}$. We can think of (43) as a version of a Fundamental Theorem of Calculus.

We will take $\mu = \mathbb{P}$, and tak $\nu(G) := \int_G X\, d\mathbb{P}$.

MARCH 3, 2008

We're doing this section on conditional expectation. Let's recall what it is. We have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, but now we'll be looking at different σ-algebras. Consider a σ-algebra $\mathcal{G} \subset \mathcal{F}$ that is coarser than $\mathcal{F}$. Then, the conditional expectation of a random variable X with respect to $\mathcal{G}$, denoted $\mathbb{E}(X|\mathcal{G})$, is a random variable Y such that

(i) $\mathcal{G}$-measurable,

(ii) $\int_A Y \, d\mathbb{P} = \int_A X \, d\mathbb{P}$ for all $A \in \mathcal{G}$.

The trivial case is when $\mathcal{G}$ is trivial. Then, this is the usual expectation.

Theorem 24.6. *The conditional expectation of an integrable random variable X exists and is unique (up to a null set[101]).*

Uniqueness will be easy. Existence will follow from the Radon-Nikodym Theorem.

Theorem 24.7 (Radon-Nikodym)**.** *If ν, μ are σ-finite measures on $(\Omega, \mathcal{F})$ and $\nu \ll \mu$, then there exists an integrable function $f = \frac{d\nu}{d\mu}$ such that*

$$\nu(A) = \int_A f \, d\mu$$

for $A \in \mathcal{F}$.

This is a generalization of the Fundamental Theorem of Calculus.

Proof of Theorem 24.6. To prove existence, the Radon-Nikodym Theorem[102] is applied. We will first do the case that $X \geq 0$. For every set A, we know $\int_A X \, d\mathbb{P}$. For every A, the RHS of the definition (part (ii)) gives us a number. This will be our μ. So, in the notation of the Radon-Nikodym Theorem for $(\Omega, \mathcal{G})$, $\mu = \mathbb{P}$. And, ν is defined by

$$\nu(A) := \int_A X \, d\mathbb{P} \text{ for } A \in \mathcal{G}.$$

Indeed, ν is a measure. The usual properties of integration will give you that this is a measure. Then, ν is a σ-finite measure (Exercise). Clearly $\nu \ll \mu$: If A is a null set with respect to μ, then integration over that set A gives you zero.

Thus, we are in the situation of the Radon-Nikodym Theorem. We conclude from it that there is a $\mathcal{G}$-measurable function f such that

$$\nu(A) = \int_A f \, d\mathbb{P} \text{ for all } A \in \mathcal{G}.$$

We'll take this function to be our random variable: Taking $Y := f$ satisfies both properties defining conditional expectation. This proves existence[103], for non-negative x.

For arbitrary X, decompose $X = X^+ - X^-$ with $X^+ \geq 0$ and $X^- \geq 0$. We find the conditional expectation of each part. Consider

$$Y_1 := \mathbb{E}(X^+|\mathcal{G})$$
$$Y_2 := \mathbb{E}(X^-|\mathcal{G}),$$

[101] up to a set of measure zero.

[102] actually, there is a version of this for signed measures.

[103] We looked at f is a density, but then we view it as a random variable in the end. This is quite interesting approach.

which exist by the above. Then define $Y = Y_1 - Y_2$. Then of course Y is $\mathcal{G}$-measureable (it is the difference of two $\mathcal{G}$-measurable functions). The second property just follows by linearity:

$$\int_A Y \, d\mathbb{P} = \int_A Y_1 \, d\mathbb{P} - \int_A Y_2 \, d\mathbb{P} = \int_A X^+ \, d\mathbb{P} - \int_A X^- \, d\mathbb{P} = \int_A X \, d\mathbb{P}.$$

Uniqueness is easy: Suppose that Y and Y' both satisfy the properties of the conditional expectation[104]. By (ii), one has

$$\int_A Y \, d\mathbb{P} = \int_A Y' \, d\mathbb{P} \text{ for all } A \in \mathcal{G}.$$

Then $\int_A (Y - Y') \, d\mathbb{P} = 0$ for all $A \in \mathcal{G}$. Consider the set A_ϵ where $Y - Y' > \epsilon$ for $\epsilon > 0$. Then

$$0 = \int_{A_\epsilon} (Y - Y') \, d\mathbb{P} \geq \mathbb{P}(A_\epsilon) \cdot \epsilon.$$

Then, $\mathbb{P}(A_\epsilon) = 0$ for all $\epsilon > 0$. Since $\{Y - Y' > 0\} = \bigcap_{n=1}^{\infty} A_{\frac{1}{n}}$. Then, the continuity property of probability measures implies

$$\mathbb{P}(Y - Y' > 0) = 0.$$

Similarly, by interchanging the roles of Y and Y', $\mathbb{P}(Y' - Y > 0) = 0$. These together imply that $Y = Y'$ almost surely[105]. □

Some examples:

Example 24.8. *1. The trivial σ-algebra $\mathcal{G} = \{\emptyset, \Omega\}$. Then, just being $\mathcal{G}$-measurable is a strong condition. Condition (i) implies Y is a constant (almost surely). We can even guess what constant it is: We would guess that it is usual expectation. In fact, $Y = \mathbb{E}X$.*

Indeed, we check condition (ii), namely, we check that

$$\int_A (\mathbb{E}X) \, d\mathbb{P} = \int_A X \, d\mathbb{P}$$

for $A = \Omega$ (the only one needing to be checked). This is true, since the RHS is expectation. Thus, both sides equal $\mathbb{E}X$, the usual definition.

2. Take the full σ-algebra $\mathcal{G} = \mathcal{F}$. Then

$$\mathbb{E}(X|\mathcal{F}) = X.$$

With this interpretation of conditional expectation that we gave last time, suppose we do not have access to X and we have to guess the random

[104] The idea then, satisfying (ii), means that we assign the same number.

[105] We should pick A_ϵ to be $\mathcal{G}$-measurable. A_ϵ is $\mathcal{G}$-measurable, because Y and Y' are $\mathcal{G}$-measurable.

variable. Suppose for every outcome of our experiment, we know whether the outcome belongs to the set $\mathcal{G}$ or not. Our guess of the random variable given this information is the conditional expectation. So, if we know nothing, then our best guess (by the first example) is the expectation. In the opposite case, if you know everything about the random variable, is to take this random variable: don't guess! There are intermediate cases of course.

3. *Discrete σ-algebra. Suppose $\Omega = \bigcup_k \Omega_k$ is a partition*[106] *of Ω such that $\mathbb{P}(\Omega_k) > 0$. Generate $\mathcal{G}$ by these sets: $\mathcal{G} = \sigma(\Omega_1, \Omega_2, \ldots)$. Then,*

$$\mathbb{E}(X|\mathcal{G})(\omega) = \mathbb{E}(X|\Omega_k)(\omega) \text{ for } \omega \in \Omega_k.$$

Let's take the verification of this as an exercise.

The Radon-Nikodym Theorem would not for us construct the random variable Y. We must first guess at Y. A random variable that is constant on the sets of the σ-algebra, and should take values as suggested by (ii) of the definition.

The conditional expectation is a random variable. In particular, it is a function. We take a guess at the values of our function, so that it is adaptive to our information.

Now, we can finally make sense of the condition expectation, even if some of the sets Ω_k (in example 3 above) is zero. In particular, we can now take the conditional expectation with respect to another random variable.

24.3 Conditional Expectation with respect to another random variable

We know what is $\mathbb{E}(X|Y \in [a,b])$. We know that this is

$$\frac{1}{\mathbb{P}(Y \in [a,b])} \int_{\{Y \in [a,b]\}} X \, d\mathbb{P}.$$

We do not know $\mathbb{E}(X|Y = y)$ when $\mathbb{P}(Y = y) = 0$. Yet, the question has a fit probabilistic interpretation. Thus, we want to understand this. Now, we can define what is this.

Definition 24.9. $\mathbb{E}(X|Y) := \mathbb{E}(X|\sigma(Y))$.

Recall that $\sigma(Y) = \{\{\omega : Y(\omega) \in B\}, B \text{ Borel}\}$. The idea is that the conditional expectation averages out the information. Here, it means that we want to average out the information about Y. We want to say that the expectation is the same number for all sets in $\sigma(Y)$ occuring. For example, Y has finitely many values $y_1, \ldots, y_n$. Then $\sigma(Y) = \sigma(\Omega_1, \ldots, \Omega_n)$, where $\Omega_k = \{Y = y_k\}$. So, $\mathbb{E}(X|Y)(\omega) = \mathbb{E}(X|\Omega_k)(\omega)$ for $\omega \in \Omega_k$. In other words, $\mathbb{E}(X|Y)(\omega) = \mathbb{E}(X|Y = y_k)$, for $\omega \in \Omega_k$.

MARCH 5, 2008

[106] We already discussed the partition $\Omega = \Omega_1 \cup \Omega_2$.

24.4 Properties of Conditional Expectation

Many of these properties are easy to guess, based on our knowledge of usual expectation.

Proposition 24.10 (Linearity)**.** $\mathbb{E}(aX + bY|\mathcal{F}) = a\mathbb{E}(X|\mathcal{F}) + b\mathbb{E}(Y|\mathcal{F})$.

The issue is that the proof is an existential statement. In all these proofs, you first guess what should be the conditional expectation. So, we propose that the correct formula for the LHS in the statement **is** the RHS of the statement:

Proof. We need to check that the RHS is the conditional expectation of $aX+bY$ given the σ-algebra $\mathcal{F}$. We verify the properties:

(i) RHS is $\mathcal{F}$-measurable: We have a linear combination of two $\mathcal{F}$-measurable functions.

(ii) For every $A \in \mathcal{F}$,

$$\begin{aligned}\int_A (a\mathbb{E}(X|\mathcal{F}) + b\mathbb{E}(Y|\mathcal{F}))\ d\mathbb{P} &= a\int_A \mathbb{E}(X|\mathcal{F})\ d\mathbb{P} + b\int_A \mathbb{E}(Y|\mathcal{F})\ d\mathbb{P} \text{ by linearity of integration} \\ &= a\int_A X\ d\mathbb{P} + b\int_A Y\ d\mathbb{P} \text{ by definition of conditional expectation} \\ &= \int_A (aX + bY)\ d\mathbb{P} \text{ by linearity.}\end{aligned}$$

□

Corollary 24.11. *Conditional expectation is a linear operator on* $L^1(\Omega)$.

We should compare this to the usual expectation, which is a linear **functional**. The difference is that conditional expectation takes a point in the space and returns a point in the space, and the expectation returns a number.

Proposition 24.12 (Monotonicity)**.** *If* $X \leq Y$ *a.s., then* $\mathbb{E}(X|\mathcal{F}) \leq \mathbb{E}(Y|\mathcal{F})$ *a.s.*

Again, we don't have full access to the conditional expectation. We can only test it using an event A.

Proof. Consider an event $A \in \mathcal{F}$. Then

$$\begin{aligned}\int_A \mathbb{E}(X|\mathcal{F}\ d\mathbb{P} &= \int_A X\ d\mathbb{P} \\ &\leq \int Y\ d\mathbb{P} \\ &= \int_A \mathbb{E}(Y|\mathcal{F})\ d\mathbb{P}\end{aligned}$$

Here, we only proved that one integral is smaller than the other.

Hence, $\int_A (\mathbb{E}(Y|\mathcal{F}) - \mathbb{E}(X|\mathcal{F}))\, d\mathbb{P} \geq 0$ for every $A \in \mathcal{F}$. Therefore, $\mathbb{E}(Y|\mathcal{F}) - \mathbb{E}(X|\mathcal{F}) \geq 0$ a.s., and let's leave that as an exercise (similar to a result in the previous lecture). Hint: consider $A_\epsilon := \{\mathbb{E}(Y|\mathcal{F}) - \mathbb{E}(X|\mathcal{F}) < -\epsilon\}$. □

Corollary 24.13. $\left|\mathbb{E}(X|\mathcal{F})\right| \leq \mathbb{E}(|X| \;\middle|\; \mathcal{F})$.

Proof. Note $X \leq |X|$ and $-X \leq |X|$. □

We have a conditional dominated convergence theorem:

Theorem 24.14 (Dominated Convergence Theorem)**.** *Suppose that random variables X_n converge to X a.s., and $|X_n| \leq Y$ a.s., with Y integrable. Then,*

$$\mathbb{E}(X_n|\mathcal{F}) \to \mathbb{E}(X|\mathcal{F}) \text{ a.s.}$$

It's interesting to note that we don't lose any power. We still have almost sure convergence. In the usual version for expectation, we have a statement about almost sure convergence of numbers, so we lose much. Here, we don't lose anything. We'll prove this.

Proof. We wish to show that a sequence of random variables convergence, so we must bound their difference.

$$\begin{aligned}
|\mathbb{E}(X_n|\mathcal{F}) - \mathbb{E}(X|\mathcal{F})| &= |\mathbb{E}(X_n - X|\mathcal{F})| \\
&\leq \mathbb{E}(|X_n - X| \;\middle|\; \mathcal{F}) \\
&\leq \mathbb{E}(Z_n|\mathcal{F}), \text{ where } Z_n = \sup_{k \geq n} |X_k - X| \text{ and clearly } Z_n \searrow .
\end{aligned}$$

In fact, $Z_n \searrow 0$ by hypothesis. We want to show that $\mathbb{E}(Z_n|\mathcal{F}) \searrow 0$ a.s. What we did here is that we replaced X by 0, essentially.

By monotonicity of the conditional expectations, $\mathbb{E}(Z_n|\mathcal{F})$ is nonincreasing a.s. Since it is non-negative, it converges to a limit:

$$\mathbb{E}(Z_n|\mathcal{F}) \downarrow Z \text{ a.s.}$$

What if Z is larger than zero? Since $Z \geq 0$, $\mathbb{E}Z = 0$ implies $Z = 0$. So, it suffices to show that $\mathbb{E}Z = 0$. For normal expectation, we already have a Dominated Convergence Theorem:

We have $0 \leq Z \leq Y + |X| \leq 2Y$.[107] Therefore

$$\begin{aligned}
\mathbb{E}Z &= \int_\Omega Z \, d\mathbb{P} \\
&\leq \int \mathbb{E}(Z_n|\mathcal{F}) \, d\mathbb{P} \\
&= \mathbb{E}Z_n, \text{ by definition of conditional expectation.}
\end{aligned}$$

Since $Z_n \downarrow 0$, the Monotone Convergence Theorem yields

$$\mathbb{E}Z_n \to 0.$$

Hence, $\mathbb{E}Z = 0$. □

[107] When did we use this fact? I leave it as a challenge.

Proposition 24.15. *If X is $\mathcal{F}$-measurable, then*

$$\mathbb{E}(XY|\mathcal{F}) = X \cdot \mathbb{E}(Y|\mathcal{F})$$

provided XY and X are integrable.

Remark 24.16. *We know this for $Y = 1$. $\mathbb{E}(X|\mathcal{F}) = X$. This is more general. This tells us that we can look at X as a constant when it's $\mathcal{F}$-measurable.*

We know this for $\mathcal{F} = \{\emptyset, \Omega\}$ the trivial σ-algebra. Then, the only choice for X is a constant: $X = a$. Then, $\mathbb{E}(aY) = a\mathbb{E}Y$.

How do we prove this? We do not know the conditional expectation itself, we only know its properties. We need to check that the RHS in the statement is the conditional expectation (that is, it satisfies the properties in the definition) of XY.

Proof. We check that $X \cdot \mathbb{E}(Y|\mathcal{F})$ is the conditional expectation of XY given $\mathcal{F}$.

(i) $X \cdot \mathbb{E}(Y|\mathcal{F})$ is $\mathcal{F}$-measurable, because X and $\mathbb{E}(Y|\mathcal{F})$ are $\mathcal{F}$-measurable.

(ii) For every $A \in \mathcal{F}$, we need to check that

$$\int_A X \ \mathbb{E}(Y|\mathcal{F}) \ d\mathbb{P} = \int_A XY \ d\mathbb{P}. \tag{44}$$

The proof of this is a bit technical. We follow the technique of Lebesgue integral, considering first indicators, then linear combinations, then positive/negative, etc. We sketch the proof:

(a) If $X = \mathbf{1}_B$ for $B \in \mathcal{F}$, then (44) becomes

$$\int_{A \cap B} \mathbb{E}(Y|\mathcal{F}) \ d\mathbb{P} = \int_{A \cap B} Y \ d\mathbb{P}$$

which is true because $A \cap B \in \mathcal{F}$.

(b) If X is a simple random variable, then (44) is true by linearity (of the integral).

(c) If $X \geq 0$, then there exists simple random variables $X_n \uparrow X$ a.s. Then we use (44) for X_n and apply the Monotone Convergence Theorem to complete the argument and achieve (44) for X.

(d) For arbitrary X, decompose $X = X^+ - X^-$ and use linearity to prove (44).

This verifies property (ii).

□

This will be useful when we deal with martingales. The interpretation: We know X perfectly. We are guessing the variable XY.

MARCH 7, 2008

This is the last lecture on conditional expectation. We will look at a geometric view. We will look at is as a projection

24.5 Conditional Expectation as a Projection

Here is one theorem that you will not necessarily view in a geometric way, but it will be useful for martingales.

Proposition 24.17 (Smoothing/Towering)**.** *If $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then*

1. $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$.
2. $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_2)|\mathcal{F}_1) = \mathbb{E}(X|\mathcal{F}_1)$.

Proof. Let $\mathcal{F}_1 \subseteq \mathcal{F}_2$.

1. $Y = \mathbb{E}(X|\mathcal{F}_1)$ is $\mathcal{F}_1$-measurable. $\mathcal{F}_1 \subseteq \mathcal{F}_2$ therefore also is $\mathcal{F}_2$-measurable. Hence
$$\mathbb{E}(Y|\mathcal{F}_2) = Y$$
by the above example.

2. We want to show that $\mathbb{E}(X|\mathcal{F}_1)$ is the conditional expectation of $\mathbb{E}(X|\mathcal{F}_2)$ given $\mathcal{F}_1$.

   (i) Indeed, $\mathbb{E}(X|\mathcal{F}_1)$ is $\mathcal{F}_1$-measurable.

   (ii) We check for every $A \in \mathcal{F}_1$, we want to show that
$$\int_A \mathbb{E}(X|\mathcal{F}_1)\ d\mathbb{P} = \int_A \mathbb{E}(X|\mathcal{F}_2). \tag{45}$$
   The LHS is the integral of X over A. So is the right: since $A \in \mathcal{F}_1 \subseteq \mathcal{F}_2$, the definition of conditional expectation shows that both sides of (45) are equal
$$\int_A X\ d\mathbb{P}.$$

Hence, (i) and (ii) are true.

□

Corollary 24.18. $\mathbb{E}(\mathbb{E}(X|\mathcal{F})|\mathcal{F}) = \mathbb{E}(X|\mathcal{F})$.

Here comes the geometric point of view, which is very useful. We know that conditional expectation is a linear operator. We also know that if we apply it twice, it's the same as applying it once. It satisfies $A^2 = A$. These are the projections. When you take a point in the image and want to project again, you're not doing anything: the image stays fixed. This corollary shows that the conditional expectation $P : X \mapsto \mathbb{E}(X|\mathcal{F})$ is a (linear) projection[108] in $L^1(\Omega)$ onto[109] the subspace of all $\mathcal{F}$-measurable random variables.

One proposition that we'll first put into more of a statistical point of view and then develop a geometric intuition for is

[108] that is, $P^2 = P$.

[109] Is it clear that it is indeed onto? One can take a random variable that is $\mathcal{F}$-measurable, P would do nothing.

Theorem 24.19 (Conditional expectation is the best estimator)**.** *For any $\mathcal{F}$-measurable random variable Y, we have*

$$\mathbb{E}(X - \mathbb{E}(X|\mathcal{F}))^2 \leq \mathbb{E}(X - Y)^2. \tag{46}$$

In other words, if you take any other $\mathcal{F}$-measurable random variable, then it will be worse (in the sense of distance[110]).

Both sides of (46) show is the mean squared error. So in terms of the mean squared error, $\mathbb{E}(X|\mathcal{F})$ gives us the best (that is, least) error.

In order to prove the theorem, we'll view it in a geometric way. So, here's a geometric interpretation: We want to view the mean square error as a true distance, as a metric space. We look at the Hilbert space $L^2(\Omega)$. The norm in the space of random variables is

$$\|X\|_2 = (\mathbb{E}X^2)^{1/2}.$$

Since it is a Hilbert space, the norm is actually given by an inner product, which is

$$\langle X, Y\rangle = \mathbb{E}XY.$$

This is good. Now, the left and right hand sides of (46) are 2-norms in this space. Consider

$$H_{\mathcal{F}} := \{X \in L^2(\Omega) \mid X \text{ is an } \mathcal{F}\text{-measurable random variable}\},$$

a closed subspace of $L^2(\Omega)$.

The picture behind the theorem is that $\mathbb{E}(X|\mathcal{F})$ is the orthogonal projection. The theorem says that

$$\|X - \mathbb{E}(X|\mathcal{F})\|_2 \leq \|X - Y\|_2,$$

so $\mathbb{E}(X|\mathcal{F})$ is the point in $H_{\mathcal{F}}$ nearest X.

In other words, because we know that the nearest point to a subspace is when the error $X - \mathbb{E}(X|\mathcal{F})$ is orthogonal, it will follow that

Corollary 24.20. *Conditional expectation $\mathbb{E}(\,\cdot\,|\,\mathcal{F})$ is the orthogonal projection in $L^2(\Omega)$ onto the subspace $H_{\mathcal{F}}$ of all $\mathcal{F}$-measurable random variables.*

We will prove Theorem 24.19 with this geometric background:

Proof. We will first prove the orthogonality: We prove that the error $X - \mathbb{E}(X|\mathcal{F})$ is orthogonal to $H_{\mathcal{F}}$.

Claim 24.21 (Orthogonality)**.** *$X - \mathbb{E}(X|\mathcal{F})$ is orthogonal to $H_{\mathcal{F}}$, i.e. for every $Z \in H_{\mathcal{F}}$, we have $\mathbb{E}Z(X - \mathbb{E}(X|\mathcal{F})) = 0$.*

Proof of claim. Indeed, $\mathbb{E}Z(X - \mathbb{E}(X|\mathcal{F})) = \mathbb{E}ZX - \mathbb{E}(Z\mathbb{E}(X|\mathcal{F}))$. Since Z is $\mathcal{F}$-measurable, we can put it on the inside on the RHS (since it acts like a constant). Thus, the expression is $= \mathbb{E}ZX - \mathbb{E}(\mathbb{E}(ZX|\mathcal{F}))$. By the smoothing property (part 2), gives $= \mathbb{E}ZX - \mathbb{E}ZX = 0$. □

[110] Think about variance for the distance. It's the expectation of the absolute value squared.

So we proved this geometric claim first, that the error is orthogonal. Now, with this geometry, we can conclude that the orthogonal amount is the shortest distance.

Now let $Y \in H_{\mathcal{F}}$. We wish to show that $\mathbb{E}(X-Y)^2 \geq \mathbb{E}(X-\mathbb{E}(X|\mathcal{F}))^2$. We will write

$$\begin{aligned}\mathbb{E}(X-Y)^2 &= \mathbb{E}(X-\mathbb{E}(X|\mathcal{F})+Z)^2, \quad Z = \mathbb{E}(X|\mathcal{F}) - Y \in H_{\mathcal{F}} \\ &= \mathbb{E}(X-\mathbb{E}(X|\mathcal{F}))^2 + \mathbb{E}Z^2 + 2\mathbb{E}Z(X-\mathbb{E}(X|\mathcal{F}))\end{aligned}$$

The second term is non-negative and the third term was shown (in "Orthogonality") to be zero. Thus

$$\mathbb{E}(X-Y)^2 \geq \mathbb{E}(X-\mathbb{E}(X|\mathcal{F}))^2.$$

This completes the proof. □

I guess we just proved Pythagoras' Theorem. With that geometric interpretation, we may want to go back to the smoothing/towering property and think of what it means.

In the Towering Proposition, note that $H_{\mathcal{F}_1} \subseteq H_{\mathcal{F}_2}$ if $\mathcal{F}_1 \subseteq \mathcal{F}_2$. The geometry of what we're doing reflects into the [relative] sizes of the spaces. The smoothing property just tells us these operators commute. That is, if $P_i = \mathbb{E}(\cdot \,|\, \mathcal{F}_i)$, then $P_1P_2 = P_2P_1 = P_1$.

The more information you have, the larger the target projection space, thus you have a better chance for less error. With the trivial σ-algebra, then the only random variables are constants, and they form a 1-dimensional subspace.

Do we have enough time for evaluations?

MARCH 10, 2008

25 Martingales

Today we are starting martingales, the last topic of this class. We'll do as much as we can. Similar to the case with conditional expectation, we'll be changing the σ-algebra. (The integrability that follows is just a technical condition.)

Definition 25.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An increasing sequence of σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ is called a filtration.*

A sequence (X_n) of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ is called a martingale relative to $(\mathcal{F}_n)$ if, for every n, $\mathbb{E}|X_n| < \infty$, and

(i) X_n is $\mathcal{F}_n$-measurable;

(ii) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$.

The condition (i) in some other terminology, is that the sequence (X_n) is *adapted* to $(\mathcal{F}_n)$.

Let's give one example with which you can temporarily think of martingales, and then we'll get to some theorems.

25.1 A small example

One example is gambling systems. Let X_n be the fortune of a gambler after n^{th} play. Here, $\mathcal{F}_n$ will be thought of some information about the game after n^{th} play. (So after, the n^{th} play, the gambler has access to just what happened).

Here, condition (i) is just saying that the gambler knows how much money he/she has at this time (that is, after the n^{th} play).

The second thing: we do not claim that X_{n+1} is $\mathcal{F}_n$-measurable. That would mean that the gambler knows the fortune after the next play, which has not yet occurred. Statement (ii) says that the gambler knows the expected fortune after the next play: The expected fortune after next play equals the present fortune. In other words, this is a fair game.

So, this is our temporary filling for all sorts of useful examples. We'll get more.

By the way, how do we interpret a filtration? The information we know keeps growing.

25.2 Martingale Theory

Definition 25.2. *A sequence of random variables (X_n) is called a martingale if it is a martingale relative to some filtration $(\mathcal{F}_n)$.*

Remark 25.3. *If this is the case, then*

$$\mathcal{G}_n = \sigma(X_1, \dots, X_n)$$

will always work.

That is, if $\mathcal{F}_n$ is not known, we can provide a filtration that will always work.

Proof. $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$ is obvious.

(i) Clearly, X_n is $\mathcal{G}_n$-measurable.

(ii) Note $\mathcal{G}_n \subseteq \mathcal{F}_n$. Why? Since X_n is $\mathcal{F}_n$-measurable and $\mathcal{G}_n$ is the smallest σ-algebra that makes $X_1, \dots, X_n$ measurable, thus $\mathcal{G}_n \subseteq \mathcal{F}_n$. $\mathbb{E}(X_{n+1}|\mathcal{G}_n) = \mathbb{E}(\mathbb{E}(X_{n+1}|\mathcal{F}_n)|\mathcal{G}_n)$, by towering. But this is $\mathbb{E}(X_n|\mathcal{G}_n)$ by definition of martingale. And finally, this is X_n, since X_n is $\mathcal{G}_n$-measurable.

Thus, $\mathcal{G}_n$ is the smallest filtration for X_n. □

In this case, (ii) can be rewritten[111] as,

$$\mathbb{E}(X_{n+1}|X_1, \dots, X_n) = X_n.$$

Here, the gambler (from the example in Section 25.1) only knows his/her fortune after first n plays, and nothing else about the game. This is the minimal information in gambling[112].

[111] Recall that $\mathbb{E}(X|\sigma(Y))$ is the definition of $\mathbb{E}(X|Y)$.

[112] We should at least suppose the gambler knows how much money he/she has!

Sometimes it is useful to replace the unknown filtration by a concrete one. However, this may not be quite possible. Sometimes, there might be other information about the game that can't be observed, for example, the fortunes of the other players of the game.

An immediate exercise here is to show

Proposition 25.4. *Suppose (X_n) is a martingale.*

1. $\mathbb{E}(X_{n+k}|\mathcal{F}_n) = X_n$ *for all* $k = 1, 2, \ldots$.
2. $\mathbb{E}X_1 = \mathbb{E}X_2 = \cdots$.

Proof. A sketch of ideas:

1. Just apply our trick repeatedly.
2. This average property follows from condition (ii).

□

25.2.1 Martingale Differences

One useful way to look at martingales is through martingale differences.

Definition 25.5. *Let (X_n) be a martingale relative to the filtration $(\mathcal{F}_n)$. Define differences*

$$\Delta_1 = X_1, \quad \Delta_2 = X_2 - X_1, \quad \Delta_3 = X_3 - X_2, \ldots$$

In terms of the gambler, this will be the gain/loss at n^{th} play (the change in fortune).

Then (ii) is equivalent to[113]

$$\mathbb{E}(\Delta_{n+1}|\mathcal{F}_n) = 0,$$

that is, the expected gains/losses at each play is zero.

Then, $X_n = \Delta_1 + \Delta_2 + \cdots + \Delta_n$. Hence (Δ_n) determine (X_n) uniquely, so

$$\sigma(X_1, \ldots, X_n) = \sigma(\Delta_1, \ldots, \Delta_n).$$

Thus, this would be an equivalent way (by specifying the Δ's) to define a martingale.

25.3 A second example

Here comes a second example: sums of independent random variables. Let (Δ_n) be independent integrable mean zero random variables. Then, they are martingale differences for some martingale: Specifically,

$$X_n = \Delta_1 + \cdots + \Delta_n$$

is a martingale[114].

Indeed, using $\mathcal{G}_n = \sigma(X_1, \ldots, X_n)$, we have that

[113] since X_n is already $\mathcal{F}_n$-measurable, so we can push it into the expectation

[114] and of course, the Δ_n are the martingale differences.

(i) X_n is obviously $\mathcal{G}_n$-measurable

(ii) $\mathbb{E}(X_{n+1}|X_1,\ldots,X_n) = \mathbb{E}((\sum_{k=1}^n \Delta_k) + \Delta_{n+1}|\Delta_1,\ldots,\Delta_n)$

Part (ii)'s "given" follows since the σ-algebras are the some. Thus, on the RHS, just the linearity gives $\mathbb{E}(\sum_{k=1}^n \Delta_n|\Delta_1,\ldots,\Delta_n)+\mathbb{E}(\Delta_{n+1}|\Delta_1,\ldots,\Delta_n)$. So, this would be $\sum_{k=1}^n \Delta_k + 0$, by the independence.

So, this game has no memory (of what happened in the past). This is the simplest type of game. In other situations, Δ_n might be information of the past. So, this shows how martingales are generalization of independent random variables. If you know Markov chains a bit, then in a Markov chain, it depends only on the most recent iteration. In a Martingale, even this might not be true. So here, the information only depends on the result, the fortune, of the player. In a Markov chain, there is a dependence on the outcome of the previous play, not just the fortune.

So this is a very illuminating example.

25.4 A third example

This is also very general. A martingale can also be defined in this way. Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{F}_n$ be a filtration in $\mathcal{F}$. We will create a martingale out of a single random variable:

For each n, define

$$X_n = \mathbb{E}(X|\mathcal{F}_n),$$

the nth random variable in the sequence (X_n). Then, (X_n) is a martingale relative to $\mathcal{F}_n$.

Proof. (i) X_n is $\mathcal{F}_n$-measurable.

(ii) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_{n+1})|\mathcal{F}_n) = \mathbb{E}(X|\mathcal{F}_n) = X_n$.

□

So, we have a martingale, obtained just from a single random variable and a filtration, by averaging. We interpret X as being all knowledge (ultimate knowledge) about the gambler's fortune at all points in time (the gambler's fortune at time ∞). $X(\omega)$ is the knowledge for outcome ω. It's the best guess about the random variable X (without the specified outcome ω).

MARCH 12, 2008

Monday (March 17th) class is cancelled. So, we are going through the theory of martingales.

Definition 25.6. *(X_n) is called a supermartingale if condition (ii) in the definition of martingale is replaced by $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$.*

Similarly, (X_n) is called a submartingale if condition (ii) in the definition of martingale is replaced by $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$.

I realize it seems strange that super comes with $\leq$ and sub comes with $\geq$. Durrett talks about this.

The interpretation of this, in terms of gambling:

1. A supermartingale is a gambler's profit in an unfavorable game.
2. A submartingale is a gambler's profit in a favorable game.

A canonical example for supermartingales: partial sums of independent random variables with negative means. Independent random variables with positive means form submartingales.

They are related to each other: There is a dual theory of submartingales and supermartingales.

Remark 25.7. *(X_n) is a supermartingale iff $(-X_n)$ is a submartingale.*

A more general source of submartingales is through a convex function:

Proposition 25.8. *Let (X_n) be a martingale and φ be a convex function s.t. $\mathbb{E}|\varphi(X_n)| < \infty$ for every n. Then $(\varphi(X_n))$ is a submartingale.*

Proof.

$$\begin{aligned}\mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) &\geq \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \\ &= \varphi(X_n)\end{aligned}$$

by a conditional version of Jensen's Inequality. □

Of course, if you take a concave function, you get a supermartingale. It's a pretty general way to construct submartingales out of martingales.

Now, we're going into gambling strategies. Actually, we're going into theory of martingales.

25.5 Strategies

So far, only the game dictates the rule. We want now to have the gambler to be able to bet. The definition related to this betting is

Definition 25.9. *A sequence of random variables (H_n) is called predictable (with respect to a filtration $(\mathcal{F}_n)$) if H_n is $\mathcal{F}_{n-1}$-measurable.*

We want to predict H_n with certainty after the previous game. So, after the $n-1$ game, we are able to determine H_n. That is, H_n can be predicted with certainty from the information available at time $n-1$.

Example 25.10. *We will think of $H_n \geq 0$ as the amount of money the gambler bets at time n.*

The gambler bets with certainty on the nth game, but the outcome of the nth game is not yet known. In the case of the example, (H_n) is the *gambler strategy*. The strategy can even be defined ahead of time: The gambler can decide what to do in all cases ahead of time. Of course, the major question is what is the optimal strategy[115].

The **simplest strategy** is to bet 1 dollar at a time. Don't think: just bet 1 every game. Namely,

$$\left.\begin{array}{l}\text{Let } X_n \text{ be the net amount of money the gambler wins at}\\ \text{time } n\text{, using the simplest strategy.}\end{array}\right\} \tag{47}$$

Thus, suppose (47) is what is known. For an arbitrary strategy (H_n), the net amount of money the gambler wins at time n is

$$(H \cdot X)_n := \sum_{k=1}^{n} H_k(X_k - X_{k-1})$$

This is the gambler's profit with arbitrary strategy.

The result, the pessimistic result, is that no strategy can beat an unfavorable game[116], that is, a supermartingale.

Theorem 25.11. *Let (X_n) be a supermartingale. Let (H_n) be a predictable sequence, and suppose each $H_n \geq 0$ is bounded. Then $(H \cdot X)_n$ is a supermartingale.*

If we keep track of how we're doing in this game, each time we're getting worse and worse and worse.

Proof.

$$\begin{aligned}\mathbb{E}((H \cdot X)_{n+1}|\mathcal{F}_n) &= \mathbb{E}((H \cdot X)_n + H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n)\\ &= (H \cdot X)_n + H_{n+1}\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n), \text{ since } (H \cdot X)_n \text{ and } H_{n+1} \text{ are } \mathcal{F}_n\text{-measurable}\\ &\leq (H \cdot X)_n, \text{ since } H_{n+1} \geq 0 \text{ and } X_n \text{ a supermartingale}\end{aligned}$$

□

So, this is pessimistic, but what about if we pick a stopping time?

25.6 Stopping times

Definition 25.12. *A random variable $N \in \mathbb{N}$ is a stopping time if the event $\{N = n\} \in \mathcal{F}_n$ for every n.*

We are able to decide whether we are going to stop or not right after the nth game. Our decision to stop right after the nth game depends only on the information about n games. (We can not mentally play more games, and decide to have stopped earlier.)

[115] I must warn you, the results are pessimistic.
[116] So there is no strategy, actually.

If we also throw in the possibility of stopping at any time, then the theorem is the same pessimistic result. Even then we can't win. No stopping time can beat an unfavorable game. First, define: Let $N \wedge n$ denote the minimum of the two.

Corollary 25.13. *Let N be a stopping time, and (X_n) be a supermartingale. Then $(X_{N\wedge n})$ is a supermartingale.*

We interpret this as our sequence $Y_n = X_{N\wedge n}$ freezes at some point, namely, it looks like

$$X_1, X_2, X_3, \ldots, X_N, X_N, X_N, \ldots.$$

Proof. Let $H_n := \mathbf{1}_{\{N\geq n\}}$. That is, will play the nth game. Obviously, H_n is $\mathcal{F}_{n-1}$-measurable because $\{N \geq n\} = \{N \leq n-1\}^c \in \mathcal{F}_{n-1}$.

Then, by the theorem above, $(H \cdot X)_n$ is a supermartingale. (If we stop at N, we have a telescoping series, and all that is left is the biggest term. We're left with $X_{N\wedge n} - X_0$.) The constant X_0 also forms a supermartingale (a trivial supermartingale). Thus

$$X_{N\wedge n} = (H \cdot X)_n + X_0,$$

the sum of two supermartingales is a supermartingale. So $X_{N\wedge n}$ is a supermartingale. □

No optimal startegy, no optimal stopping time. Now, a paradox.

25.7 The Martingale Paradox

We're going to describe a simple winning strategy. It's very old. This describes the etymology of the word. This describes a fake belt. Another theory is something that can hold a horse from running too fast. A martingale has this sort of chain property.

Here's the strategy: Double the bet if we loose, at every game. So, enter with one dollar. Suppose, for simplicity, I lose or win with probability $\frac{1}{2}$. If I win, they'll give me the amount of my bet. If I lose, I lose my bet. I enter with one dollar. If I lose, I bet two. If I lose the two, I bet four dollars. Next time I lose, I bet eight dollars. Suppose I'm rich enough. Every time I lose, I'll bet double. And once I win, I'll quit when I win.

What is our profit? If we loose k times and win the next time, we compute:

$$-1 - 2 - 2^2 - \cdots - 2^k + 2^{k+1} = 1 > 0.$$

Of course, I if quadruple each time, I can win more. This seems to be in contradiction with the theory. I encourage you to think about this paradox. It's actually ruined so many people in the casinos. See movie Los Alamos (sp?).

MARCH 14, 2008

Today's our last class. There is no class on Monday. The final exam is posted. Please do not be late on turning in the final exam. I will collect it precisely at noon. Please turn it into my mailbox. If some typo is discovered, I will put a note in red. If you have any difficulty, you may wish to check for notes in red.

25.8 Two results of Doob

Today, we cover two classical results due to Doob. One of them is the Upcrossing Inequality. We were talking about gambling, for a bit. Now, we'll look at stock prices.

25.8.1 The Upcrossing Inequality

Let (X_n) be stock prices at time n. What is the simplest strategy in the market? Buy low and sell high. The broker's strategy, which we are examining here, is precisely this.

We will assume that the market is favorable. That is, assume (X_n) is a submartingale. We expect the broker to earn money in the long run. We analyze this strategy. We buy at one [lower] level, and sell at a higher level. The broker buys at level a and sells at level b.

The period of time when the broker holds onto a share is called an upcrossing.

Definition 25.14. *A sequence $(X_k, X_{k+1}, \dots, X_m)$ is an upcrossing if*

$$X_k \le a \le X_{k+1} \le X_{k+2} \le \cdots \le X_{m-1} \le b \le X_m.$$

In the definition, this does not quite perfectly reflect our notion, because the upcrossing only concerns the price growth of the share within $[a, b]$.

Here is the Broker's Strategy: When the price plunges below a first time, buy a share. Hold onto it until the price jumps above b first time; repeat. During every upcrossing, the broker earns at least $b - a$ dollars per share. Now, the biggest question is: how many crossings U_n are there during time n (that is, from 1 to n). Then, our profit guarantee is at least $(b - a)n$.

Another paradoxical result is

Theorem 25.15 (Upcrossing Inequality)**.** *Let (X_n) be a submartingale, and U_n be the number of upcrossings from a to b in time n. Then,*

$$\mathbb{E}U_n \le \frac{\mathbb{E}(X_n - X_0)}{b - a}.$$

This is another pessimistic result. What does it say? If we multiply by $b-a$, we obtain

$$\mathbb{E}[(b - a)U] \le \mathbb{E}[X_n - X_0].$$

The LHS is the profit while using the strategy "buy low, sell high." What is the RHS? X_0 is the price to buy at the beginning. X_n is the price to sell at time

n. So, the RHS is the profit using the ignorant strategy "sit and wait," where we sell after time n regardless of X_n. I'll leave for you as a challenge to think of how this is possible.

Proof. We realize the broker's strategy "buy low, sell high" as a predictable sequence. After time k, what the broker is doing is

$$H_k := \begin{cases} 1 & \text{if broker holds a share} \\ 0 & \text{otherwise} \end{cases}$$

Then, what is the profit?

$$\text{Profit } = (H \cdot X)_n = \sum_{k=1}^{n} H_k (X_k - X_{k-1}).$$

When the H is 1, then the RHS above is a telescoping sum, which just gives $X_{\text{last}} - X_{\text{first}}$.

We know that the profit is bounded below:

$$(H \cdot X)_n \geq (b-a) U_n. \tag{48}$$

Now, we use the theorem that this is a supermartingale. We will look at the empty spaces, where we do not hold a share. Define

$$\overline{H_k} = 1 - H_k,$$

which indicates the times we do not hold a share.

Easily, we have

$$(H \cdot X)_n + (\overline{H} \cdot X)_n = (1 \cdot X)_n \tag{49}$$

The first term in the LHS is the profit in "buy low, sell high." The RHS is the profit in "sit and wait". We need to analyze the difference, given by $(\overline{H} \cdot X)_n$.

(X_n) is a submartingale, and $(\overline{H_n})$ is predictable. By the dual formulation of Theorem 25.11, $(\overline{H} \cdot X)_n$ is a submartingale.

Thus, $\mathbb{E}(\overline{H} \cdot X)_n \geq \mathbb{E}(\overline{H} \cdot X)_0 = 0$. We apply expectation to (49) and use the above to obtain

$$\mathbb{E}(H \cdot X)_n \leq \mathbb{E}(X_n - X_0).$$

Together with (48), this completes the proof. □

The martingale is either bounded in the end, or it's not. If it's unbounded, then it will not respect the midpoint of $[a, b]$. Then, the number of crossings will be small.

So, all these gambling and stock prices have remarkable theorems, which complete the course.

25.8.2 The Martingale Convergence Theorem

It's one of the rare consequences of taking a really applied problem and apply it to a pure problem.

Let's give a simple formulation of it.

Theorem 25.16 (Martingale Convergence Theorem)**.** *Let (X_n) be a submartingale and $\sup_n \mathbb{E}|X_n| < \infty$. Then $X_n \to X$ a.s. for some random variable X with $\mathbb{E}|X| < \infty$.*

The proof is a simple application of the Upcrossing Inequality. What can go wrong, if X_n does not converge? Then X_n oscillates. The number of upcrossings will be infinite. But this can not be, because the RHS of the previous theorem (by hypothesis) would be finite, so this is impossible. Here is the formal proof.

Proof. Let K be the level that bounds all the expectations. That is $\mathbb{E}|X_n| \leq K$ for all n. So the Upcrossing Inequality (with the triangle inequality) implies

$$\mathbb{E}U_n \leq \frac{\mathbb{E}|X_n| + \mathbb{E}|X_0|}{b-a}$$

for all a, b.

We know that $\mathbb{E}|X_n| \leq K$. If (X_n) is a submartingale, then $(|X_n|)$ is as submartingale (since $x \mapsto |x|$ is a convex function). Thus $\mathbb{E}|X_0| \leq \mathbb{E}|X_n| \leq K$. Therefore

$$\mathbb{E}U_n \leq \frac{2K}{b-a}.$$

We need something that works for infintely many U_ns. What can we say about the total number of upcrossings $\sup_n U_n$? U_n is integrable and monotone increasing. So $\sup_n U_n$ is integrable by the Monotone Convergence Theorem.

In particular, an integrable function is finite almost everywhere. That is,

$$\sup U_n < \infty \text{ a.s.}$$

The number of upcrossings (in such a market) will be finite. So, how can the conclusion be false? That is, how can X_n not converge to anything? We consider

$$X_* = \liminf X_n \qquad X^* = \limsup X_n.$$

The crucial observation is that if $X_* < a < b < X^*$, then this sequence must go between the two infinitely many times. If this is true, then the number of upcrossings must be infinite, and this does not happen. Hence,

$$\mathbb{P}(X_* < a < b < X^*) = 0$$

for every $a < b$. Now, we can tighten a and b together. We can represent the event

$$\{X_* < X^*\} = \bigcup \{X_* < a < b < X^*\},$$

where the union is over all $a, b \in \mathbb{Q}$, $a < b$ (for countability). Every event in the union on the RHS has measure zero. It follows that

$$\mathbb{P}(X_* < X^*) = 0.$$

This, it must be the case that $X_n \to X$ for some random variable.

Let's leave $\mathbb{E}|X| < \infty$ as an exercise. □

This is an application to pure subject matter from an applied topic.

This is remarkable.

Corollary 25.17. *For every non-negative martingale (X_n), $X_n \to X$ a.s. to some random variable X with $\mathbb{E}|X| < \infty$.*

We haven't even said anything about integrability of X_ns.

Proof. If (X_n) is non-negative, then $\mathbb{E}|X_n| = \mathbb{E}X_n$. But since this is a martingale, this is $\mathbb{E}X_0$, which is finite. Thus, the condition of the previous theorem is satisfied. □

It's very surprising is that we've done almost nothing, and we've created this random variable to which it converges almost surely. Level 0 is a strong boundedness property.

The theorem "almost" shows that (X_n) can be obtained from a single random variable X by taking expectations. Recall that $X_n = \mathbb{E}(X|\mathcal{F}_n)$ is always a martingale. In the theorem, we already have a candidate for X.

This is not quite true. Why? The simplest example is a random walk. The floor is level zero. If I hit the floor, I stay there. The expectation of every increment is finite. So, at some point (almost surely), I will hit the floor. And of course, from zero, you can not recover the martingale.

This completes our year, and I wish to you the best.