Contents

Chapter 1 Probability Measures

Lecture 1 09/05

Definition. σ-field: A class $\mathcal{F} \subseteq \mathfrak{P}(\Omega)$ of subsets of a universe Ω is called a σ-field if:

(1) $\Omega \in \mathcal{F}$.

(2) If $A \in \mathcal{F}$ then $A^c = \Omega \setminus A \in \mathcal{F}$.

(3) If we have $A_1, A_2, \ldots \in \mathcal{F}$ (countably infinitely many sets), then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Example: The set $\mathcal{F} = \{\emptyset, \Omega\}$ is called the trivial σ-field.

Example: The power set $^{\Omega} = \{A : A \subseteq \Omega\}$ is a σ-field.

Example: The set $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ is a σ-field for any $A \subseteq \Omega$.

Example: The set $\mathcal{F} = \{A \subseteq \Omega : A \text{ is countable or } A^c \text{ is countable}\}$ is a σ-field. If any set has A^c is countable, then the union has a smaller (therefore countable) complement. If all sets are countable, then their countable union is countable.

Example: Consider the following probability space with $\Omega = (0, 1]$ and $\mathscr{B}_0 = \{$the finite union of disjoint subintervals of Ω of the form $(a, b]$ of $\Omega\}$. We see then $\Omega \in \mathscr{B}_0$ and if A is of the form $(a_1, b_1] \cup \cdots \cup (a_n, b_n]$, then $A^c = (0, a_1] \cup (b_1, a_2] \cup \cdots \cup (b_n, 1] \in \mathscr{B}_0$. Moreover, we see that for any two sets $A_1, A_2 \in \mathscr{B}_0$, we can combine them by adding non-overlapping intervals and adding the union of overlapping intervals, so $A_1 \cup A_2 \in \mathscr{B}_0$. By induction, we can show that for $A_1, \ldots, A_n \in \mathscr{B}_0$ then $A_1 \cup \cdots \cup A_n \in \mathscr{B}_0$, however, this is only for finite (not countable) unions. Indeed, a countable union of disjoint intervals is not in $\mathscr{B}_0$. E.g., where $A_n = (\frac{1}{2n}, \frac{1}{2n+1}]$, the union $\bigcup_{i=1}^{\infty} A_i$ has no finite union (it has infinitely many connected components). Another counterexample is $A_n = (0, 1 - \frac{1}{n}]$ since $\bigcup_{i=1}^{\infty} A_i = (0, 1) \notin \mathscr{B}_0$ since it is right open.

Definition. σ-field Generators: Let $\mathscr{A}$ be a class of sets. Then the intersection of all the σ-fields containing $\mathscr{A}$ is called the σ-field generated by $\mathscr{A}$, defined as $\sigma(\mathscr{A})$. One can check that this is indeed a σ-field. Moreover, $\sigma(\mathscr{A})$ is the smallest σ-field containing $\mathscr{A}$.

Example: The trivial σ-field can be generated as $\{\emptyset, \Omega\} = \sigma(\{\emptyset\})$.

Example: The following σ-field is generated $\{\emptyset, A, A^c, \Omega\} = \sigma(\{A\}) = \sigma(\{A^c\})$.

Example: The following σ-field is generated $\{A \subseteq \Omega : A \text{ is countable or } A^c \text{ is countable}\} = \sigma(\{\{w\} : w \in \Omega\})$ is generated by the set of all singleton sets.

Example. Borel σ-field: The Borel σ-field $\mathscr{B}$ defined on $\Omega = (0, 1]$ is the σ-field generated by $\mathscr{B}_0 = \{$the finite union of disjoint subintervals of Ω of the form $(a, b]\}$. It can also be generated by $\mathscr{B} = \sigma(\{(a, b] \in \Omega\}) = \sigma(\{[a, b) \in \Omega\}) = \sigma(\{[a, b] \in \Omega\}) = \sigma(\{(a, b) \in \Omega\})$. We can define $\mathscr{B}$ on $\Omega = \mathbb{R}$ in the same way. For a general topological space, the Borel σ-field is generated by the set of all open (or equivalently all closed) sets.

Exercise: Check that $\mathscr{B} = \sigma(\{(a,b] \in \Omega\}) = \sigma(\{[a,b) \in \Omega\}) = \sigma(\{[a,b] \in \Omega\}) = \sigma(\{(a,b) \in \Omega\})$

Definition. Probability Measure: A set function defined on a σ-field $\mathbb{P} : \mathcal{F} \to [0,1]$ is a probability measure if

(1) $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}$.

(2) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(3) $\mathbb{P}$ must be countably-additive. In particular, if $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then $\lim_{n\to\infty} \mathbb{P}(\bigcup_{i=1}^{n} A_i) = \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Remark: Some of the conditions of the above definition for probability measures are redundant. In particular, the following is a more succinct definition that is equivalent to the above.

(1') $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$.

(2') $\mathbb{P}(\Omega) = 1$.

(3') $\mathbb{P}$ is countably-additive.

Remark: Probability can be interpreted from a frequentist perspective as the proportion of the times a given event occurs (or "long-time frequency") when you repeat an experiment over and over. Using such a perspective, conditions (1) and (2) are immediate, however, (3) does not hold since this only yields finite additivity, not countable additivity. Condition (3) is a technical condition which allows us to take limits.

Remark: We often have two large interpretations (from math to the real world) of probability. Objective probability is what is defined above, this yields the frequentist perspective. Subjective probability, on the other hand, is a scale of what events are more likely than which other events, then rescaled to $[0,1]$, this yields the Bayesian perspective. Note that going from the real world back to mathematics is modelling.

Definition. Probability space: Defined as the triplet $(\Omega, F, \mathbb{P})$.

- Ω: A set/collection of points $\{\omega\}$, called the *sample space*. We liken it to the set which contains all possible outcomes of some random experiment/variable. However, mathematically it is just some set.
- $\mathcal{F}$: A σ-field (or sometimes called a σ-algebra). It is a class of subsets of Ω and satisfies the properties of the above definition. The elements of $\mathcal{F}$ are called events.
- $\mathbb{P}$: Is the probability measure assigning probabilities to events of $\mathcal{F}$.

Example: Consider the experiment measuring the outcomes of flipping a coin n times, then our sample space $\Omega = \{\text{sequences of } H \text{ and } T \text{ of length } n\}$.

Example: Consider the experiment measuring the time until some random event happens (e.g., time until an atom decays), then our sample space $\Omega = \{0\} \cup \mathbb{R}^+$.

Example: Consider the experiment of what will tomorrow's weather be, then our sample space $\Omega = \{\text{sunny, rainy, cloudy}, \ldots\}$.

Lecture 2 09/10

Example: Let Ω be countable (or finite). Let $p : \Omega \to \mathbb{R}$ be a function such that $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$. Then define $\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$ for all $A \in \mathcal{F}$, then $\mathbb{P}$ is a probability measure. The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is called the *discrete probability space*. Note that here $\mathcal{F}$ is not needed, and just specifies for which sets $\mathbb{P}$ is defined. Note moreover that p is not a probability measure.

Remark: Why do we need $\mathcal{F}$?

1. We can define a probability measure for any $\mathfrak{P}(S)$, however, we may not be able to define a good probability measure. For instance, for a given $\omega \in \Omega$, we could define $\mathbb{P}(A) = 1$ if $\omega \in A \subseteq \Omega$ and $\mathbb{P}(A) = 0$ otherwise. In particular, we may not have translation invariance $\mu(A) = \mu(A + b)$, such as for the defined probability measure above. In fact, there does not necessarily exist a translation invariant probability measure whenever $\mathcal{F} = \mathfrak{P}(\Omega)$. Translation invariance is a desirable property, as we hope a measure roughly defines the size of a set.

2. $\mathcal{F}$ represents the "information" available for an experiment. For instance, there may be events which are unsure whether or not they occur and which we may not know their probability. In particular, $\mathcal{F}$ is the collection of sets for which we can talk about probability and will know if/when they occur. For different people and at different points in time we may know more information and be able to speak to more events and therefore have different σ-fields $\mathcal{F}$.

Definition. Set Limits: Recall for a sequence of numbers a_n, we have

$$\liminf_{n\to\infty} a_n = \lim_{n\to\infty} \inf_{k \geq n} a_k \qquad \text{and} \qquad \limsup_{n\to\infty} a_n = \lim_{n\to\infty} \sup_{k \geq n} a_k$$

We can extend this to sets as follows. Let $A_1, A_2, \ldots, \subseteq \Omega$ be an enumeration of sets. We define the following

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{j \geq n} A_j = \{x \in \Omega : \exists n, \forall j \geq n, x \in A_j\}$$

$\liminf$, therefore, represents the part of the set which stops changing (the smallest subset that is contained in every tail subsequence). Then $\omega \in \liminf_{n\to\infty} A_n$ if and only if $\omega \in A_k$ for all but finitely many k. You might therefore write

$$\liminf_{n\to\infty} A_n = \{A_n \underbrace{\text{a.a.}}_{\text{almost always}}\}$$

I.e., there are only finitely many k where $\omega \notin A_k$. We also define

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{j\geq n} A_j = \{x \in \Omega : \forall n, \exists j \geq n, x \in A_j\}$$

lim sup, therefore, represents the part of the set which will always eventually be attainable (the largest subset that is contained in every tail subsequence). Then $\omega \in \limsup_{n\to\infty} A_n$ if and only if ω happens infinitely often. You might therefore write

$$\limsup_{n\to\infty} A_n = \{A_n \underbrace{\text{ i.o. }}_{\text{infinitely often}}\}$$

I.e., there are infinitely many k where $\omega \in A_k$. When we $\liminf_{n\to\infty} A_n = \limsup_{n\to\infty} A_n$, then we call it $\lim_{n\to\infty} A_n$.

Proposition 1. Limits of Chains: When $A_1 \subseteq A_2 \subseteq \cdots$, then

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n = \lim_{n\to\infty} \bigcup_{k=1}^{n} A_k$$

When $A_1 \supseteq A_2 \supseteq \cdots$, then

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n = \lim_{n\to\infty} \bigcap_{k=1}^{n} A_k$$

Proposition 2. Properties of Probability Measures:

(1) Probability measures are monotonic, in particular if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

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

(3) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Exercise: Prove the above

Proposition 3. Inclusion-Exclusion Formula: The inclusion-exclusion formula specifies the probability of the union of sets: $A_1, A_2, \ldots, A_n \in \mathcal{F}$

$$\begin{aligned}\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{1\leq i\leq n} \mathbb{P}(A_i) &- \sum_{1\leq i<j\leq n} \mathbb{P}(A_i \cap A_j) \\ &+ \sum_{1\leq i<j<k\leq n} \mathbb{P}(A_i \cap A_j \cap A_k) \\ &- \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n)\end{aligned}$$

Proof. For $n = 2$ the result holds by property (3). Assume that the result holds for n. Then note

$$\begin{aligned}
\mathbb{P}(\bigcup_{i=1}^{n+1} A_i) &= \mathbb{P}((\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) \\
&= \mathbb{P}(\cup_{i=1}^{n} A_i) + \mathbb{P}(A_{n+1}) - \mathbb{P}((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
&= \mathbb{P}(\cup_{i=1}^{n} A_i) + \mathbb{P}(A_{n+1}) - \mathbb{P}(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})) \\
&= \sum_{1\le i\le n} \mathbb{P}(A_i) - \sum_{1\le i<j\le n} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n) \\
&\qquad + \mathbb{P}(A_{n+1}) - \sum_{1\le i\le n} \mathbb{P}(A_i \cap A_{n+1}) + \sum_{1\le i<j\le n} \mathbb{P}(A_i \cap A_j \cap A_{n+1}) \\
&\qquad - \cdots - (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_{n+1}) \\
&= \sum_{1\le i\le n} \mathbb{P}(A_i) - \sum_{1\le i<j\le n} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+2}\mathbb{P}(A_1 \cap \cdots \cap A_{n+1})
\end{aligned}$$

Thus the result holds by induction. □

Proposition 4. Continuity of Probability Measures: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability sapce. Let $A_1, A_2, \ldots \in \mathcal{F}$. If $A_1 \subseteq A_2 \subseteq \cdots$ is an increasing sequence or $A_1 \supseteq A_2 \supseteq \cdots$ is a decreasing sequence, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\lim_{n\to\infty} A_n)$. That is, probability measures are continuous from below and from above.

Proof. Suppose A_n is increasing, i.e., $A_1 \subseteq A_2 \subseteq \cdots$. Define $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i = A_n \setminus A_{n-1}$ to be the portion added by A_n. Note then that $B_1, B_2, \ldots$ is a disjoint sequence and $A_n = \bigcup_{i=1}^{n} B_i$, and in particular $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} B_i$. Then by $\mathbb{P}$'s countable-additivity

$$\mathbb{P}(A) = \mathbb{P}(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \lim_{n\to\infty} \mathbb{P}(\bigcup_{i=1}^{n} B_i) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Suppose instead that A_n is decreasing, i.e., $A_1 \supseteq A_2 \supseteq \cdots$. Then instead, considering complements we see that the sequence A_n^c is increasing and with $\lim_{n\to\infty} A_n^c = A^c$. Then from the previous result, we see that

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = \lim_{n\to\infty} (1 - \mathbb{P}(A_n^c)) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

This proves the result from above and from below. □

Proposition 5. Boole's Inequality: If $A_1, A_2, \ldots \in \mathcal{F}$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) \le \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Proof. Define $B_1 = A_1$ and $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i \subseteq A_n$ to be what is newly added to $\bigcup_{i=1}^{n} A_i$ by A_n. Then we see that $B_1, B_2, \ldots$ are disjoint. Moreover, we see that $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$,

and in particular $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. Therefore

$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \mathbb{P}(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

since each $B_i \subseteq A_i$, as desired. □

Definition. Field: A class $\mathcal{F}_0 \subseteq \mathfrak{P}(\Omega)$ of subsets of Ω is called a field if it satisfies the following:

(1) $\Omega \in \mathcal{F}_0$.

(2) If $A \in \mathcal{F}_0$ then $A^c \in \mathcal{F}_0$.

(3) If $A, B \in \mathcal{F}_0$ then $A \cup B \in \mathcal{F}_0$.

Note that property (3) shows closedness under finite unions by induction, however, unlike a σ-field, we do not have closedness under *countable* unions. Note that we see then that if $\mathcal{F}$ is a σ-field then it is also a field.

Lecture 3 09/12

Definition. Outer Measure: Let Ω be a sample space with field $\mathcal{F}_0$ and measure $\mathbb{P}$. For a set $A \subseteq \Omega$ (not necessarily in $\mathcal{F}_0$), we define

$$\mathbb{P}^*(A) := \inf \left\{ \sum_{n}^{\infty} \mathbb{P}(A_n) : A \subseteq \bigcup_{n=1}^{\infty} A_n, \quad A_1, A_2, \ldots \in \mathcal{F}_0, \right\}$$

That is, the outer measure of a set A is the smallest sum of the measures of a cover of A.

Definition. $\mathbb{P}^*$-measurable Set: Let Ω be a sample space with field $\mathcal{F}_0$ and measure $\mathbb{P}$. A set $A \subseteq \Omega$ is said to be $\mathbb{P}^*$-*measureable* if $\mathbb{P}^*(A \cap E) + \mathbb{P}^*(A^C \cap E) = \mathbb{P}^*(E)$ for all $E \subseteq \Omega$. This is called Carathéodory's criterion in real analysis. Intuitively, this means that for any set E the boundary between $A \cap E$ and $A^c \cap E$ does not exist, then A is $\mathbb{P}^*$-measurable. In a sense, this means that the boundary of E with the rest of the space may not be perfectly coverable, but the boundary within E (cast by A and A^c) is perfectly coverable. We define $\mathcal{M}$ to be the class of all $\mathbb{P}^*$ measurable subsets of Ω.

Proposition 6: Let Ω be a sample space with field $\mathcal{F}_0$ and measure $\mathbb{P}$. Then $\mathcal{M}$ is a σ-field, and $\mathbb{P}^*$ is countably additive on $\mathcal{M}$ and $\mathbb{P}^*$ is a probability measure. Therefore, $(\Omega, \mathcal{M}, \mathbb{P}^*)$ is a probability space.

Proof. The proof is omitted. However, we can show conditions (1') and (2') to see it is a probability measure. (1') $\mathbb{P}^*(A) \geq 0$ since it is a sum of values of $\mathbb{P}$ and $\mathbb{P}(B) \geq 0$ for all $B \in \mathcal{F}_0$. (2') $\Omega \in \mathcal{F}_0$ by definition and so $\mathbb{P}^*(\Omega) = \mathbb{P}(\Omega) = 1$. □

Proposition 7: Let Ω be a sample space with field $\mathcal{F}_0$ and measure $\mathbb{P}$. Then $\mathcal{F}_0 \subseteq \mathcal{M}$.

Proof. Let $A \in \mathcal{F}_0$. For any $E \subseteq \Omega$, we must show that $\mathbb{P}^*(A \cap E) + \mathbb{P}^*(A^c \cap E) = \mathbb{P}^*(E)$. Let $\varepsilon > 0$. Let $A_1, A_2, \ldots \in \mathcal{F}_0$ be a cover of E, i.e., such that $E \subseteq \bigcup_{n=1}^{\infty} A_n$, and such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq \mathbb{P}^*(E) + \varepsilon$. Note such a cover exists by the definition of $\mathbb{P}^*(E)$. Define $B_n := A_n \cap A \in \mathcal{F}_0$ and $C_n = A_n \cap A^c \in \mathcal{F}_0$. By countable additivity (and since B_n and C_n are disjoint), we see that $\mathbb{P}(B_n) + \mathbb{P}(C_n) = \mathbb{P}(A_n)$. Moreover, we see then that

$$E \cap A \subseteq \bigcup_{n=1}^{\infty} B_n \qquad \text{and} \qquad E \cap A^c \subseteq \bigcup_{n=1}^{\infty} C_n.$$

This implies that

$$\mathbb{P}^*(E \cap A) \leq \mathbb{P}^*(\bigcup_{n=1}^{\infty} B_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_n) \qquad \text{and} \qquad \mathbb{P}^*(E \cap A^c) \leq \mathbb{P}^*(\bigcup_{n=1}^{\infty} C_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(C_n)$$

since $\{B_n\}_{n=1}^{\infty}$ is one such cover and by Boole's inequality. Thus,

$$\mathbb{P}^*(E \cap A) + \mathbb{P}^*(E \cap A^c) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_n) + \sum_{n=1}^{\infty} \mathbb{P}(C_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq \mathbb{P}^*(E) + \varepsilon$$

Since this holds for all $\varepsilon > 0$, taking a limit we see that $\mathbb{P}^*(E \cap A) + \mathbb{P}^*(E \cap A^c) \leq \mathbb{P}^*(E)$.

On the other hand, we trivially have that $\mathbb{P}^*(E \cap A) + \mathbb{P}^*(E \cap A^c) \geq \mathbb{P}^*(E)$ since the union of the smallest covers of $E \cap A$ and $E \cap A^c$ necessarily also forms a cover of E. By the definition of $\mathbb{P}^*(E)$, the cover used must have a smaller (summed) measure than the union of the covers of E.

Combining the two results, we get $\mathbb{P}^*(A \cap E) + \mathbb{P}^*(A^c \cap E) = \mathbb{P}^*(E)$, and so $A \in \mathcal{M}$. This shows that $\mathcal{F}_0 \subseteq \mathcal{M}$. □

Definition. π-system: A class $\mathcal{P} \subseteq \mathfrak{P}(\Omega)$ of subsets of Ω is called a *π-system* if whenever $A, B \in \mathcal{P}$ then $A \cap B \in \mathcal{P}$ (i.e., $\mathcal{P}$ is closed under finite intersection).

Definition. λ-system: A class $\mathcal{L} \subseteq \mathfrak{P}(\Omega)$ of subsets of Ω is called a *λ-system* if the following are satisfied:

(1) $\Omega \in \mathcal{L}$.

(2) if $A \in \mathcal{L}$ then $A^c \in \mathcal{L}$.

(3) If $A_1, A_2, \ldots \in \mathcal{L}$ are disjoint, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$.

The only difference between a λ-system and a σ-field is that a λ-system requires the union to be of disjoint sets (whereas an σ-field does not have this requirement).

Proposition 8: A class $\mathcal{P}$ that is both a π-system and a λ-system is a σ-field.

Proof. We only need to show that $\mathcal{C}$ is closed under countable (not necessarily disjoint) union. Let $A_1, A_2, \ldots \in \mathcal{C}$. Define $B_1 = A_1$, $B_2 = A_2 \setminus A_1 = A_2 \cap A_1^c$, and in general

$$B_n = A_n \setminus (\bigcup_{i=1}^{n-1} A_i) = A_n \cap A_{n-1}^c \cap A_{n-2}^c \cap \cdots \cap A_1^c.$$

Note that each B_n is in $\mathcal{C}$ since each one is formed by finite intersections (and $\mathcal{C}$ is closed under complementation). Note then that each $B_1, B_2, \ldots$ are all also disjoint and so $\bigcup_{n=1}^{\infty} B_n \in \mathcal{C}$. We see then that $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n \in \mathcal{C}$, and so $\mathcal{C}$ is a σ-field. □

Proposition 9: If a λ-system $\mathcal{L}$ contains both $A \in \mathcal{L}$ and $A \cap B \in \mathcal{L}$ for some set $B \subseteq \Omega$, then $A \cap B^c \in \mathcal{L}$.

Proof. Note that since $A \in \mathcal{L}$ then $A^c \in \mathcal{L}$. Moreover, A^c and $A \cap B$ are disjoint. So

$$A^c \cup (A \cap B) \in \mathcal{L} \quad \Longrightarrow \quad (A^c \cup (A \cap B))^c = A \cap (A^c \cup B^c) = (A \cap A^c) \cup (A \cap B^c) = A \cap B^c \in \mathcal{L},$$

as desired. □

Lecture 4 09/17

Theorem 10. π-λ Theorem: Also called the monotone class theorem. If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system, and $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.

Proof. Let $L(\mathcal{P}) = \bigcap\{\mathcal{L}' : \mathcal{P} \subseteq \mathcal{L}', \mathcal{L}' \text{ is a } \lambda\text{-system}\} \subseteq \mathfrak{P}(\Omega)$ be the intersection of all the λ-systems containing $\mathcal{P}$. One can check that $L(\mathcal{P})$ is also a λ-system containing $\mathcal{P}$. For $A \subseteq \Omega$, define $\mathcal{G}_A$ to be the class of sets $B \subseteq \Omega$ such that $A \cap B \in L(\mathcal{P})$.

Step 1. We will show that for any $A \in L(\mathcal{P})$, then $\mathcal{G}_A$ is a λ-system. (1) Note that $A \cap \Omega = A \in \mathcal{L}(P)$ therefore $\Omega \in \mathcal{G}_A$. (2) If $B \in \mathcal{G}_A$, then $L(\mathcal{P})$ is a λ-system containing both A and $A \cap B$ (by definition of $\mathcal{G}_A$), and so by the previous proposition $A \cap B^c \in \mathcal{L}(\mathcal{P})$ and thus $B^c \in \mathcal{G}_A$. (3) If $B_1, B_2, \ldots \in \mathcal{G}_A$ are disjoint, then $A \cap B_1, A \cap B_2, \ldots \in L(\mathcal{P})$ are disjoint. Therefore $\bigcup_{i=1}^{\infty}(A \cap B_i) = A \cap (\bigcup_{i=1}^{\infty} B_i) \in L(\mathcal{P})$ since $B_1, B_2, \ldots$ are disjoint and $L(\mathcal{P})$ is a λ-system. Thus $\bigcup_{n=1}^{\infty} B_n \in \mathcal{G}_A$, which shows that $\mathcal{G}_A$ is a λ-system.

Step 2. Next, we will show that if $A \in \mathcal{P}$ then $L(\mathcal{P}) \subseteq \mathcal{G}_A$. For any $B \in \mathcal{P}$, then $A \cap B \in \mathcal{P} \subseteq L(\mathcal{P})$ since $\mathcal{P}$ is a π-system. Thus, we see that $B \in \mathcal{G}_A$ by the definition of $\mathcal{G}_A$. Thus $\mathcal{P} \subseteq \mathcal{G}_A$ and since $\mathcal{G}_A$ is a λ-system, then since $L(\mathcal{P})$ is the smallest λ-system containing $\mathcal{P}$, we have that $L(\mathcal{P}) \subseteq \mathcal{G}_A$.

Step 3. Finally, we will show that if $B \in L(\mathcal{P})$ then $L(\mathcal{P}) \subseteq \mathcal{G}_B$. For any $A \in \mathcal{P}$, then $B \in L(\mathcal{P}) \subseteq \mathcal{G}_A$ from step 2. Thus $A \cap B \in L(\mathcal{P})$ by definition of $\mathcal{G}_A$, and so this also means $A \in \mathcal{G}_B$ by definition of $\mathcal{G}_B$. This shows that $\mathcal{P} \subseteq \mathcal{G}_B$, and therefore that $L(\mathcal{P}) \subseteq \mathcal{G}_B$ since $L(\mathcal{P})$ is the smallest λ-system containing $\mathcal{P}$.

Now, we are ready to show the main result. For any $A, B \in L(\mathcal{P})$, we have $A \in \mathcal{G}_B$ by step 3 and so by definition of $\mathcal{G}_B$ we also have $A \cap B \in \mathcal{L}(P)$. Thus $L(\mathcal{P})$ is also a π-system. Then by proposition 8, $L(\mathcal{P})$ is a σ-field. Moreover, we see that

$$\mathcal{P} \subseteq \sigma(\mathcal{P}) = L(\mathcal{P}) \subseteq \mathcal{L}.$$

The equality follows since $L(\mathcal{P})$ is a σ-field containing $\mathcal{P}$ and $\sigma(\mathcal{P})$ is the smallest σ-field containing $\mathcal{P}$, this shows $\sigma(\mathcal{P}) \subseteq L(\mathcal{P})$. But also, $\sigma(\mathcal{P})$ is a λ-system containing $\mathcal{P}$ and $L(\mathcal{P})$ is the smallest λ-system containing $\mathcal{P}$, so $L(\mathcal{P}) \subseteq \sigma(\mathcal{P})$. The last containment follows since $\mathcal{L}$ is some λ-system. □

Corollary 11: Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability measures that agree on a π-system $\mathcal{P}$. That is $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \mathcal{P}$. Then they agree on $\sigma(\mathcal{P})$.

Proof. Let $\mathcal{L} = \{A \in \sigma(\mathcal{P}) : \mathbb{P}_1(A) = \mathbb{P}_2(A)\}$ be the class of sets where $\mathbb{P}_1$ and $\mathbb{P}_2$ agree. We check $\mathcal{L}$ is a λ-system. (1) Note clearly $\Omega \in \sigma(P)$ and $\mathbb{P}_1(\Omega) = 1 = \mathbb{P}_2(\Omega)$ so $\Omega \in \mathcal{L}$. (2) Suppose $A \in \mathcal{L}$, then $\mathbb{P}_1(A^c) = 1 - \mathbb{P}_1(A) = 1 - \mathbb{P}_2(A) = \mathbb{P}_2(A^c)$ and so $A^c \in \mathcal{L}$. (3) Suppose $A_1, A_2, \ldots \in \mathcal{L}$ are disjoint, then

$$\mathbb{P}_1(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}_1(A_n) = \sum_{n=1}^{\infty} \mathbb{P}_2(A_n) = \mathbb{P}_2(\bigcup_{n=1}^{\infty} A_n)$$

by countably-additivity of probability measures, so $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$. We see then that $\mathcal{L}$ is a λ-system.

Then, by applying the π-λ theorem, we see that $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. Then by the definition of $\mathcal{L}$, we see that $\mathbb{P}_1$ and $\mathbb{P}_2$ agree on $\sigma(\mathcal{P})$. □

Theorem 12. Existence and Uniqueness of Probability Measures: Let $\mathcal{F}_0$ be a field on Ω and $\mathbb{P}$ be a set function defined on $\mathcal{F}_0$ such that it satisfies the probability axioms on $\mathcal{F}_0$. I.e.,

(1) $0 \leq \mathbb{P}(A) \leq 1$ for all $A \in \mathcal{F}_0$.

(2) $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

(3) If $A_1, A_2, \ldots \in \mathcal{F}_0$ are disjoints sets, and if $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}_0$ then $\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$.

That is, if $\mathcal{F}_0$ satisfies all of the properties for a σ-field (but $\mathcal{F}_0$ is not closed under countable union), then there exists a unique probability measure $\mathbb{Q}$ on $\sigma(\mathcal{F}_0)$ such that whenever $A \in \mathcal{F}_0$ (and therefore $A \in \sigma(\mathcal{F}_0)$), then $\mathbb{Q}(A) = \mathbb{P}(A)$. We say that $\mathbb{Q}$ is an extension of $\mathbb{P}$. That is, for every field and probability measure, there is exists a unique extension of the probability to the σ-field generated by the field.

Proof. Since $\mathcal{F}_0 \subseteq \mathcal{M}$ and $\mathcal{M}$ is a σ-field then $\mathcal{M}$ is part of the intersection which yields $\sigma(\mathcal{F}_0)$. Therefore, since $\mathbb{P}^*$ is a probability measure on the σ-field $\mathcal{M}$ we see that $\mathbb{P}^*$ must be

a probability measure on $\sigma(\mathcal{F}_0) \subseteq \mathcal{M}$. Moreover, for any $A \in \mathcal{F}_0$, the smallest cover of A is A itself, and so $\mathbb{P}^*(A) = \mathbb{P}(A)$. This shows the existence of an extension of $(\Omega, \mathcal{F}_0, \mathbb{P})$.

Uniqueness follows from corollary 11‘ since $\mathcal{F}_0$ is a π-system. Therefore, any probability measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ that agree on $\mathcal{F}_0$, then they also agree on $\sigma(\mathcal{F}_0)$, and so $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are actually the same probability measure. □

Remark: Members of σ-fields are typically hard to describe, making it also hard to describe probability measures on such a σ-field directly. The idea of the above theorem is to define a probability measure on a smaller/simpler collection of sets and extend the result to the whole σ-field.

Definition. Lebesgue Measure: Define $\Omega = (0, 1]$ and use the Borel-σ-field $\mathscr{B}$. Recall $\mathscr{B}_0 = \mathscr{B}_0((0, 1]) = \{\text{finite unions of disjoint intervals } (a, b] \text{ on } (0, 1]\}$ and $\mathscr{B} = \sigma(\mathscr{B}_0)$. Define a set function λ on $\mathscr{B}_0$ such that $\lambda((a, b]) = b - a$ and for other members of $\mathscr{B}_0$ it is the sum of λ applied to each disjoint interval. One can check that λ satisfies the probability axioms on $\mathscr{B}_0$. Then by theorem 12, there is a unique extension of λ which is a probability measure defined on $\mathscr{B}$ since $\mathscr{B} = \sigma(\mathscr{B}_0)$ (we will continue to denote it by λ for simplicity). We call λ the Lebesgue measure on $\mathscr{B}$.

Moreover, the Lebesgue measure is the only probability measure on $((0, 1], \mathscr{B})$ such that the measure of an interval is equal to its length for all measures.

Remark: Any measure can be transformed into a probability measure by $\mu'(A) = \mu(A)/\mu(\Omega)$ assuming $\mu(\Omega) < \infty$. However, this implies that the Lebesgue measure defined on $\mathbb{R}$ is fundamentally different, since $\lambda(\mathbb{R}) = \infty$.

Definition. Null Set: For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a set $A \in \mathcal{F}$ is called a null set if $\mathbb{P}(A) = 0$.

Definition. Completeness: A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete if whenever $A \subseteq B$ where $B \in \mathcal{F}$ but A may or may not be in $\mathcal{F}$, and $\mathbb{P}(B) = 0$ (i.e., B is a null set), then $\mathbb{P}(A) = 0$ (i.e., A is a null set and $A \in \mathcal{F}$).

Proposition 13: If $(\Omega, \mathcal{F}, \mathbb{P})$ is complete, then for a set A', if there exists an $A \in \mathcal{F}$ such that $A \Delta A' \subseteq B$ for some with $\mathbb{P}(B) = 0$, then $A' \in \mathcal{F}$ and $\mathbb{P}(A') = \mathbb{P}(A)$. Recall $A \Delta A' = (A \cap A'^c) \cup (A^c \cap A')$ is the symmetric difference.

Proposition 14: For any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there exists a complete probability space $(\Omega, \mathcal{F}', \mathbb{P}')$ which extends $(\Omega, \mathcal{F}, \mathbb{P})$. That is, $\mathcal{F} \subseteq \mathcal{F}'$ and $\mathbb{P}'(A) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$.

Proof. Recall the outer measure $\mathbb{P}^*$ defined earlier which is a probability measure on the class $\mathcal{M}$ of all $\mathbb{P}^*$-measurable sets. Recall also that A is $\mathbb{P}^*$-measurable if and only if $\mathbb{P}^*(A \cap E) + \mathbb{P}^*(A^c \cap E) = \mathbb{P}^*(E)$ for all $E \subseteq \Omega$.

We will show that $(\Omega, \mathcal{M}, \mathbb{P}^*)$ is a complete probability space. Note also that $\mathcal{F} \subseteq \mathcal{M}$ and by definition $\mathbb{P}^*$ is an extension of $\mathbb{P}$. Let $\mathbb{P}^*(B) = 0$ and $A \subseteq B$. For any $E \subseteq \Omega$, note that

$$\mathbb{P}^*(\underbrace{A \cap E}_{\subseteq A \subseteq B}) + \mathbb{P}^*(\underbrace{A^c \cap E}_{\subseteq B}) \leq \mathbb{P}^*(B) + \mathbb{P}^*(E) \leq \mathbb{P}^*(E)$$

Note this follows by the monotonicity of $\mathbb{P}^*$, which follows since for any $A \subseteq B$, any cover of B is also a cover of A, and so by taking the infimum, we have $\mathbb{P}^*(A) \leq \mathbb{P}^*(B)$. Note $\mathbb{P}^*(E) \leq \mathbb{P}^*(A \cap E) + \mathbb{P}^*(A^c \cap E)$ since any cover of E is also a cover of both $A \cap E$ and $A^c \cap E$. Thus, $A \in \mathcal{M}$ and $\mathbb{P}^*(A) = 0$ follows by the monotonicity of $\mathbb{P}^*$. □

Definition. Conditional Probability: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $A, B \in \mathcal{F}$. Then the conditional probability of B given A is $\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$.

Proposition 15. Chain Rule: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. An immediate result of conditional probability is the chain rule. For $A_1, A_2, \ldots, A_n \in \mathcal{F}$, then

$$\mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}(A_n|A_1 \cap \cdots \cap A_{n-1})$$

Proposition 16. Law of Total Probability: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose $A_1, A_2, \ldots \in \mathcal{F}$ is a partition of Ω, i.e., $A_1, A_2, \ldots$ are disjoint and $\bigcup_{n=1}^{\infty} A_n = \Omega$. Then for any event $B \in \mathcal{F}$, we have $\mathbb{P}(B) = \sum_{n=1}^{\infty} \mathbb{P}(B|A_n)\mathbb{P}(A_n)$.

Lecture 5 09/19

Theorem 17: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For a sequence of events $A_1, A_2, \ldots \in \mathcal{F}$, then we have

(1) $\mathbb{P}(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mathbb{P}(A_n) \leq \limsup_{n\to\infty} \mathbb{P}(A_n) \leq \mathbb{P}(\limsup_{n\to\infty} A_n)$

(2) If $\lim_{n\to\infty} A_n = A$ then $\lim_{n\to\infty} \mathbb{P}(A_n) \to \mathbb{P}(A)$

Proof. (1) Define $B_n = \bigcap_{k=n}^{\infty} A_k$ and $C_n = \bigcup_{k=n}^{\infty} A_k$. Then we note B_n is monotonically increasing since we intersect with fewer events, i.e., $B_1 \subseteq B_2 \subseteq \cdots$. We also note C_n is monotonically decreasing since we union with fewer events, i.e., $C_1 \supseteq C_2 \supseteq \cdots$. Thus, $\lim_{n\to\infty} B_n = \liminf_{n\to\infty} A_n$ and $\lim_{n\to\infty} C_n = \limsup_{n\to\infty} A_n$. We also see $B_n \subseteq A_n \subseteq C_n$. And so,

$$\liminf_{n\to\infty} \mathbb{P}(A_n) \geq \lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}(\lim_{n\to\infty} B_n) = \mathbb{P}(\liminf_{n\to\infty} A_n)$$

and

$$\limsup_{n\to\infty} \mathbb{P}(A_n) \leq \lim_{n\to\infty} \mathbb{P}(C_n) = \mathbb{P}(\lim_{n\to\infty} C_n) = \mathbb{P}(\limsup_{n\to\infty} A_n)$$

(2) follows immediately from (1). □

Definition. Independence of Events: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two events A, B are called independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. When $\mathbb{P}(A) > 0$, then this is equivalent to $\mathbb{P}(B|A) = \mathbb{P}(B)$. We write $A \perp\!\!\!\perp B$ in this case. For n events $A_1, \ldots, A_n$, they are called (mutually) independent if $\mathbb{P}(\bigcap_{i\in I} A_i) = \prod_{i\in I} \mathbb{P}(A_i)$ holds for all $I \subseteq \{1, 2, \ldots, n\}$. An infinite (possibly uncountable) number of events $\{A_t\}_{t\in T}$ are said to be independent if every finite subset $\{A_{i_1}, A_{i_2}, \ldots, A_{i_n}\} \subseteq \{A_t\}_{t\in T}$ is mutually independent.

Note: For $n \geq 3$, mutual independence is different (and stronger) than the pairwise independence of all pairs (i.e., $A_i \perp\!\!\!\perp A_j$ for all $i, j \in \{1, \ldots, n\}$ with $i \neq j$). Mutual independence is also different (and stronger) than the condition $\mathbb{P}(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$.

Exercise: Find examples of events $A_1, \ldots, A_n$ that are pairwise independent but not mutually independent. Find examples of events $B_1, \ldots, B_n$ that have $\mathbb{P}(\bigcap_{i=1}^n B_i) = \prod_{i=1}^n \mathbb{P}(B_i)$, but are not mutually independent.

Definition. Independence Between Collections of Events: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A (potentially infinite) number of classes of events (i.e., sets of events) $\{\mathscr{A}_\theta\}_{\theta \in \Theta}$ such that $\mathscr{A}_\theta \subseteq \mathcal{F}$ for all $\theta \in \Theta$ are called independent if every collection $\{A_\theta : A_\theta \in \mathscr{A}_\theta\}_{\theta \in \Theta}$ is mutually independent. I.e., every collection that can be formed by picking one element from $\mathscr{A}_{\theta_1}$, one element from $\mathscr{A}_{\theta_2}$, so and so forth is mutually independent. Two σ-fields, $\mathcal{F}$ and $\mathcal{G}$, are called independent, if they form independent classes of events. That is, for every $A \in \mathcal{F}$ and $B \in \mathcal{G}$, then $A \perp\!\!\!\perp B$. Similarly, if $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are σ-fields, then they are independent if every set of events $A_1, \ldots, A_n$ is mutually independent for all $A_1 \in \mathcal{F}_1, \ldots, A_n \in \mathcal{F}_n$. This continues naturally for infinitely many σ-fields.

Proposition 18: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $\mathscr{A}_\theta$ for $\theta \in \Theta$ are independent and each $\mathscr{A}_\theta$ is a π-system, then $\sigma(\mathscr{A}_\theta)$ for $\theta \in \Theta$ are independent. That is, σ-fields generated by independent π-systems are independent.

Proof. It is sufficient to show that for any $n \in \mathbb{N}$ and any $A_1, \in \sigma(\mathscr{A}_1), \ldots, \mathscr{A}_n \in \sigma(\mathscr{A}_n)$ we have $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdots \mathbb{P}(A_n)$. Fix $n \in \mathbb{N}$ and $A_2 \in \mathscr{A}_2, \ldots, A_n \in \mathscr{A}_n$. Define $\mathcal{L}_1 = \{A_1 \in \sigma(\mathscr{A}_1) : \mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^n \mathbb{P}(A_i)\}$. Note then that $\mathscr{A}_1 \subseteq \mathcal{L}_1$ since by assumption $\mathscr{A}_1, \ldots, \mathscr{A}_n$ are independent. Moreover, one can check that $\mathcal{L}_1$ is a λ-system. Then by the π-λ theorem, we have that $\sigma(\mathscr{A}_1) \subseteq \mathcal{L}_1$, hence $\mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^n \mathbb{P}(A_i)$ for any $A_1 \in \sigma(\mathscr{A}_1, A_2 \in \mathscr{A}_2, \ldots, A_n \in \mathscr{A}_n$. This shows that $\sigma(A_1), A_2, \ldots, A_n$ are independent classes. We can then repeat this process, fixing $A_1 \in \sigma(\mathscr{A}_1), A_3 \in \mathscr{A}_3, \ldots, A_n \in \mathscr{A}_n$ and defining $\mathcal{L}_2 = \{A_2 \in \sigma(\mathscr{A}_n) : \mathbb{P}(A_1 \cap \cdots \cap A_n) = \prod_{i=1}^n \mathbb{P}(A_i)\}$. This will show us that $\sigma(\mathscr{A}_1), \sigma(\mathscr{A}_2), A_3, \ldots, A_n$ are independent classes. We then repeat this for all remaining A_i, to show that $\sigma(\mathscr{A}_1), \ldots, \sigma(\mathscr{A}_n)$ are independent classes. □

Exercise: Show that the π-system condition in the above proposition is necessary.

Lecture 6 09/24

Proposition 19: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let

$$\begin{matrix} A_{11} & A_{12} & \cdots \\ A_{21} & A_{22} & \cdots \\ \vdots & \vdots & \ddots \end{matrix}$$

be an array of independent events. If $\mathcal{F}_i = \sigma(\{A_{i,j} : j \in \mathbb{N}\})$ is the σ-field generated by the i-th row, then $\mathcal{F}_i, \mathcal{F}_2, \ldots$ are independent. This could also be extended to the uncountable case, but here we assume the array has cardinality $|\mathbb{N} \times \mathbb{N}|$. This condition differs from the

previous proposition since we don't only have independence within one generating class, but also between all generating classes.

Proof. Let $\mathscr{A}_i$ be the class of all finite intersections of elements of the ith row. One can check that $\mathscr{A}_i$ is then a π-system, and obviously $\sigma(\mathscr{A}_i) = \mathcal{F}_i$. It suffices to show then that $\{\mathscr{A}_i\}_{i=1}^{\infty}$ is independent. That is, for any finite set of indices I, and any $i \in I$, pick $C_i \in \mathscr{A}_i$, we must then show that $\mathbb{P}(\bigcap_{i\in I} C_i) = \prod_{i\in I} \mathbb{P}(C_i)$. Note since $C_i \in \mathscr{A}_i$, there is a finite set J_i of indices such that $C_i = \bigcap_{j\in J_i} A_{ij}$.

$$\mathbb{P}(\bigcap_{i\in I} C_i) = \mathbb{P}(\bigcap_{i\in I}\bigcap_{j\in J_i} A_{ij}) = \prod_{i\in I}\prod_{j\in J_i} \mathbb{P}(A_{ij}) = \prod_{i\in I} \mathbb{P}(\bigcap_{j\in J_i} A_i) = \prod_{i\in I} \mathbb{P}(C_i)$$

Thus $\mathscr{A}_1, \mathscr{A}_2, \ldots$ are independent, so by proposition 18, we see that $\mathcal{F}_1, \mathcal{F}_2, \ldots$ are independent. □

Theorem 20. 1st Borel-Contelli Lemma: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A_1, A_2, \ldots \in \mathcal{F}$ be a sequence of events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$

Proof. Recall $\{A_n \text{ i.o.}\} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k$. Then, $\limsup_{n\to\infty} A_n \subseteq \bigcup_{k=m}^{\infty} A_k$ for any m. So we see,

$$\mathbb{P}(\limsup_{n\to\infty} A_n) \leq \mathbb{P}(\bigcup_{k=m}^{\infty} A_n) \leq \sum_{k=m}^{\infty} \mathbb{P}(A_k)$$

for all $m \in \mathbb{N}$. Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, we know that $\sum_{k=m}^{\infty} \mathbb{P}(A_k) \to 0$ as $m \to \infty$. Thus, by taking the limit $m \to \infty$, we see that $\mathbb{P}(\limsup_{n\to\infty} A_n) \to 0$ by the squeeze theorem. Thus $\mathbb{P}(A_n \text{ i.o.}) = 0$ by taking the limit. □

Theorem 21. 2nd Borel-Contelli Lemma: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A_1, A_2, \ldots$ be an independent sequence of events. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Proof. Recall $\{A_n \text{ i.o.}\} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k$. It suffices to show then that

$$1 - \mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}((\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k)^c) = 0 \quad \iff \quad \mathbb{P}(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k^c) = 0$$

by De Morgan's laws. We will show that $\mathbb{P}(\bigcap_{k=n}^{\infty} A_k^c) = 0$ for all n. In particular, let n be given. Then for any $j = 1, 2, \ldots$, note

$$\mathbb{P}(\bigcap_{k=n}^{n+j} A_k^c) = \prod_{k=n}^{n+j} \mathbb{P}(A_k^c) = \prod_{k=n}^{n+j} (1 - \mathbb{P}(A_k)) \leq \prod_{k=n}^{n+j} \exp(-\mathbb{P}(A_k)) = \exp\left(-\sum_{k=n}^{n+j} \mathbb{P}(A_k)\right)$$

since $1 - x \leq e^{-x}$ for all x (the inequality is strict if $x \neq 0$). Then, since $\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \infty$, we see that if $\sum_{k=n}^{n+j} \mathbb{P}(A_k) \to \infty$ as $j \to \infty$, since we only ever drop the first $n < \infty$ terms, and so $\sum_{k=1}^{n-1} \mathbb{P}(A_k) < \infty$. This implies then that $\mathbb{P}(\bigcap_{k=n}^{n+j} A_k^c) \to 0$ as $j \to \infty$. Finally, this implies that $\mathbb{P}(\bigcap_{k=n}^{\infty} A_k^c) = 0$. But then, $\mathbb{P}(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k^c) \leq \sum_{n=1}^{\infty} \mathbb{P}(\bigcap_{k=n}^{\infty} A_k^c) = \sum_{n=1}^{\infty} 0 = 0$, as desired. □

Exercise: Find a counterexample where $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = 1$ but $\mathbb{P}(A_n \text{ i.o.}) \neq 1$ by picking a sequence $A_1, A_2, \ldots$ which is not necessarily independent.

Example: Suppose you have a box containing m balls. Each time you draw a ball at random, and put it back in the box along with an extra ball. Thus any ball will always remain in the box, and the box will have increasingly many new balls. For any given ball, how many times will it be picked in total?

Let A_n denote the event that the ball was picked at time n (where we start counting draws from $n = 1$). Assume the ball was first put into the box immediately prior to time K (if it was in the box since the start, then $K = 1$). Then

$$\mathbb{P}(A_n) = \begin{cases} 0 & \text{if } n < K \\ \frac{1}{m+n-1} & \text{if } n \geq K \end{cases}$$

Note also that $\{A_n\}_{n=1}^{\infty}$ is independent. Note also that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ by the divergence of the harmonic series. Then by the 2nd Borel-Contelli lemma, $\mathbb{P}(A_n \text{ i.o.}) = 1$. Thus with probability 1, the chosen ball will be picked infinitely many times.

Example: Suppose you have a box containing 1 ball. This time, each time you draw a ball at random, and put it back in the box along with enough new balls so that there are exactly 2^n new balls. Thus any ball will always remain in the box, and the box will have increasingly many new balls. For any given ball, how many times will it be picked in total?

Let A_n denote the event that the ball was picked at time n (where we start counting draws from $n = 1$). Assume the ball was first put into the box immediately prior to time K (if it was in the box since the start, then $K = 1$). Then

$$\mathbb{P}(A_n) = \begin{cases} 0 & \text{if } n < K \\ \frac{1}{2^n} & \text{if } n \geq K \end{cases}$$

Then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, and so by the 1st Borel-Contelli lemma, $\mathbb{P}(A_n \text{ i.o.}) = 0$. So with probability 1, the chosen ball will be picked only finitely many times.

Remark. Zero-One Laws: Many results in probability theory take the form of zero-one laws, i.e., $\mathbb{P}(A_n \text{ i.o.}) = 0$ or $\mathbb{P}(A_n \text{ i.o.}) = 1$. They assert that certain events take place with probability 1 (almost always happen) or probability 0 (almost never happen). The above are some examples of zero-one laws.

Example. DTMC Recurrence: A state i in a discrete-time Markov chain (DTMC) is said to either be recurrent or transient. If i is recurrent, then the chain will visit i infinitely many times (starting from state i) with probability 1, and if i is transient, then the chain will only visit i finitely many times (starting from state i) with probability 1. Let N_i denote the number of visits to state i after infinitely many steps, starting from the initial state i. Then if i is recurrent $\mathbb{P}(N_i = \infty) = 1$ or if i is transient $\mathbb{P}(N_i = \infty) = 0$. Thus recurrence is a zero-one law.

Definition. Tail σ-field: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A_1, A_2, \ldots \in \mathcal{F}$ be events. Then the tail σ-field of $\{A_n\}_{n=1}^{\infty}$ is $\mathscr{T} = \bigcap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, \ldots)$. We call elements of $\mathscr{T}$ tail events.

Example: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A_1, A_2, \ldots \in \mathcal{F}$ be an independent sequence of events. Let $\mathscr{T}$ be their tail σ-field. Consider $\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$. Then since $\{\bigcup_{k=n}^{\infty} A_i\}_{n=1}^{\infty}$ is a decreasing sequence of events. So we can write

$$\limsup_{n\to\infty} A_n = \bigcap_{n=m}^{\infty} \bigcup_{k=n}^{\infty} A_n \in \sigma(A_m, A_{m+1}, \ldots) \in \mathscr{T}$$

for any m. We can similarly show that $\liminf_{n\to\infty} A_n \in \mathscr{T}$

Remark: We'll now introduce a mechanism which leads to many zero-one laws.

Theorem 22. Kolmogorov's Zero-One Law: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $A_1, A_2, \ldots \in \mathcal{F}$ be an independent sequence of events. Let $\mathscr{T}$ be their tail σ-field. Then any event $A \in \mathscr{T}$ either has probability $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

Proof. By independence of $A_1, A_2, \ldots$, we know that $\sigma(\{A_1\}), \ldots, \sigma(\{A_{n-1}\}), \sigma(\{A_n, A_{n+1}, \ldots\})$ are all independent for any n. For $A \in \mathscr{T}$, $A \in \sigma(A_n, A_{n+1}, \ldots)$ and hence, A is independent of $A_1, \ldots, A_{n-1}$. Since this holds for all n, $\sigma(A)$ and $\sigma(A_1, A_2, \ldots)$ are independent. However, we also have $A \in \mathscr{T} \subseteq \sigma(A_1, A_2, \ldots)$. Obviously $A \in \sigma(A)$ and $A \in \sigma(A_1, A_2, \ldots)$, which by the independence of the two σ-fields implies A is independent of itself. This implies $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A) \cdot \mathbb{P}(A)$, whose only solutions are $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. □

Chapter 2 Random Variables

Lecture 7 09/26

Example: Consider the sample space $\Omega = \{\text{sunny}, \text{rainy}, \text{cloudy}, \ldots\}$ describing tomorrow's weather. Then we can map these to some variable X with domain $\{1, 2, 3, \ldots\}$, to make talking about these easier. E.g., we can take probabilities $P\mathbb{P}(X = 1), \mathbb{P}(X \in [1, 2])$, etc.

Definition. Measurable Mappings: Let $(\Omega, \mathcal{F})$ and $(S, \mathscr{A})$ be measurable spaces. A mapping $X : \Omega \to \mathscr{A}$ is said to be measurable if for any $A \in \mathscr{A}$, if the pre-image $X^{-1}(A) := X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ of A under X, is such that $X^{-1}(A) \in \mathcal{F}$. That is, the pre-image of any measurable set is measurable. If $(S, \mathscr{A}) = (\mathbb{R}, \mathscr{B})$, then X is called a *random variable*.

Notation. Random Variable Notation: Suppose $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathscr{B})$ are measurable spaces and $X : \Omega \to \mathbb{R}$ is a random variable. When we discuss events defined by the value of X, we often use the shorthand $\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B)$. We also often use the shorthand $\mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$.

Example. Discrete Probability Space: If Ω is a countable sample space, then $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$ is said to be a discrete probability space. Then, any mapping $X : \Omega \to \mathbb{R}$ is a random variable. As always, the probability measure induced by X is $\mathbb{P}(X \in B) = \sum_{b \in B} \mathbb{P}(X = b) = \mathbb{P}(X^{-1}(B))$.

Example. Indicator Random Variable: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let A be an event in $\mathcal{F}$. The indicator of A is defined as

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \text{ i.e., } A \text{ "happens"} \\ 0 & \text{if } \omega \notin A \end{cases}$$

Then $\mathbb{1}_A$ is a random variable.

Definition. Distribution of a Random Variable: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then X *induces* a probability measure μ on $(\mathbb{R}, \mathscr{B})$ by setting $\mu(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$. Using this definition, μ is called the distribution of X.

Exercise: Check that μ as defined above is indeed a probability measure.

Remark: The probability measure μ induced by the random variable X is also called a *push-forward measure*, since it "pushes" the probability measure $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$ to the measurable space $(\mathbb{R}, \mathscr{B})$.

Definition. Cumulative Distribution Function: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The function $F : \mathbb{R} \to [0, 1]$ defined through $F(x) = P(X \leq x) = \mu((-\infty, x])$ is called the (cumulative) distribution function (often abbreviated c.d.f.). Since $\{(-\infty, x], x \in \mathbb{R}\}$ is a π-system, it generates $\mathscr{B}$, and thus by extension, $F(x)$ uniquely determines μ. I.e., the distribution of a random variable is fully characterized by its (cumulative) distribution function.

Proposition 23: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The following are some properties of the distribution function F of X.

(1) F is non-decreasing.

(2) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

(3) F is right-continuous, which is to say $F(a) = \lim_{x \to a^-} F(x)$ for all $a \in \mathbb{R}$.

(4) $\lim_{x \to a^+} F(x) = \mathbb{P}(X < a) = \mu((-\infty, a))$ we also write $F(x^-) = \mathbb{P}(X < x)$.

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

Proof. (1) Follows by the monotonicity of probability measures.

(2) Note that $\lim_{x \to \infty}(-\infty, x) = \mathbb{R}$ and $\lim_{x \to -\infty}(-\infty, x)$, then the result follows by the continuity of probability measures.

(3) By the continuity (from above) of probability measures, we have

$$\lim_{x \to a^-} F(x) = \lim_{x \to a^-} \mathbb{P}(\{X \leq x\}) = \mathbb{P}\Big(\bigcap_{x \to a^-} \{X \leq x\}\Big) = \mathbb{P}(\{X \leq a\}) = F(a)$$

(4) By the continuity (from below) of probability measures, we have

$$\lim_{x \to a^+} F(x) = \lim_{x \to a^+} \mathbb{P}(\{X \leq x\}) = \mathbb{P}\Big(\bigcup_{x \to a^+} \{X \leq x\}\Big) = \mathbb{P}(\{X < a\}) = F(a)$$

(5) Follows by taking the difference between properties (3) and (4). □

Lecture 8 10/01

Theorem 24: Suppose a function $F : \mathbb{R} \to \mathbb{R}$ satisfies (1), (2), and (3) of the properties in proposition 23, above. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on that space such that F is the distribution function of X. That is, the properties (1), (2), and (3) fully characterize distribution functions.

Proof. Take $\Omega = (0,1)$, $\mathcal{F} = \mathscr{B}$ and $\mathbb{P}$ to be the Lebesgue measure λ. In particular, for $\omega \in (0,1)$ define $X(\omega) := \sup\{y \in \mathbb{R} : F(y) < \omega\} =: F^{-1}(\omega)$, called the *generalized inverse* of F. Note for the generalized inverse it is not the case that $F \circ F^{-1}$ is the identity. Then we want to show that the events $A = \{\omega \in (0,1) : \omega \leq F(x)\}$ and $B = \{x \in \mathbb{R} : X(\omega) \leq x\}$ are the same.

Let $\omega \in (0,1)$ and $x \in \mathbb{R}$. If $\omega \leq F(X)$ so that $\omega \in A$, then we know that for any y such that $F(y) < \omega$ (since F is non-decreasing), then $x > y$. Then taking the limit, we have $x \geq \sup\{y \in \mathbb{R} : F(y) < \omega)\}$ and so $X(\omega) \leq x$, and so $x \in B$.

On the other hand, if $\omega > F(x)$ so that $\omega \in A^c$, then since F is right-continuous (and the inequality is strict), we can move slightly to the right of x to get $x' > x$ still satisfying $\omega > F(x')$. Then $x' \in \{x \in \mathbb{R} : F(x) < \omega\}$ and so $x < x' \leq \sup\{x \in \mathbb{R} : F(x) < \omega\} = X(\omega)$.

We see then that $\omega \leq F(x)$ if and only if $X(\omega) \leq x$, so the events are identical. So, $\mathbb{P}(X(\omega) \leq x) = \mathbb{P}(\omega \leq F(x)) = \lambda((0, F(x))) = F(x)$, so the distribution function of X is $F(x)$, as desired. □

Notation: If X and Y induce the same distribution on $(\mathbb{R}, \mathscr{B})$, equivalently, if X and Y have the same distribution functions $F_X = F_Y$, then we say that X and Y *have the same distribution* or that they are *equal in distribution*. We denote this by $X \overset{\text{D}}{=} Y$. Note that this does not necessarily imply that $X = Y$, or equivalently that $X(\omega) = Y(\omega)$ for all ω or that they are equal almost surely (i.e., for all but a set of measure 0).

Definition. Probability Density/Mass Function: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. If there exists a function f such that for any $x \in \mathbb{R}$ we have $\mathbb{P}(X \leq x) = F(x) = \int_{-\infty}^{x} f(y)dy$, then f is called the (probability) density function of X (often abbreviated p.d.f.). If f is only defined on a countable set $A \in \mathscr{B}$, then f is called the (probability) mass function of X (often abbreviated p.m.f.).

Remark: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with density function f. Note that by the Fundamental Theorem of Calculus, if F is well-behaved, then $\frac{d}{dx}F(x) = f(x)$. Moreover,

$$\mathbb{P}(X \in (a,b])) = F(b) - F(a) = \int_{-\infty}^{b} f(y)dy - \int_{\infty}^{a} f(y)dy = \int_{a}^{b} f(y)dy$$

and as a result,

$$\mathbb{P}(X = x) = \lim_{\varepsilon \to 0} \mathbb{P}(X \in (x - \varepsilon, x + \varepsilon)) = \lim_{\varepsilon \to 0} \int_{x-\varepsilon}^{x+\varepsilon} f(y)dy = 0.$$

Note this also implies then that $\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b)$.

Definition. Continuous Distribution: If a distribution has a density function, then it is called an *absolutely continuous* distribution. If a distribution has $\mathbb{P}(X = a) = F(a) - \lim_{x \to a^-} F(x) = 0$ for all a, i.e., its distribution function is continuous, then the distribution is called *continuous* This also implies that all absolutely continuous distributions are continuous, though the converse is not necessarily true.

Definition. Discrete Distribution: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. If there is a countable set $A \in \mathscr{B}$ such that $\mathbb{P}(X \in A) = 1$, then X is said to be discrete. Note if X is not continuous, then it is discrete and has a probability mass function.

Definition. Singular Distribution: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with density function f. If there is a set $A \in \mathscr{B}$ such that $\lambda(A) = 0$ for the Lebesgue measure λ, but $\mathbb{P}(X \in A) = 1$ and X is continuous, then X is said to be singular.

Remark: The existence of singular distributions is why we cannot assume that a general distribution is a mixture of a part with density and a part with probability mass, as there may be a singular part. However, any distribution can be decomposed into an absolutely continuous part, a discrete part, and a singular part.

Example: Suppose X has a p.d.f. $f(x) = 1$ for $x \in (0, 1)$. Then X has c.d.f.

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

In this case, we say that X is a uniform $(0, 1)$ distribution and write $X \sim U(0, 1)$

Example: Suppose X has a p.d.f. $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $\lambda > 0$). Then we say X is an exponential λ distribution and write $X \sim Exp(\lambda)$.

Example: Suppose X has a p.d.f. $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for all $x \in \mathbb{R}$. Then we say X is a standard normal distribution and write $X \sim \mathcal{N}(0, 1)$.

Example: Suppose X has point mass at 0, i.e., $\mathbb{P}(X = 0) = 1$ so that $F(x) = \mathbb{1}_{x \geq 0}$. Then we say X has a Dirac distribution.

Definition. σ-field Generated by a Mapping: Let $X : \Omega \to \mathbb{R}$ be a mapping. Consider the class of sets $\{\{\omega \in \Omega : X(\omega) in \mathscr{B}\} : B \in \mathscr{B}\}$. This class is a σ-field, and in fact $\{\{\omega \in \Omega : X(\omega) \in B\} : B \in \mathscr{B}\}$ is the smallest σ-field on Ω such that X is a measurable mapping. We call it the σ-field generated by X and denote it by $\sigma(X)$. Similarly, $\sigma(X_1, X_2, \ldots)$ is the smallest σ-field such that all of $X_1, X_2, \ldots$ are measurable.

Exercise: Prove $\sigma(X) = \{\{\omega \in \Omega : X(\omega) \in B\} : B \in \mathscr{B}\}$ is a σ-field and the smallest σ-field such that X is a measurable mapping.

Theorem 25: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $X : (\Omega, \mathcal{F}) \to (S, \mathscr{A})$ and $f : (S, \mathscr{A}) \to (T, \mathscr{B})$ are measurable mappings, then $f \circ X : (\Omega, \mathcal{F}) \to (T, \mathscr{B})$ (defined by $(f \circ X)(\omega) = f(X(\omega))$) is a measurable mapping.

Proof. Let $B \in \mathscr{B}$. Since f is a measurable mapping and $B \in \mathscr{B}$, then $f^{-1}(B) \in \mathscr{A}$. Since X is a measurable mapping and $f^{-1}(B) \in A$, then $X^{-1}(f^{-1}(B)) \in \mathcal{F}$. Then note that

$$\{\omega : f(X(\omega)) \in B\} = \{\omega : X(\omega \in \underbrace{f^{-1}(B)}_{\in \mathscr{A}}\} \in \mathcal{F}$$

so $f \circ X$ is measurable. As a result, a measurable function of random variable(s) is a random variable. □

Note: Let $X_1, X_2, \ldots, X_n$ be random variables. Consider $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ be a function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathscr{B}^n)$ and $f : (\mathbb{R}^n, \mathscr{B}^n) \to (\mathbb{R}, \mathscr{B})$. Then $f \circ X$ is a random variable. Then, for instance, $-X_1$, $X_1 + \cdots + X_n$, $X_1 \cdots X_n$, e^{X_1}, $\sin(X_1 + X_2)$, etc. are all random variables.

Theorem 26: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $X_1, X_2, \ldots$ are random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then

$$\inf\{X_n : n \in \mathbb{N}\}, \quad \sup\{X_n : n \in \mathbb{N}\}, \quad \liminf_{n\to\infty} X_n, \quad \limsup_{n\to\infty} X_n$$

are all random variables (on the extended reals $\overline{\mathbb{R}}$).

Proof. Note that $\inf X_n < a$ if and only if $X_n < a$ for some n. Note that $\{X_n < a\} \in \mathcal{F}$ for all n since each X_n is a random variable. Hence, $\{\inf X_n < a\} = \bigcup_{n=1}^{\infty}\{X_n < a\} \in \mathcal{F}$ Similarly, $\{\sup X_n > a\} = \bigcup\{X_n > a\} \in \mathcal{F}$.

Now $\liminf_{n\to\infty} X_n = \sup_n \inf_{m\geq n} X_m$, however, $\inf_{m\geq n} X_m$ is a random variable, and thus $\sup_n \inf_{m\geq n} X_m = \liminf_{n\to\infty} X_n$ is a random variable. Similarly, $\limsup_{n\to\infty} X_n$ is also a random variable. □

Chapter 3 Lebesgue Integration

Lecture 9 10/03

Definition. σ-finite Measure: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space (μ is not necessarily a probability measure, i.e., it may not have total mass 1). Then μ is σ-finite if there are $A_1, A_2, \ldots \in \mathcal{F}$ such that $\mu(A_n) < \infty$ for all $n = 1, 2, \ldots$ and $\bigcup_n A_n = \Omega$.

Definition. Simple Function: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. A function $\varphi : \Omega \to \mathbb{R}$ is a simple function if $\varphi(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{\omega\in A_i}$ where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$ with $\mu(A_i) < \infty$ for all

$i = 1, \ldots, n$. Simple functions are in a sense, a generalization of the bins used in Riemann integrals.

Definition. Lebesgue Integral of Simple Function: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $\varphi(\omega) = \sum_{i=1}^n a_i \mathbb{1}_{\omega \in A_i}$ be a simple function. We define the Lebesgue integral of φ to be

$$\int \varphi d\mu = \sum_{i=1}^n a_i \mu(A_i)$$

Lemma 27: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let φ and ψ be simple functions. Then the following holds

(1) $\varphi \geq 0$ almost everywhere (a.e.) (i.e., $\mu(\varphi < 0) = 0$), then $\int \varphi d\mu \geq 0$.

(2) For any $a \in \mathbb{R}$, then $\int a\varphi d\mu = a \int \varphi d\mu$.

(3) $\int \varphi + \psi d\mu = \int \varphi d\mu + \int \psi d\mu$.

(4) If $\varphi \leq \psi$ almost everywhere, then $\int \varphi d\mu \leq \int \psi d\mu$.

(5) If $\varphi = \psi$ almost everywhere, then $\int \varphi d\mu = \int \psi d\mu$

(6) $|\int \varphi d\mu| \leq \int |\varphi| d\mu$.

Proof. (1) Trivial.

(2) Trivial.

(3) Suppose $\varphi = \int_i^n a_i \mathbb{1}_{A_i}$ and $\psi = \sum_j^m b_j \mathbb{1}_{B_j}$. Note for $\omega \in A_i \cap B_j$ then $(\varphi + \psi)(\omega) = a_i + b_j$. So we may write $\varphi + \psi = \sum_{i=1}^n \sum_{j=1}^m (a_i + b_j) \mathbb{1}_{A_i \cap B_j}$. Then

$$\begin{aligned}
\int \varphi + \psi d\mu &= \sum_{i=1}^n \sum_{j=1}^m (a_i + b_j) \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n \sum_{j=1}^m a_i \mu(A_i \cap B_j) + \sum_{i=1}^n \sum_{j=1}^m b_j \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n a_i \sum_{j=1}^m \mu(A_i \cap B_j) + \sum_{j=1}^m b_j \sum_{i=1}^n \mu(A_i \cap B_j) \\
&= \sum_{i=1}^n a_i \mu(A_i) + \sum_{j=1}^m b_j \mu(B_j) \\
&= \int \varphi d\mu + \int \psi d\mu,
\end{aligned}$$

(4) By (3), we may write $\int \psi d\mu = \int \varphi d\mu + \int (\psi - \varphi) d\mu$. But $(\psi - \varphi) \geq 0$ almost everywhere, and so by (1) we have that

$$\int \psi d\mu = \int \varphi d\mu + \int (\psi - \varphi) d\mu \geq \int \varphi d\mu + 0$$

(5) Since $\varphi = \psi$ almost everywhere, necessarily both $\varphi \leq \psi$ almost everywhere and $\psi \leq \varphi$ almost everywhere. Then by (4) we have equality of their integrals.

(6) Note that $|\varphi| = \max(\varphi, -\varphi)$. Then necessarily $\varphi \leq |\varphi|$ and so $\int \varphi d\mu \leq \int |\varphi| d\mu$ by (4). Similarly, $-\varphi \leq |\varphi|$ and so $-\int \varphi d\mu = \int -\varphi d\mu \leq \int |\varphi| d\mu$. This shows that $|\int \varphi d\mu| \leq \int |\varphi| d\mu$. □

Proposition 28: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be a bounded function such that $f(x) = 0$ for $x \in E^c$ for some E with $\mu(E) < \infty$. Then

$$\sup\left\{\int \varphi d\mu : \varphi \leq f \text{ and } \varphi \text{ satisfies } (*)\right\} = \sup\left\{\int \psi d\mu : \psi \geq f \text{ and } \psi \text{ satisfies } (*)\right\}$$

where $(*)$ is the condition that φ (resp. ψ) is simple and $\varphi(x) = 0$ (resp. $\psi(x) = 0$) for all $x \in E^c$.

Proof. Then for any such φ and ψ we have $\varphi \leq f \leq \psi$ and they are simple so that $\int \varphi d\mu \leq \int \psi d\mu$. Thus, we also see $\sup_{\varphi \leq f} \int \varphi d\mu \leq \inf_{\psi \geq f} \int \psi d\mu$.

On the other hand, as f is bounded, there is some M such that $|f(\omega)| \leq M$ for all $\omega \in \Omega$. Let $n \in \{1, 2, \ldots\}$ be given. Define $E_k = \{x \in E : \frac{kM}{n} \geq f(x) \geq \frac{(k-1)M}{n}\}$ for $-n \leq k \leq n$. Now define

$$\psi_n(x) = \sum_{k=-n}^{n} \frac{kM}{n} \mathbb{1}_{x \in E_k} \quad \text{and} \quad \varphi_n(x) = \sum_{k=-n}^{n} \frac{(k-1)M}{n} \mathbb{1}_{x \in E_k}$$

We also know then that $\psi_n(x) - \varphi_n(x) = \frac{M}{n} \mathbb{1}_{x \in E}$. We see then $\int (\varphi_n - \psi_n) d\mu = \frac{M}{n} \mu(E)$. Therefore, we see that

$$\sup_{\varphi \leq f} \int \varphi d\mu \geq \int \varphi_n d\mu = \int \psi_n d\mu - \int \psi_n - \varphi_n d\mu = \int \psi_n d\mu - \frac{M}{n}\mu(E) \geq \inf_{\psi \geq f} \int \psi d\mu - \frac{M}{n}\mu(E)$$

since φ_n (resp. ψ_n) is only one such $\varphi \leq f$ (resp. $\psi \geq f$) in the supremum (resp. infimum). Therefore, taking $n \to \infty$ we see that

$$\sup_{\varphi \leq f} \int \varphi d\mu = \lim_{n \to \infty} \sup_{\varphi \leq f} \int \varphi d\mu \geq \lim_{n \to \infty} \inf_{\psi \geq f} \int \psi d\mu - \frac{M}{n}\mu(E) = \inf_{\psi \geq f} \int \psi d\mu$$

We have thus shown both sides of the inequality, and so we see that

$$\sup_{\varphi \leq f} \int \varphi d\mu = \inf_{\psi \geq f} \int \psi d\mu.$$

□

Definition. Lebesgue Integral of a Bounded Function: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let f be a bounded function such that $f(x) = 0$ for $x \in E^c$ for some E with $\mu(E) < \infty$. Then we define the integral of f to be $\int f d\mu = \sup_{\varphi \leq f} \int \varphi d\mu = \inf_{\psi \geq f} \int \psi d\mu$.

Lemma 29: The properties (1)–(6) of integrals for simple functions in lemma 27 also hold for bounded functions.

Proof. Exercise. □

Definition. Lebesgue Integral of Non-negative Function: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be a non-negative function. Then we define the integral of f to be

$$\int f d\mu = \sup\left\{\int h d\mu : 0 \leq h \leq f,\ h \text{ is bounded, and } \mu(\{x : h(x) > 0\}) < \infty\right\}.$$

Lemma 30: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be a non-negative function. Define $(f \wedge n) := \min(f, n)$. Define $h_n = (f \wedge n)\mathbb{1}_{E_n}$ where $E_1 \subseteq E_2 \subseteq \cdots$ and $\lim_{n\to\infty} E_n = \Omega$ but $\mu(E_n) < \infty$ for $n = 1, 2, \ldots$. Then $\lim_{n\to\infty} \int h_n d\mu = \int f d\mu$.

Proof. Clearly, $\int h_n d\mu$ is non-decreasing (each h_n itself gets larger and larger, and the area of integration gets larger and larger). Thus the limit $\lim_{n\to\infty} \int h_n d\mu$ exists. Let

$$H := \{h : \Omega \to \mathbb{R} \ : \ 0 \leq h \leq f,\ h \text{ is bounded, and } \mu(\{x : h(x) > 0\}) < \infty\}$$

so that $\int f d\mu = \sup\{\int h d\mu : h \in H\}$. For any $h \in H$, let M be an upper bound of h. Then for any $n \geq M$, we have that

$$\int h_n d\mu = \int_{E_n} f \wedge n d\mu \geq \int_{E_n} h d\mu = \int h d\mu - \int_{E_n^c} h d\mu$$

since $h \leq f$ and h is bounded by M so that $h \leq \min\{f, M\} = f \wedge M \leq f \wedge n$. Let $E = \{x : h(x) > 0\}$ with $\mu(E) < \infty$. Then

$$\int_{E_n^c} h d\mu = \int_{E_n^c \cap E} h d\mu \leq \int_{E_n^c \cap E} M d\mu = M\mu(E_n^c \cap E) = M\mu(E \setminus E_n)$$

But, $E_n \to \Omega$ so that $\mu(E \setminus E_n) \to 0$. This implies then that $\int_{E_n^c} h d\mu \to 0$. So, taking $n \to \infty$, we have $\lim_{n\to\infty} \int h_n d\mu \geq \int h d\mu$. Since this is true for any $h \in H$, we have that

$$\lim_{n\to\infty} \int h_n d\mu \geq \sup\left\{\int h d\mu : h \in H\right\} = \int f d\mu.$$

On the other hand, $h_n \in H$ for all n, and so necessarily

$$\lim_{n\to\infty} \int h_n d\mu \leq \sup\left\{\int h d\mu : h \in H\right\} = \int f d\mu.$$

Proving the equality, as desired. □

Note: This proof does not necessarily mean that $\lim_{n\to\infty} \int h_m n d\mu = \int f d\mu$ is finite. In fact, both sides can simultaneously be infinite.

Lemma 31: The properties (1)–(6) of integrals for simple functions in lemma 27 also hold for non-negative functions.

Proof. Exercise. □

Definition. Lebesgue Integral: Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Let $f : \Omega \to \mathbb{R}$ be any measurable function. If $\int |f| d\mu < \infty$ (which is defined since $|f|$ is non-negative), then we say f is *integrable* and define $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$, where $f^+(x) = \max\{f(x), 0\}$ and $f^-(x) = -\min\{f(x), 0\} = \max\{-f(x), 0\}$ are the positive and negative parts of f, respectively. Each of f^+ and f^- is non-negative, however, and thus have defined integrals. Note that $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

Theorem 32: The properties (1)–(6) of integrals for simple functions in lemma 27 hold for all integrable functions.

Proof. Exercise. □

TWO PROBABILITY INEQUALITIES, PART 5 OF LECTURE 8

Theorem 33. Chebyshev's Inequality (General): Let X be an random variable and g be a non-negative function. For some $B \in \mathscr{B}$, let $l := \inf\{g(x) : x \in B\}$ (a lower bound of $g(x)$ on B). Then $l \cdot \mathbb{P}(X \in B) \leq \mathbb{E}[g(X)]$ or equivalently $\mathbb{P}(X \in B) \leq \frac{1}{l}\mathbb{E}[g(X)]$.

Proof. Define $Y = l \cdot \mathbb{1}_{X \in B}$. Then

$$\mathbb{P}(Y = y) = \begin{cases} \mathbb{P}(X \in B) & \text{if } y = l \\ 1 - \mathbb{P}(X \in B) & \text{if } y = 0 \end{cases}$$

Then, clearly $Y \leq g(x)$ for any $x \in B$ and so we see that

$$l \cdot P(X \in B) = \mathbb{E}[Y] \leq \mathbb{E}[g(X)]$$

as desired. □

Corollary 34. Markov Inequality: A result of the above theorem is that if X is a non-negative random variable, then $\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}[X]$ for any $a \geq 0$.

Proof. Take $B = [0, \infty) \in \mathscr{B}$ and $g(x) = x^+ = \max\{x, 0\}$. Then the result holds by Chebyshev's inequality. □

Corollary 35. Chebyshev's Inequality (Strict): A result of the above theorem is that

$$a^2\mathbb{P}(|X| \geq a) \leq \mathbb{E}[X^2] \quad \text{and} \quad a^2\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \text{Var}(X)$$

Proof. Take $g(x) = x^2$ and $B = (-\infty, -a] \cup [a, infty)$. Then the result holds by Chebyshev's inequality in the general case. □

Chapter 4 Convergence of Random Variables

Lecture 10 10/22

Note: When $\limsup_{n\to\infty} X_n = \liminf_{n\to\infty} X_n$, then $\lim_{n\to\infty} X_n$ exists.

$$\Omega_0 := \{\omega : \lim_{n\to\infty} X_n \text{ exists}\} = \underbrace{\{\limsup_{n\to\infty} X_n - \liminf_{n\to\infty} X_n = 0\}}_{\text{Random Variable}}$$

is measurable. If $\mathbb{P}(\Omega_0) = 1$, then we say that X_n converges almost surely (a.s.)

Example. Binomial Distribution: Let $X \sim Bin(n,p)$ be a random variable with distribution defined by $\mathbb{P}(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$. This is also the distribution of the number of successes in n independent trials, each having a probability of success p. As a result, $X \overset{\text{D}}{=} Y_1 + \cdots + Y_n$ where $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Bern(p)$. We see, moreover, that $\mathbb{E}[X] = \mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_n] = n\mathbb{E}[Y_1] = np$. Since each $Y_1, \ldots, Y_n$ is independent, variance is also additive, so $\text{Var}(X) = \text{Var}(Y_1) + \cdots = \text{Var}(Y_n) = n\,\text{Var}(Y_1) = np(1-p)$.

Definition. Characteristic Function: Let X be a random variable. Its characteristic function (ch.f.) is defined as

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)] = \int_{\mathbb{R}} e^{itx} f(x)dx = \int_{\mathbb{R}} \cos(tx)f(x)dx + i\int_{\mathbb{R}} \sin(tx)f(x)dx.$$

Note: The characteristic function of X is also the same as the Fourier transform of the probability density function of X.

Note: Unlike the moment generating function $\mathbb{E}[e^{tX}]$ or the generating function of non-negative integer-valued random variables $\mathbb{E}[s^X]$, the characteristic function always exists on $\mathbb{R}$. So we, don't need to worry about checking its existence. This is why, to prove general results about random variables, we should always use the characteristic function instead of the (moment) generating function.

Theorem 36. Jensen's Inequality: Let X be a random variable and φ be a convex function. Then $\varphi(\mathbb{E}[X]) \le \mathbb{E}[\varphi(X)]$.

Theorem 37. Properties of Characteristic Function: Let $\varphi = \varphi_X$ be the characteristic function of a random variable X. Then we have,

(1) $\varphi(0) = 1$.

(2) $\varphi(-t) = \overline{\varphi(t)}$ (i.e., the complex conjugate $\overline{a+bi} = a - bi$).

(3) $|\varphi(t)| = |\mathbb{E}[e^{itX}]| \le \mathbb{E}[|e^{itX}|] = 1$.

(4) $\varphi_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = e^{itb}\mathbb{E}[e^{itaX}] = e^{itb}\varphi_X(at)$.

Proof. (1) Trivially, $e^{0\cdot iX} = 1$.

(2) Trivially, $\varphi(-t) = \mathbb{E}[e^{-itX}] = \mathbb{E}[\cos(-tX)] + i\mathbb{E}[\sin(-tX)] = \mathbb{E}[\cos(tX)] - i\mathbb{E}[\sin(tX)]$ since cos is even and sin is odd.

(3) Note that $f(x,y) = (x^2+y^2)^{1/2}$ is a convex function. Thus, by Jensen's inequality $f(\mathbb{E}[X], \mathbb{E}[Y]) \leq \mathbb{E}[f(X,Y)]$. Define $X' = \cos(tX)$ and $Y' = \sin(tX)$. Then

$$|\varphi(t)| = |\mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)]| = f(\mathbb{E}[X'], \mathbb{E}[Y']) \leq \mathbb{E}[f(X',Y')] = \mathbb{E}[|\varphi(t)|] = 1.$$

(4) Follows trivially.

□

Theorem 38: Let X and Y be independent random variables with characteristic functions φ_X and φ_Y, respectively. Then $X+Y$ has characteristic function $\varphi_{X+Y} = \varphi_X \cdot \varphi_Y$.

Proof. Notice

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}] = \varphi_X(t)\varphi_Y(t)$$

since $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$ when X and Y are independent. □

Example. Poisson Distribution: Let $X \sim Poi(\lambda)$ be a random variable with distribution defined by $\mathbb{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, \ldots$. Then

$$\varphi(t) = \mathbb{E}[e^{itX}] = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k e^{itk}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{it})^k}{k!} = e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

since $\sum_{k=0}^{\infty}\frac{x^k}{k!} = e^x$ by the Taylor series expansion of e^x. Now let $Y \sim Poi(\eta)$ be independent of X. Then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{\lambda(e^{it}-1)}e^{\eta(e^{it}-1)} = e^{(\lambda+\eta)(e^{it}-1)}$$

We see that this is the characteristic function of a $Poi(\lambda+\eta)$. We may wonder then if this guarantees $X+Y \sim Poi(\lambda+\eta)$? More generally, does the characteristic function fully characterize/determine the distribution? Yes!

Theorem 39. Inversion Formula: Let $\varphi_X(t) = \int e^{itx}\mu(dx) = \int e^{itx}d\mu = \mathbb{E}[e^{itX}]$ where μ is the probability measure induced by X. Then for $a < b$,

$$\lim_{T\to\infty}(2\pi)^{-1}\underbrace{\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt}_{:=I_T} = \mu((a,b)) + \tfrac{1}{2}\mu(\{a,b\})$$

Proof. Define I_T as above. Namely,

$$\begin{aligned}
I_T &= \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t)\, dt \\
&= \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \int e^{itx}\, \mu(dx)\, dt \\
&= \int \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} e^{itx}\, dt\, \mu(dx) && \text{by Fubini's} \\
&= \int \int_{-T}^{T} \frac{e^{it(x-a)} - e^{it(x-b)}}{it}\, dt\, \mu(dx) \\
&= \int \left(- \int_{-T}^{T} \frac{\sin(t(x-a))}{t}\, dt + \int_{-T}^{T} \frac{\sin(t(x-b))}{t} dt \right) \mu(dx) && (*)
\end{aligned}$$

Note equation (*) follows since

$$\frac{e^{it(x-a)}}{i} = \frac{-\cos(t(x-a)) + i\sin(t(x-a))}{i} = -\sin(t(x-a)) + i\cos(t(x-a))$$

but, since cos is an even function and our integration interval $[-T, T]$ is symmetric, the cos cancels itself out. Now, for any $\theta > 0$, note that

$$\int_{-T}^{T} \frac{\sin(\theta t)}{t} dt = \int_{-T}^{T} \frac{\sin(\theta t)}{\theta t} d(\theta t) = \int_{-\theta T}^{\theta T} \frac{\sin(y)}{y} dy$$

by a u-substitution with $y = \theta t$. Now taking the limit

$$\lim_{T\to\infty} \int_{-T}^{T} \frac{\sin(\theta t)}{t} dt = \lim_{T\to\infty} \int_{-\theta T}^{\theta T} \frac{\sin y}{y} dy = \int_{-\infty}^{\infty} \frac{\sin y}{y} dy = \pi$$

Similarly, for $\theta < 0$ we get that

$$\lim_{T\to\infty} \int_{-T}^{T} \frac{\sin(\theta t)}{t} dt = \lim_{T\to\infty} \int_{-\theta T}^{\theta T} \frac{\sin y}{y} dy = \int_{-\infty}^{\infty} \frac{\sin y}{y} dy = -\pi$$

Therefore, applying this above result, we see that

$$g(x) := \lim_{T\to\infty} \int_{-T}^{T} \frac{\sin(t(x-a))}{t}\, dt - \int_{-T}^{T} \frac{\sin(t(x-b))}{t} dt = \begin{cases} 2\pi & \text{if } a < x < b \\ \pi & \text{if } x = a \text{ or } x = b \\ 0 & \text{if } x < a \text{ or } x > b \end{cases}$$

Note that

$$\int_{-T}^{T} \frac{\sin(t(x-a))}{t}\, dt \le \sup_c \int_{-c}^{c} \frac{\sin(y)}{y} dy =: M < \infty$$

This implies

$$\left| \int_{-T}^{T} \frac{\sin(t(x-a))}{t}\, dt - \int_{-T}^{T} \frac{\sin(t(x-b))}{t} dt \right| \le 2M$$

Then by the dominated convergence theorem,

$$\lim_{T\to\infty} I_T = \int g(x)\mu(dx) = 2\pi\mu((a,b)) + \pi\mu(\{a,b\})$$

Hence, $\frac{1}{2\pi} I_T \to \mu((a,b)) + \frac{1}{2}\mu(\{a,b\})$. □

Lecture 11 10/29

Remark: From the inversion formula, we know that if $X_1, X_2, \ldots, X_n$ are Poisson random variables with $X_i \sim Poi(\lambda_i)$ then $\sum_{i=1}^n X_i \sim Poi(\sum_{i=1}^n \lambda_i)$. It's worth noting that we rarely apply the inversion formula directly, rather it's primary use is in ensuring that characteristic functions fully characterize their distributions.

Example. Normal Distribution: A standard normal distribution $X \sim \mathcal{N}(0,1)$ has probability density function $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for $x \in \mathbb{R}$. Then the characteristic function is

$$\begin{aligned}
\varphi(t) &= \mathbb{E}[e^{itX}] \\
&= \int_{\mathbb{C}} e^{itx} f(x) dx \\
&= \int_{\mathbb{C}} \frac{1}{\sqrt{2\pi}} e^{itx - x^2/2} dx \\
&= e^{-t^2/2} \int_{\mathbb{C}} \frac{1}{\sqrt{2\pi}} e^{(x-it)^2/2} dx \\
&= e^{-t^2/2} \int_{\mathbb{C}} \underbrace{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}_{=f(y)} dy \qquad\qquad y = x - it \\
&= e^{-t^2/2}
\end{aligned}$$

Note, however, that we simplified this problem by performing a u-sub, ignoring the fact that it is an integral over the complex plane. However, doing so is valid since we are taking the analytic continuation of $\mathbb{E}[e^{tX}] = e^{-1/2t^2}$ to the complex plane.

Example. General Normal Distribution: A general normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ has probability density function $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$. Let $Y = \mu + \sigma Z$ for $Z \sim \mathcal{N}(0,1)$. Then, we can easily show $Y \overset{\mathrm{D}}{=} X$ by property (4) of characteristic functions. In particular,

$$\varphi_Y(t) = e^{i\mu t - \sigma^2 t^2/2}$$

Notice also that if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with X_1 and X_2 being independent, then

$$\varphi_{X_1+X_2}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t) = \exp(i\mu_1 t - \tfrac{1}{2}\sigma_1^2 t^2) \exp(i\mu_2 t - \tfrac{1}{2}\sigma_2^2 t^2) = \exp(i(\mu_1+\mu_2)t - \tfrac{1}{2}(\sigma_1^2+\sigma_2^2)t^2)$$

We notice this is the characteristic function of a $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ distribution. Thus, by the inversion formula, we see that $Y_1 + Y_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Remark. Multivaraite Characteristic Functions: Suppose $\vec{X} = (X_1, \ldots, X_n)$ is a multivariate random variable. Then for any $\vec{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n$ the characteristic function of $\vec{X}$ is defined to be

$$\varphi_{\vec{X}}(\vec{T}) = \mathbb{E}[e^{i\langle \vec{t}, \vec{X}\rangle}] = \mathbb{E}[\exp(i \sum_{j=1}^n (t_i X_i))]$$

where $\langle a, b\rangle$ is the inner-product of a and b. In $\mathbb{R}^n$ this inner product is usually the dot product $a \cdot b = \sum_{i=1}^n a_i b_i$.

Example. Multivariate Normal: An $\mathbb{R}^n$-valued multivariate random variable $\vec{X} = (X_1, \ldots, X_n)$ is said to be normal if any linear combination of the coordinates $X_1, \ldots, X_n$ follows a (univariate) normal distribution. Note that this univariate normal distribution may or may not be degenerate, which is to say it may have variance 0, and therefore take on a single constant value almost surely. Recall any such linear combination would be of the form $\sum_{i=1}^n a_i X_i \in \mathbb{R}$ where $a_1, \ldots, a_n \in \mathbb{R}$.

We see then that $\vec{X}$ is a $\mathbb{R}^n$-valued normal random variable if and only if its characteristic function has the form

$$\varphi_{\vec{X}}(\vec{t}) = \exp(i\langle \vec{t}, \vec{\mu}\rangle - \tfrac{1}{2}\langle \vec{t}, \Sigma, \vec{t}\rangle) \qquad (*)$$

where $\vec{\mu} \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}^{n\times n}$ is a symmetric positive semi-definite matrix. Recall a matrix $A \in \mathbb{R}^{n\times n}$ is said to be positive semi-definite if and only if $xAx^T \geq 0$ for all $x \in \mathbb{R}^n$. In this case, $\vec{\mu}$ is the mean of $\vec{X}$:

$$\vec{\mu} = (\mu_1, \ldots, \mu_n) = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]) = \mathbb{E}[\vec{X}]$$

and Σ is the (variance-)covariance matrix of X:

$$\Sigma_{ij} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

where $\sigma_{ij}^2 = \mathrm{Cov}(X_i, X_j)$.

Proof. ($\Longleftarrow$) Assume equation (*) holds. Then for $Y = \sum_{i=1}^n A_i X_i = \langle \vec{a}, \vec{X}\rangle$, its characteristic function is given by

$$\begin{aligned} \varphi_Y(t) &= \mathbb{E}[e^{itY}] & &= \mathbb{E}[e^{i(ta_1X_1 + \cdots + ta_nX_n)}] \\ &= \mathbb{E}[e^{i\langle t\vec{a}, \vec{X}\rangle}] \\ &= \varphi_{\vec{X}}(t\vec{a}) \\ &= \exp(i\langle t\vec{a}, \vec{\mu}\rangle - \tfrac{1}{2}\langle t\vec{a}, \Sigma t\vec{a}\rangle) \\ &= \exp(it\langle \vec{a}, \vec{\mu}\rangle - \tfrac{t^2}{2}\langle \vec{a}, \Sigma \vec{a}^T\rangle) \end{aligned}$$

We remark, however, that this is the characteristic function of a univariate normal distribution $\mathcal{N}(\langle \vec{a}, \vec{\mu}\rangle, \langle \vec{a}, \Sigma\vec{a}\rangle)$. Note that $\langle \vec{a}\Sigma\vec{a}\rangle = \vec{a}^T\Sigma\vec{a}$.

($\Longrightarrow$) Suppose $\vec{X}$ is a multivariate normal. Then for any $\vec{a} \in \mathbb{R}^n$, consider $Y = \langle \vec{a}, \vec{X}\rangle = a_1X_1 + \cdots + a_nX_n \in \mathbb{R}$. Then Y is a univariate normal distribution, and in particular, from an n-dimensional generalization of the sum of normals seen in the previous example, we will have $Y \sim \mathcal{N}(\langle \vec{a}, \vec{\mu}\rangle, \langle \vec{a}, \Sigma\vec{a}\rangle)$. From this, we know that the characteristic function of Y will be

$$\varphi_Y(t) = \exp(it\langle \vec{a}, \vec{\mu}\rangle - \tfrac{t^2}{2}\langle \vec{a}, \Sigma\vec{a}\rangle)$$

notice also that

$$\varphi_{\vec{X}}(\vec{a}) = \mathbb{E}[\exp(i\langle\vec{a}, \vec{X}\rangle)] = \mathbb{E}[\exp(iY)] = \mathbb{E}[\exp(i \cdot 1 \cdot Y)] = \varphi_Y(1) = \exp(i\langle\vec{a}, \vec{\mu}\rangle - \tfrac{1}{2}\langle\vec{a}, \Sigma\vec{a}\rangle)$$

We see, however, that this is exactly (*) (if we take $\vec{a} = \vec{t}$ for whichever $\vec{t} \in \mathbb{R}^n$ for which we'd like to evaluate $\varphi_{\vec{t}}(\vec{t})$). □

Remark: We can see that the distribution of a multivariate normal random variable is completely determined by its mean vector and covariance matrix (its first two moments).

Corollary 40: If $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ then $A\vec{X} + \vec{c} \sim \mathcal{N}(A\vec{\mu} + \vec{c}, A\Sigma A^T)$ for any $A \in \mathbb{R}^{m\times n}$ and $\vec{c} \in \mathbb{R}^m$. Note $\vec{X}$ is an $\mathbb{R}^n$-valued multivariate normal random variable and $A\vec{X}$ is an $\mathbb{R}^m$-value multivariate normal random variable.

Corollary 41: If $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, then any two components X_i, X_j are independent if and only if $\mathrm{Cov}(X_i, X_j) = 0$.

Proof. It is easiest to prove the reverse direction by analyzing their characteristic functions. □

Remark: In the general case, while $X \perp\!\!\!\perp Y$ implies $\mathrm{Cov}(X, Y) = 0$, it is not necessarily the case that $\mathrm{Cov}(X, Y)$ implies $X \perp\!\!\!\perp Y$.

Lecture 12 10/31

Definition. Almost Sure Convergence: Let $X_1, X_2, \ldots$ be a sequence of random variables. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ converges almost surely to a random variable X if and only if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$. This effectively checks that the realizations of each random variable are the same. This can be alternatively written as $\mathbb{P}(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$. Note you can also check that the complement has probability measure 0. In this case, we write $\{X_n \to X\}$ a.s.or $X_n \overset{\text{a.s.}}{\to} X$.

Definition. Convergence Everywhere: Let $X_1, X_2, \ldots$ be a sequence of random variables. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ converges everywhere to a random variable X if and only if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$. Note this differs from converging almost surely since $X_n(\omega) \to X(\omega)$ has to hold for all $\omega \in \Omega$, not only on a set of measure 1.

Definition. Convergence in $\mathcal{L}^p$: Defined for $1 \leq p < \infty$. Let $X_1, X_2, \ldots$ be a sequence of random variables. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ converges everywhere to a random variable X if and only if $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$. Recall that $\mathcal{L}^P$ is the normed vector space with norm $\|x\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ for vectors, $\|f\|_p = (\int |f|^p d\mu)^{1/p}$ for functions, and $\|X\|_p = \mathbb{E}[|X|^p]^{1/p}$ for random variables. Thus, $X_n \overset{\mathcal{L}^p}{\to} X$ if and only if $\|X_n - X\|_p \to 0$. A useful case is that if $X_n \overset{\mathcal{L}^1}{\to} X$ then $\mathbb{E}[|X_n - X|] \to 0$ which in turn implies that $\mathbb{E}[X_n] \to \mathbb{E}[X]$, this is necessary but not sufficient to imply convergence. Similarly $X_n \overset{\mathcal{L}^2}{\to} X$ if and only if $\mathbb{E}[(X_n - X)^2] \to 0$ which implies $\mathbb{E}[X_n^2] \to \mathbb{E}[X^2]$ which is again necessary but not sufficient to imply convergence. Note that $\mathcal{L}^2$ is an especially nice space to analyze convergence in since

it is a Hilbert space (has an inner product structure). Other $\mathcal{L}^p$ spaces are Banach spaces, but not necessarily Hilbert spaces, and so they do not necessarily have inner products.

Exercise: Prove that $X_n \xrightarrow{\mathcal{L}^1} X$ implies that $\mathbb{E}[X_n] \to \mathbb{E}[X]$. Give a counterexample where $\mathbb{E}[X_n] \to \mathbb{E}[X]$ but $X_n \not\xrightarrow{\mathcal{L}^1} X$.

Definition. Convergence in Probability: Let $X_1, X_2, \dots$ be a sequence of random variables. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ converges in probability to a random variable X if and only if $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$ for any $\varepsilon > 0$. This is equivalent to saying $\lim_{n\to\infty} \mathbb{P}(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0$ for all $\varepsilon > 0$. In this case we write $X_n \xrightarrow{\mathbb{P}} X$

Theorem 42:

(1) If $X_n \xrightarrow{\text{a.s.}} X$, then $X_n \xrightarrow{\mathbb{P}} X$.

(2) If $X_n \xrightarrow{\mathcal{L}^p} X$, then $X_n \xrightarrow{\mathbb{P}} X$.

Proof. (1) Fix $\varepsilon > 0$. Let $A_n = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \leq \varepsilon \text{ for all } m \geq n\}$. Then A_n is non-decreasing (i.e., $A_1 \supseteq A_2 \supseteq A_3$). Notice then that

$$\begin{aligned}\lim_{n\to\infty} A_n &= \bigcup_{n=1}^{\infty} A_n \\ &= \{\text{exists } n \text{ such that } |X_m - X| \leq \varepsilon \text{ for all } m \geq n\} \\ &\supseteq \{X_n \to X\} = \{\omega \in \Omega : X_n(\omega) \to X(\omega)\}\end{aligned}$$

Then, since $P(\{X_n \to X\}) = 1$, we see that

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\lim_{n\to\infty} A_n) = \mathbb{P}(\{X_n \to X\}) = 1$$

by the continuity of convergence. On the other hand, $A_n \subseteq \{|X_n - X| \leq \varepsilon\}$ which shows that $1 \leq \lim_{n\to\infty} \mathbb{P}(A_n) \leq \lim_{n\to\infty} \mathbb{P}(|X_n - X| \leq \varepsilon) \leq 1$.

(2) Fix $\varepsilon > 0$. Then

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p) \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p}$$

by the Markov Inequality. We know, however, that $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$ since $X_n \xrightarrow{\mathcal{L}^p} X$ and so

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} = \frac{0}{\varepsilon^p} = 0$$

□

Proposition 43: Let $1 \leq p < q < \infty$. If $X_n \xrightarrow{\mathcal{L}^q} X$ then also $X_n \xrightarrow{\mathcal{L}^p} X$.

Proof. For any $\varepsilon < 1$,

$$\mathbb{E}[|X_n - X|^p] = \mathbb{E}[|X_n - X|^p \mathbb{1}_{\{|X_n - X| \geq \varepsilon\}}] + \mathbb{E}[|X_n - X|^p \mathbb{1}_{\{|X_n - X| < \varepsilon\}}]$$

Note that whenever $|X_n - X| \geq \varepsilon$, then $|X_n - X|^p \leq \varepsilon^{p-q}|X_n - X|^q$ since $p - q < 0$ and x^{p-q} is therefore a decreasing function. We also clearly see that whenever $|X_n - X| < \varepsilon$ then $|X_n - X|^p < \varepsilon^p$. So, we have that

$$\begin{aligned}\mathbb{E}[|X_n - X|^p] &= \mathbb{E}[|X_n - X|^p \mathbb{1}_{\{|X_n - X| \geq \varepsilon\}}] + \mathbb{E}[|X_n - X|^p \mathbb{1}_{\{|X_n - X| < \varepsilon\}}] \\ &\leq \varepsilon^{p-q}\mathbb{E}[|X_n - X|^q \mathbb{1}_{\{|X_n - X| \geq \varepsilon\}}] + \varepsilon^p \\ &\leq \varepsilon^{p-q}\mathbb{E}[|X_n - X|^q] + \varepsilon^p\end{aligned}$$

This implies

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] \leq \varepsilon^{p-q} \limsup_{n\to\infty} \mathbb{E}[|X_n - X|^q] + \varepsilon^p = \varepsilon^p$$

Taking $\varepsilon \to 0$ completes the proof. □

Example: In general, $X_n \xrightarrow{\mathbb{P}} X$ does not imply $X_n \xrightarrow{\text{a.s.}} X$ or $X_n \xrightarrow{\mathcal{L}^p} X$. As a counterexample, take the probability space $(\Omega, \mathcal{F}, \mu) = ((0,1), \mathscr{B}, \lambda)$ where λ is the Lebesgue measure.

Consider taking $X_1 = \mathbb{1}_{0\leq\omega\leq 1}$, then $X_2 = \mathbb{1}_{0\leq\omega\leq\frac{1}{2}}$ and $X_3 = \mathbb{1}_{\frac{1}{2}\leq\omega\leq 1}$, then $X_4 = \mathbb{1}_{0\leq\omega\leq\frac{1}{3}}$ and $X_5 = \mathbb{1}_{\frac{1}{3}\leq\omega\leq\frac{2}{3}}$ and $X_6 = \mathbb{1}_{\frac{2}{3}\leq\omega\leq 1}$. I.e., partitioning the space into smaller and smaller splits on which X_n is 1. We see then that $X_n \xrightarrow{\mathbb{P}} 0$ but $X_n \overset{\text{a.s.}}{\not\to} 0$. To see $X_n \xrightarrow{\mathbb{P}} 0$, let ε be given, then the probabilities are $\mathbb{P}(|X_n| > \varepsilon) = 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots \to 0$ To see $X_n \overset{\text{a.s.}}{\not\to} 0$, notice that for all irrational numbers $\omega \in (0,1)$ there will always be some n such that $X_n(\omega) = 1$. Thus $\mathbb{P}(\{X_n \to 0\}) < 1$.

Consider now $X_n = n\mathbb{1}_{0<\omega<\frac{1}{n}}$. Similarly to above, $X_n \xrightarrow{\mathbb{P}} 0$, in particular for any $0 < \varepsilon < 1$ and any n, $\mathbb{P}(|X_n - 0| > \varepsilon) = \frac{1}{n} \to 0$. But $\mathbb{E}[|X_n - 0|^p] = \int n \cdot \frac{1}{n} d\lambda = 1$ for all n, and so $X_n \overset{\mathcal{L}^1}{\not\to} 0$, and as a result $X_n \overset{\mathcal{L}^p}{\not\to} 0$ for any $p \geq 1$.

Remark: The above examples also show that $X_n \xrightarrow{\text{a.s.}} X$ does not imply $X_n \xrightarrow{\mathcal{L}^1} X$ or vice-versa. In fact, the first example above had $X_n \xrightarrow{\mathcal{L}^1} X$ despite $X_n \overset{\text{a.s.}}{\not\to} X$ and the second example had $X \xrightarrow{\text{a.s.}} X$ despite $X_n \overset{\mathcal{L}^1}{\not\to} X$.

Theorem 44: If $X_n \xrightarrow{\mathbb{P}} X$, then there exists a subsequence $\{X_{i_m}\}_{m=1}^{\infty}$ for $i_1 < i_2 < i_3 < \cdots$ where $X_{i_m} \xrightarrow{\text{a.s.}} X$

Proof. Set $i_0 = 0$. For any $m \geq 1$, set $i_m = \inf\{i \geq i_{m-1} : \mathbb{P}(|X_i - X| > \frac{1}{m}) \leq 2^{-m}\}$. Notice this infimum is always exists since $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$ for any $\varepsilon > 0$. We see then that

$$\sum_{m=1}^{\infty} \mathbb{P}(|X_{i_m} - X| > \frac{1}{m}) \leq \sum_{m=1}^{\infty} 2^{-m} = 1 < \infty$$

Then, by the first Borel-Contelli Lemma,

$$\mathbb{P}(|X_{i_m} - X| > \frac{1}{m} \text{ i.o.}) = 0$$

this in turn implies that $|X_{i_m} - X| > \frac{1}{m}$ almost surely, only for finitely many m. Finally, this means that $|X_{i_m} - X| \leq \frac{1}{m}$ starting from some term $m \geq M$ almost surely. This therefore means that $X_{i_m} \overset{\text{a.s.}}{\to} X$. □

Definition. Uniform Integrability: Let $X_1, X_2, \ldots$ be a sequence of random variables. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ is uniformly integrable (denote U.I.) if

$$\lim_{k\to\infty} \sup_n \mathbb{E}[|X_n| \mathbb{1}_{\{|X_n|>k\}}] = 0$$

Intuitively, if you only had 1 function X in your sequence and you only look at the part $|X| > k$, then the area will eventually be 0. So uniform integrability implies that our integrals don't grow faster than linearly. This means that our random variables grow nicely and we don't always have huge integrals. For instance, in our counterexample with rectangles of shape $[0, \frac{1}{n}] \times [0, n]$, these grow too quickly for uniform integrability.

Lecture 13 11/05

Theorem 45: If $X_n \overset{\mathbb{P}}{\to} X$ and $\{X_n\}_{n=1}^{\infty}$ is uniformly integrable, then $X_n \overset{\mathcal{L}^1}{\to} X$.

Proof. Without loss of generality, assume $X \equiv 0$ (as we can just subtract both sides of the convergence by X). For any fixed $\varepsilon > 0$, since $\{x_n\}$ is uniformly integrable, there exists some $t_\varepsilon > 0$ such that $\mathbb{E}[|X_n| \mathbb{1}_{\{|X_n|>t_\varepsilon\}}] \leq \varepsilon$ for any n (by definition of the limit). Since $X_n \overset{\mathbb{P}}{\to} X = 0$, there exists an N_ε such that for any $n \geq N_\varepsilon$, we have $\mathbb{P}(|X_n| > \varepsilon) \leq \frac{\varepsilon}{t_\varepsilon}$. Hence, for any $n \geq N_\varepsilon$,

$$\begin{aligned}
\mathbb{E}[|X_n|] &= \int_{\{|X_n|<\varepsilon\}} |X_n| d\mathbb{P} + \int_{\{\varepsilon<|X_n|<t_\varepsilon\}} |X_n| d\mathbb{P} + \int_{\{|X_n|>t_\varepsilon\}} |X_n| d\mathbb{P} \\
&\leq \int_{\{|X_n|<\varepsilon\}} \varepsilon d\mathbb{P} + t_\varepsilon \mathbb{P}(|X| > \varepsilon) + \mathbb{E}[|X_n| \mathbb{1}_{\{|X_n|>t_\varepsilon\}}] \\
&\leq \varepsilon + \varepsilon \cdot \frac{\varepsilon}{t_\varepsilon} + \varepsilon \\
&= 3\varepsilon
\end{aligned}$$

Since this holds for all $\varepsilon > 0$ and $n \geq N_\varepsilon$, this shows $\lim_{n\to\infty} \mathbb{E}[|X_n|^1] = 0$. Thus, $X_n \overset{\mathcal{L}^1}{\to} X$ by definition. □

Remark: We have seen so far that almost sure convergence and convergence in $\mathcal{L}^p$ both imply convergence in probability. We also saw that convergence in probability implies the

existence of a subsequence that converges almost surely. Finally, we just saw that for uniformly integrable sequences, convergence in probability implies convergence in $\mathcal{L}^p$. We will not be able to find any relation between almost sure convergence and convergence in $\mathcal{L}^p$.

Definition. Convergence in Distribution: (Also called weak convergence.) Let $X_1, X_2, \ldots$ be a sequence of random variables with distribution functions $F_1, F_2, \ldots$, respectively. We say that the sequence $\{X_n\}_{n=1}^{\infty}$ converges in distribution to a random variable X with distribution function F if and only if $F_n \to F$ pointwise in all points where F is continuous. This is equivalent to saying that $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ such that F is continuous at x. In this case, we write $X_n \xrightarrow{d} X$.

Remark: Convergence in distribution is different from the previous three types of convergences in the sense that we don't care about the random variables themselves converging, only about their distribution functions. In particular, the previous convergences told us something about the pointwise behaviour of the random variables. Convergence in distribution only gives us information about the pointwise behaviour of the *distribution functions*. I.e., it is not about relation of $X_n(\omega)$ and $X(\omega)$, rather only about the distributions $F_n(x)$ and $F(x)$.

Example: If $X_1, X_2, \ldots, \overset{\text{i.i.d.}}{\sim} X$, then $X \xrightarrow{d} X$, but $X_1, X_2, \ldots$ does not have any other convergence in random variables.

Example: Consider independent trials which we repeat until we get our first success. If the probability of a success in a given trial is p, then the number of trials until a success, X_p, has a geometric distribution with parameter p. Note also that $\mathbb{P}(X_p = n) = p(1-p)^{n-1}$ and so this implies that $\mathbb{P}(X_p > n) = (1-p)^n$. Consider taking $p \to 0$. Then

$$\lim_{p\to 0} \mathbb{P}(p \cdot X_p > x) = \lim_{p\to 0} \mathbb{P}(X_p > \frac{x}{p}) \approx \lim_{p\to 0} (1-p)^{x/p} = e^{-x} \qquad \forall x \geq 0$$

(where the approximation comes from the fact that $\frac{x}{p}$ is not necessarily an integer). We see then that $\mathbb{P}(pX_p \leq x) \to 1 - e^{-x}$, which is the distribution function of an Exp(1) random variable. Thus, we see that $pX_p \xrightarrow{d} W$ where $W \sim \text{Exp}(1)$.

Proposition 46: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables. If $X_n \xrightarrow{\mathbb{P}} X$ for some random variable X then $X_n \xrightarrow{d} X$.

Proof. Let F_n be the distribution function of X_n. Let F be the distribution function of X. Let $a \in \mathbb{R}$ be any point such that F is continuous at a. For any $\varepsilon > 0$, since F is continuous at a and F is non-decreasing, there exists a $\delta > 0$ such that

$$F(a) - \varepsilon < F(a-\delta) \leq F(a) \leq F(a+\delta) < F(a) + \varepsilon$$

Now since $X_n \xrightarrow{\mathbb{P}} X$, there exists an $N_\varepsilon > 0$ such that $\mathbb{P}(|X_n - X| > \delta) < \varepsilon$ for all $n \geq N_\varepsilon$. Now for such an $n \geq N_\varepsilon$, we have that

$$F_n(a) = \mathbb{P}(X_n \leq a) = \mathbb{P}(X_n \leq a, |X_n - X| \leq \delta) + \mathbb{P}(X_n \leq a, |X_n - X| > \delta)$$

Note, however, that

$$0 \leq \mathbb{P}(X_n \leq a, |X_n - X| > \delta) \leq \mathbb{P}(|X_n - X| > \delta) < \varepsilon$$

Notice also that we have

$$\begin{aligned}\mathbb{P}(X_n \le a, |X_n - X| \le \delta) &\in \Big[\mathbb{P}(X \le a-\delta, |X_n - X| \le \delta),\ \mathbb{P}(X \le a+\delta, |X_n - X| \le \delta)\Big] \\ &\subseteq \Big[\mathbb{P}(X \le a-\delta) - \mathbb{P}(|X_n - X| > \delta),\ F(a+\delta)\Big] \\ &\subseteq \Big[F(a-\delta) - \varepsilon, F(a+\delta)\Big]\end{aligned}$$

Thus, this implies that $F_n(a) \in [F(a-\delta)-\varepsilon, F(a+\delta)+\varepsilon] \subseteq (F(a)-2\varepsilon, F(a)+2\varepsilon)$. Since this holds for all $\varepsilon > 0$, we conclude that $F_n(a) \to F(a)$ for all $a \in \mathbb{R}$ such that F is continuous at a. Thus, by definition $X_n \xrightarrow{d} X$. □

Proposition 47: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables (defined on the same probability space). If $X_n \xrightarrow{d} c$ for some constant c, then $X_n \xrightarrow{\mathbb{P}} c$.

Proof. For any $\varepsilon > 0$, we need to show that $\mathbb{P}(|X_n - c| \le \varepsilon)$ converges to 1. Let F_n denote the distribution function of X_n. Now notice that

$$\mathbb{P}(|X_n - c| \le \varepsilon) = \mathbb{P}(c-\varepsilon \le X_n \le c+\varepsilon) \ge \mathbb{P}(c-\varepsilon, X_n \le c+\varepsilon) = F_n(c+\varepsilon) - F_n(c-\varepsilon)$$

Since the distribution of $X \equiv c$ is $F(x) = \mathbb{1}_{\{x>c\}}$, we see that $F_n(c+\varepsilon) \to F(c+\varepsilon) = 1$ and $F_n(c-\varepsilon) \to F(c-\varepsilon) = 0$ (since F is continuous at $c-\varepsilon$ and $c+\varepsilon$ and by definition of convergence in distribution). Thus, this implies that

$$\liminf_{n\to\infty} \mathbb{P}(|X_n - c| \le \varepsilon) \ge \liminf(F_n(c+\varepsilon) - F_n(c-\varepsilon)) \ge 1 - 0 = 1$$

Since $\mathbb{P}(|X_n - c| \le \varepsilon) \to 1$ holds for all ε, then $X \xrightarrow{\mathbb{P}} c$. □

Remark: We note that convergence in distribution is the weakest form of convergence. In particular, every other form of convergence implies convergence in distribution. However, convergence in distribution implies convergence in probability only if X is constant.

Lecture 14 11/07

Theorem 48. Skorokhod's Theorem: Let $X_1, X_2, \ldots$ be a sequence of random variables such that $X_n \xrightarrow{d} X$ for a random variable X. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of a random variables $Y_1, Y_2, \ldots$ and a random variable Y such that $X_n \overset{\mathrm{D}}{=} Y_n$ and $X \overset{\mathrm{D}}{=} Y$ such that $Y_n \xrightarrow{\text{a.s.}} Y$.

Proof. Let F_n be the distribution functions of each X_n, respectively. Let F be the distribution function of X. Let the probability space be $((0,1), \mathscr{B}, \lambda)$ where λ is Lebesgue measure. Recall that we can define $Y_1, Y_2, \ldots$ and Y on this space such that each distribution function F_n has

$$Y_n(x) := F_n^{-1}(x) = \sup\{y : F_n(y) < x\}$$

This is the generalized inverse of F_n. From theorem 24, we had seen that Y_n defined as above has distribution function F_n. We can similarly define $Y(x) := F^{-1}(x)$. We will show that $Y_n \overset{\text{a.s.}}{\to} Y$.

For any $x \in (0,1)$, define

$$a_x := \sup\{y : F(y) < x\}\ (= F^{-1}(x)) \qquad \text{and} \qquad b_x := \inf\{y : F(y) > x\}$$

Also define $\Omega_0 := \{x : (a_x, b_x) = \emptyset\} = \{x : a_x = b_x\}$. Note that if $x \in \Omega_0$, then $y < F^{-1}(x)$ implies $F(y) < x$ and $y > F^{-1}(x)$ implies $F(y) > x$.

We claim for any $x \in \Omega_0$ that

(1) $\liminf_{n\to\infty} F_n^{-1}(x) \geq F^{-1}(x)$.

(2) $\limsup_{n\to\infty}^{-1}(x) \leq F^{-1}(x)$.

This will be sufficient to imply almost sure convergence since $\Omega \setminus \Omega_0$ is at most countable (which will in turn imply it has measure zero, i.e., $\mathbb{P}(\Omega \setminus \Omega_0) = 0$). To see that $\Omega \setminus \Omega_0$, note each $x \in \Omega \setminus \Omega_0$ corresponds to a disjoint interval. Each such interval contains a different rational number, so consider mapping rational numbers to the intervals containing them. Since there are countably many rational numbers, there are only countably many such intervals, and so $\Omega \setminus \Omega_0$ is countable.

Now we prove (1). Let $y < F^{-1}(x)$ be such that F is continuous at y. Since $x \in \Omega_0$, we know that $F(y) < x$. By the weak convergence of $X_n \overset{d}{\to} X$, we know that $F_n(y) \to F(y)$. Thus, $\lim_{n\to\infty} F_n(y) = F(y) < x$ implies $F_n(y) < x$ for sufficiently large n. This implies that $y \leq F_n^{-1}(x)$ and so that $y \leq \liminf_{n\to\infty} F_n^{-1}(x)$. Taking the limit $y \to F^{-1}(x)$ from below gives $F^{-1}(x) \leq \liminf_{n\to\infty} F_n^{-1}(x)$. (Note this limit is valid since there are also only countably many points of discontinuity in F. The proof follows by a similar argument to showing $\Omega \setminus \Omega_0$ is countable.)

We will not prove (2), as it holds by a similar argument. □

Definition. **μ-continuity Set:** A set A is said to be a μ-continuity set if $\mu(\partial(A)) = 0$ where $\partial(A)$ is the boundary of A (i.e., $\partial(A) = \overline{A} \setminus \text{Int}(A)$).

Theorem 49. Portmanteau Theorem: Let $X_1, X_2, \ldots$ be a sequence of random variables and be another random variable X. The following are equivalent:

(1) $X_n \overset{d}{\to} X$.

(2) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded and continuous function f.

(3) $\mu_n(A) \to \mu_(A)$ for every μ-continuity set A, where μ_n is the probability measure induced by X_n and μ is the probability measure induced by X. This is equivalent to saying $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$ for all sets A such that $\mathbb{P}(X \in \partial(A)) = 0$.

Proof. We will show (1) $\iff$ (2) and (1) $\iff$ (3).

(1) $\implies$ (2) Let f be any bounded and continuous function f. By Skorohod's theorem, there are $Y_1, Y_2, \dots$ and Y such that $Y_n \overset{\text{D}}{=} X_n$ and $Y \overset{\text{D}}{=} X$ where $Y_n \overset{\text{a.s.}}{\to} Y$. Note in this case $\mathbb{E}[f(Y_n)] = \mathbb{E}[f(X_n)]$ and $\mathbb{E}[f(Y)] = \mathbb{E}[f(X)]$. Since f is continuous and $Y_n \overset{\text{a.s.}}{\to} Y$, we know that $f(Y_n) \overset{\text{a.s.}}{\to} f(Y)$. Moreover, f is bounded and so we can use the dominated convergence theorem. This theorem in turn implies that $\mathbb{E}[f(Y_n)] \to \mathbb{E}[f(Y)]$.

(1) $\implies$ (3) A similar idea applies here, but we instead use the function $f = \mathbb{1}_{\{x \in A\}}$, which is not continuous at $\partial(A)$. We know, however, that $\mathbb{P}(Y \in \partial(A)) = \mathbb{P}(X \in \partial(A)) = 0$ since A is a μ-continuity set, and so f is almost surely continuous for all Y. Hence, we still have $f(Y_n) \overset{\text{a.s.}}{\to} f(Y)$. Thus, since f is clearly bounded, we again can apply dominated convergence theorem. This theorem in turn implies that

$$\mathbb{P}(X_n \in A) = \mathbb{P}(Y_n \in A) = \mathbb{E}[f(Y_n)] \to \mathbb{E}[f(Y)] = \mathbb{P}(Y \in A) = \mathbb{P}(X_n \in A).$$

(3) $\implies$ (1) Take $A = (-\infty, x]$ so that $\partial(A) = \{x\}$. Clearly A is a μ-continuity if and only if $\mathbb{P}(X = x) = 0$. This is true, however, if and only if F is continuous at X. So our statement is that $F_n(x) = \mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A) = F(x)$ for all μ-continuity sets implies, or equivalently that $F_n(x) \to F(x)$ for all continuity points of F.

(2) $\implies$ (1) For any x such that F is continuous at x and $y > x$, define f as

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

which is linear between x and y. We see then that f is both bounded and continuous. Notice also that

$$F_n(x) = \mathbb{P}(X_n \leq x) = \mathbb{E}[\mathbb{1}_{(-\infty, x]}(X_n)] \leq \mathbb{E}[f(X_n)]$$

and

$$\mathbb{E}[f(X)] \leq \mathbb{E}[\mathbb{1}_{(\infty, y)}(X)] = F(y)$$

So by (2), taking $n \to \infty$ we see that $\limsup_{n\to\infty} F_n(x) \leq F(y)$. Now since F is continuous at X, taking $y \to x$ from above we have $\limsup_{n\to\infty} F_n(x) \leq F(x)$.

Consider instead taking $y < x$ and defining a similar f. We can similarly get $F(x) \leq \liminf F_n(x)$ by taking $y \to x$ from below. We have seen then that

$$F(x) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(x)$$

which shows that $\lim_{n\to\infty} F_n(x) = F(x)$, as desired. □

Lecture 15 11/12

Theorem 50. Helly's Selection Theorem: Let $X_1, X_2, \dots$ be a sequence of random variables. Let F_n be the distribution functions of each X_n, respectively. Then there exists a subsequence $F_{n(1)}, F_{n(2)}, \dots$ where $n(1) < n(2) < \cdots$ of $F_1, F_2, \dots$ and a right continuous non-decreasing function F such that $\lim_{m\to\infty} F_{n(m)}(y) = F(y)$ for all y at which F is continuous.

Proof. Consider an enumeration $q_1, q_2, \ldots$ of the rational numbers. Thus, since $F_n(q_1), F_n(q_2), \ldots$ is a bounded sequence of real numbers, then by Heine-Borel theorem, there is a subsequence $F_{n_1(1)}, F_{n_1(2)}, \ldots$ which converges to G (for some G) at q_1. Now, take a further subsequence $F_{n_2(1)}(q_2), F_{n_2(2)}(q_2), \ldots$ (where $\{n_2(i)\}_{i=1}^{\infty} \subseteq \{n_1(i)\}_{i=1}^{\infty}$) which converges to G at q_1 and q_2. We continue this for each rational number q_k. Consider now a diagonal argument:

$$\begin{matrix} F_{n_1(1)} & F_{n_1(2)} & F_{n_1(3)} & \cdots \\ F_{n_2(1)} & F_{n_2(2)} & F_{n_2(3)} & \cdots \\ F_{n_3(1)} & F_{n_3(2)} & F_{n_3(3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{matrix}$$

Where row i converges to G at $q_1, q_2, \ldots, q_i$ and row j is a subsequence of row i for $i < j$. Consider taking the sequence $F_{n_1(1)}, F_{n_2(2)}, F_{n_3(3)}, \ldots$. Then this sequence is a subsequence of each row (since the tail of $n > i$ is a subsequence of row i). As a result, $F_{n_1(1)}, F_{n_2(2)}$ converges to G at each rational number. Since each F_i is non-decreasing, and G is formed as the limit of these non-decreasing functions, it is also non-decreasing. To make G right-continuous, consider defining $F(x) := \inf\{G(q) : q \in \mathbb{Q}, q > x\}$. F is still non-decreasing, but now also right-continuous by the construction with an infimum. Now we show that $F_{n_k(k)}(x) \to F(x)$ at all $x \in \mathbb{R}$ such that F is continuous at x. Let $\varepsilon > 0$ be given. Take rational numbers $r_1, r_2, s \in \mathbb{Q}$ such that $r_1 < r_2 < x < s$ and

$$F(x) - \varepsilon < F(r_1) \leq F(r_2) \leq F(x) \leq F(s) < F(x) + \varepsilon$$

Now since $F_{n_k(k)}(s) \to G(s)$ and $G(s) \leq F(s) < F(x) + \varepsilon$, (in particular taking the limit we have $G(s) \leq G(q)$ for all $q \in Q$ with $q > s$ and $G(s) \leq \inf\{G(q) : q \in \mathbb{Q}, q > s\} = F(s)$), there exists N_1 such that

$$F_{n_k(k)}(x) \leq F_{n_k(k)}(s) < F(x) + \varepsilon$$

for all $k \geq N_1$. Similarly, since $F_{n_k(k)}(r_2) \to G(r_2)$ we have $G(r_2) \leq F(r_2)$ (however this inequality isn't in the direction we'd like). Now we have that $F(r_1) = \inf\{G(q) : q \in \mathbb{Q}, q > r_1\} \leq G(r_2)$ since $r_2 > r_1$. But now we have that $G(r_2) \geq F(r_1) > F(x) - \varepsilon$. Thus, there exists an N_2 such that

$$F_{n_k(k)}(x) \geq F_{n_k(k)}(r_2) > F(x) - \varepsilon$$

for all $k \geq N_2$. Thus, setting $N = \max\{N_1, N_2\}$, we have that whenever $k \geq N$,

$$F(x) - \varepsilon < F_{n_k(k)}(r_2) \leq F_{n_k(k)}(x) \leq F_{n_k(k)}(s) < F(x) + \varepsilon$$

That is, for any $\varepsilon > 0$, there is an N such that $|F_{n_k(k)}(x) - F(x)| < \varepsilon$ whenever $k \geq N$. And so, $F_{n_k(k)} \to F$ for all $x \in \mathbb{R}$, as desired. □

Remark: Note that in the above theorem, the limit function F may not necessarily be a distribution function.

Definition. Tightness of Measures: A sequence of probability measures $\mu_1, \mu_2, \ldots$ on $(\mathbb{R}, \mathbf{b})$ is called tight if for any $\varepsilon > 0$, there is an $M > 0$ such that $\liminf_{n\to\infty}(\mu_n(-M, M]) \geq 1 - \varepsilon$. This is equivalent to $\mu_n((-M, M]) \geq 1 - \varepsilon$ for a large enough n. This is equivalent to $\mu_n((-M', M']) \geq 1 - \varepsilon$ for all n for some M' (different from M). This is equivalent to $\limsup_{n\to\infty} \mu(\mathbb{R} \setminus (-M, M]) < \varepsilon$. For distribution functions $F_1, F_2, \ldots$, we say that the

sequence is tight if and only if the sequence of probability measures $\mu_1, \mu_2, \ldots$ induced by the distribution functions is tight. Note that for distribution functions the last equivalence may be rewritten as $\limsup_{n\to\infty} \mu(1 - F(m) + F(-m)) < \varepsilon$. Intuitively, this means that the majority of the mass is not found in the tails of the distributions.

Theorem 51: The sequence of distribution functions $F_1, F_2, \ldots$ is tight if and only if each of its subsequences has a further subsequence that converges weakly (pointwise at points of continuity) to a probability measure.

Proof. ($\implies$) Suppose $\mu_1, \mu_2, \ldots$ is a tight sequence of probability measures. By Helly's selection theorem, we already know that every subsequence has a further subsequence $F_{n(1)}(y), F_{n(2)}(y)$ which converges to $F(y)$ for all points y at which F is continuous. It remains to show that F is in fact a distribution function.. In particular, the only missing properties are $\lim_{y\to-\infty} F(y) = 0$ and $\lim_{y\to\infty} F(y)$. Recall that in theorem 48 we had shown that a non-decreasing function has only countably many points of discontinuity. Thus, $F_{n(m)} \to F(y)$ almost everywhere. Since $F_{n(m)}$ are distribution functions, we know that $\lim_{y\to-\infty} F(y) \geq 0$ and $\lim_{y\to\infty} F(y) \leq 1$. Thus, it suffices to show that $\lim_{y\to\infty}(F(y) - F(-y)) = 1$. By the tightness of the sequence, we know that for $\varepsilon > 0$ there is an M_ε such that $\limsup_{n\to\infty}(1 - F_n(M_\varepsilon) + F_n(-M_\varepsilon)) \leq \varepsilon$. Then, taking $r < -M_\varepsilon$ and $s > M_\varepsilon$ that are continuity points of F, we have that

$$1 - F(s) + F(r) = \lim_{m\to\infty}(1 - F_{n(m)}(s) + F_{n(m)}(r)) \leq \limsup_{m\to\infty}(1 - F_{n(m)}(M_\varepsilon) + F_{n(m)}(-M_\varepsilon)) \leq \varepsilon$$

But then, we have that

$$\lim_{y\to\infty}(F(y) - F(-y)) \geq F(s) - F(r) \geq 1 - \varepsilon$$

However, since $\varepsilon > 0$ was arbitrary, we have that $1 \leq \lim_{y\to\infty}(F(y) - F(-y)) \leq 1$, as desired.

($\impliedby$) By way of contrapositive, suppose that the sequence of probability distributions $F_1, F_2, \ldots$ is not tight. Then there exists an $\varepsilon > 0$ and indices $n(m) \to \infty$ such that $1 - F_{n(k)}(k) + F_{n(k)}(-k) \geq \varepsilon$ for all $m = 1, 2, \ldots$. By Helly's selection theorem, there is a subsequence $F_{m(1)}, F_{m(2)}, \ldots$ (where $\{m(i)\}_{i=1}^\infty \subseteq \{n(i)\}_{i=1}^\infty$) and a non-decreasing right-continuous function F such that $F_{m(k)}(y) \to F(y)$ for all points y at which F is continuous. Then, for any $r < s$ such that r and s are points of continuity of F, then

$$1 - F(s) + F(r) = \lim_{k\to\infty}(1 - F_{m(k)}(s) + F_{m(k)}(r)) \geq \liminf_{k\to\infty}(1 - F_{m(k)}(k) + F_{m(k)}(-k)) \geq \varepsilon$$

Thus, taking $r \to -\infty$ and $s \to \infty$, we get that $\lim_{r\to-\infty, s\to\infty} F(s) - F(r) \leq 1 - \varepsilon$. We see then that F cannot be a valid distribution function. Then, any further subsequence of $F_{n(m)}$ cannot converge (weakly) to the distribution function either. □

Lecture 16 11/14

Corollary 52: If $\mu_1, \mu_2, \ldots$ is a tight sequence of probability measures, and each weakly converging subsequence converges to the same limit, μ, then $\mu_n \xrightarrow{d} \mu$.

Proof. By way of contradiction, suppose that $\mu_n \overset{d}{\not\to} \mu$. Let $F, F_1, F_2, \ldots$ be the distribution functions induced by $\mu, \mu_1, \mu_2, \ldots$ respectively. Then there is a point x such that F is continuous at x but $F_n(x) \not\to F(x)$. That is, there is an $\varepsilon > 0$ such that $|F_n(x) - F(x)| \geq \varepsilon$ for infinitely many n. Let $n_1, n_2, \ldots$ be the enumeration of all n_k such that $|F_{n_k}(x) - F(x)| \geq \varepsilon$. Consider the subsequence $\mu_{n_1}, \mu_{n_2}, \ldots$. By tightness, this subsequence has a further subsequence that converges weakly. This subsequence cannot, however, converge to μ, as the corresponding distribution functions by construction do not converge to $F(x)$. This is a contradiction by the assumption that each weakly convergent subsequence has the same limit. Thus, it must be that $\mu_n \overset{d}{\to} \mu$. □

Proposition 53: Let φ be the characteristic function of a probability measure μ. Then φ is continuous on $\mathbb{R}$.

Proof. Recall, $\varphi(t) = \mathbb{E}[e^{itX}]$. Note that as $s \to t$ we have $e^{isX} \to e^{itX}$ pointwise. Moreover, since $|e^{itX}| = 1$, then by the dominated convergence theorem $\varphi(s) \to \varphi(t)$ as $s \to t$. This shows that φ is continuous on $\mathbb{R}$. □

Theorem 54. Continuity Theorem: Let $\mu_1, \mu_2, \ldots$ and μ be probability measures on $(\mathbb{R}, \mathscr{B})$. Let $\varphi, \varphi_1, \varphi_2, \ldots$ be the chracteristic functions for $\mu, \mu_1, \mu_2, \ldots$ respectively. Then $\mu_n \overset{d}{\to} \mu$ if and only if $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$.

Proof. ($\implies$) Note that e^{itX} is a continuous, bounded function of X. Therefore, by the portmanteau theorem (note that while e^{itX} is complex not real, we can apply portmanteau theorem to the real and imaginary parts separately) we have that $\mathbb{E}[e^{itX_n}] \to \mathbb{E}[e^{itX}]$ where $X_n \sim \mu_n$ and $X \sim \mu$. That is, $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$.

($\impliedby$) We will first show that $\mu_1, \mu_2, \ldots$ is tight. Note

$$\begin{aligned}
\frac{1}{u}\int_{-u}^{u} \underbrace{(1 - \varphi_n(t))}_{\mathbb{E}[1-e^{itX}]} dt &= \mathbb{E}\left[\tfrac{1}{u}\int_{-u}^{u}(1 - e^{itX_n})dt\right] && \text{hard} \\
&= 2\mathbb{E}\left[1 - \frac{\sin(uX_n)}{uX_n}\right] && \text{by hyperbolic sine} \\
&\geq 2\mathbb{E}\left[\left(1 - \frac{1}{|uX_n|}\right)\cdot \mathbb{1}_{\{|X_n| \geq \frac{2}{u}\}}\right] && \text{since } -1 \leq \frac{\sin x}{x} \leq 1 \\
&\geq 2\mathbb{E}\left[\left(1 - \tfrac{1}{2}\right)\cdot \mathbb{1}_{\{|X_n| \geq \frac{2}{u}\}}\right] && \text{since } |X_n| \geq \frac{2}{u} \\
\frac{1}{u}\int_{-u}^{u}(1 - \varphi_n(t))dt &\geq \mathbb{P}(|X_n| \geq \tfrac{2}{u}) && (*)
\end{aligned}$$

Note that the above expectation (and in particular getting to sin) is quite difficult to show, and there are complete proofs available elsewhere. Now since φ is continuous by the previous proposition, and $\varphi(0) = 1$, we know that

$$\lim_{u \to 0} u^{-1}\int_{-u}^{u}(1 - \varphi(t))dt = 0$$

We can view this as the average around 0, which since $\varphi(t)$ is continuous and $\varphi(0) = 1$, this is 0. Thus, for any $\varepsilon > 0$, there is a u such that

$$u^{-1} \int_{-u}^{u} (1 - \varphi(t))dt < \varepsilon$$

Now since $\varphi_n \to \varphi$ pointwise, by dominated convergence theorem, we have that

$$u^{-1} \int_{-u}^{u} (1 - \varphi_n(t))dt \to u^{-1} \int_{-u}^{u} (1 - \varphi(t))dt$$

Since this is a pointwise converges, there is an N such that whenever $n \geq N$, then

$$u^{-1} \int_{-u}^{u} (1 - \varphi_n(t))dt < \varepsilon$$

Set $M = \frac{2}{u}$ in equation (*) to get

$$\mathbb{P}(|X_n| \geq M) \leq \frac{1}{u} \int_{-u}^{u} (1 - \varphi_n(t))dt < \varepsilon$$

or equivalently, $P(X_n \in (-M, M)) > 1 - \varepsilon$ for any $n \geq N$. This is one of our equivalent definition for tightness, thus we see that $\mu_1, \mu_2, \ldots$ is tight.

Now assume that there is a subsequence $\mu_{n_1}, \mu_{n_2}, \ldots$ converges weakly to a probability measure μ', i.e., $\mu_{n_k} \xrightarrow{d} \mu'$. By the other direction of this theorem, we know that $\varphi_{n_k}(t) \to \varphi'(t)$ for all $t \in \mathbb{R}$ where φ' is the characteristic function of μ'. However, we know that $\varphi_{n_k}(t) \to \varphi(t)$, and so we must have that $\varphi' = \varphi$. Since the characteristic function completely determines the distribution, we must have that $\mu = \mu'$. Hence, all weakly convergent subsequences converge to the same limit μ. By the previous corollary, we therefore also have that $\mu_n \xrightarrow{d} \mu$. □

Chapter 5 Big Theorems in Probability

Theorem 55: Let $X_1, X_2, \ldots$ be a sequence of uncorrelated random variables (note this is strictly weaker than independence) with $\mathbb{E}[X_i] = \mu$ and such that $\text{Var}(X_i) \leq c < \infty$ for all $i = 1, 2, \ldots$. Let $S_n = X_1 + \cdots + X_n$. Then $\frac{1}{n}S_n \xrightarrow{\mathcal{L}^2} \mu$ (and as a result $\frac{1}{n}S_n \xrightarrow{\mathbb{P}} \mu$).

Proof. Note that $\mathbb{E}[\frac{S_n}{n}] = \mu$. Hence, consider the second moment

$$\mathbb{E}\left[\left(\frac{S_n}{n} - \mu\right)^2\right] = \text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}(\text{Var}(X_1) + \cdots + \text{Var}(X_n)) \leq \frac{nc}{n^2} \to 0$$

Thus, we see that the first two moments of $\frac{S_n}{n}$ converge to μ, or $\frac{S_n}{n} \xrightarrow{\mathcal{L}^2} \mu$ (which is a strictly stronger result than $\frac{1}{n}S_n \xrightarrow{\mathbb{P}} \mu$). □

Remark: If $X_1, X_2, \ldots$ are i.i.d. with finite variance (or second moment), then $\frac{1}{n}S_n \xrightarrow{\mathcal{L}^2} \mu$.

Theorem 56: Consider a triangular array of random variables

$$\begin{array}{cccc} X_{1,1} & & & \\ X_{2,1} & X_{2,2} & & \\ X_{3,1} & X_{3,2} & X_{3,3} & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

where $X_{n,1}, \ldots, X_{n,n}$, the variables in the same row, are all independent. Define $S_n = \sum_{i=1}^n X_{n,i}$ to be the sum of random variables in row n. Let $\mu_n = \mathbb{E}[S_n]$ and $\sigma_n^2 = \mathrm{Var}(S_n)$. If $\frac{\sigma_n^2}{b_n^2} \to 0$, then $\frac{S_n - \mu_n}{b_n} \xrightarrow{\mathcal{L}^2} 0$ (and as result $\frac{S_n-\mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$).

Proof. Note that

$$\mathbb{E}\left[\left(\frac{S_n - \mu_n}{b_n}\right)\right] = \frac{1}{b_n^2}\underbrace{\mathbb{E}[(S_n - \mu_n)^2]}_{\mathrm{Var}(S_n)} = \frac{\sigma_n^2}{b_n^2} \to 0$$

Thus $\frac{S_n-\mu_n}{b_n} \xrightarrow{\mathcal{L}^2} 0$ (which is a strictly stronger result showing $\frac{S_n-\mu_n}{b_n} \xrightarrow{\mathbb{P}} 0$). □

Remark: This extends our previous result, since now we can have multiple separate sequences that show convergence. Previously, we had to have one contiguous sequence which is convergent, whereas now we have increasingly large but separate sequences.

Remark: If $X_{n,k}$ are all identically distributed with mean μ and variance σ^2 and the random variables in the same row are independent (i.e., $X_{n,1}, \ldots, X_{n,n} \overset{\text{i.i.d.}}{\sim} X$), then $\bar{X}_n := \frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$ (by taking $b_n = n$). This is the most common formulation of the law of large numbers.

Lecture 17 11/19

Theorem 57: Consider a triangular array of random variables

$$\begin{array}{cccc} X_{1,1} & & & \\ X_{2,1} & X_{2,2} & & \\ X_{3,1} & X_{3,2} & X_{3,3} & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

where $X_{n,1}, \ldots, X_{n,n}$, the variables in the same row, are all independent. Let $b_1, b_2, \ldots$ be a sequence of real numbers such that each $b_n > 0$ and $b_n \to \infty$. Now consider truncating the random variables by $\overline{X_{n,k}} := X_{n,k}\mathbb{1}_{\{|X_{n,k}| \le b_n\}}$. Suppose that as $n \to \infty$ then

(1) $\sum_{k=1}^n \mathbb{P}(|X_{n,k}| > b_n) \to 0$

(2) $b_n^{-2}\sum_{k=1}^n \mathbb{E}[\overline{X_{n,k}}^2] \to 0$

Now further define $S_n = X_{n,1} + \cdots X_{n,n}$ to be the row sum for the nth row and define $a_n = \sum_{k=1}^n \mathbb{E}[\overline{X_{n,k}}]$. Then $\frac{S_n - a_n}{b_n} \xrightarrow{\mathbb{P}} 0$.

Proof. Define $\overline{S_n} = \overline{X_{n,1}} + \cdots + \overline{X_{n,n}}$. Let $\varepsilon > 0$. Then

$$\mathbb{P}\Bigg(\underbrace{\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon}_{\overline{S_n} \neq S_n \text{ or } |\frac{\overline{S_n} - a_n}{b_n}| > \varepsilon} \Bigg) \leq \mathbb{P}(S_n \neq \overline{S_n}) + \mathbb{P}\left(\left|\frac{\overline{S_n} - a_n}{b_n}\right| > \varepsilon\right)$$

Now note that

$$\begin{aligned}
\mathbb{P}(S_n \neq \overline{S_n}) &\leq \mathbb{P}(\exists k = 1, \ldots, n \; \overline{X_{n,k}} \neq X_{n,k}) \\
&\leq \sum_{k=1}^{n} \mathbb{P}(\overline{X_{n,k}} \neq X_{n,k}) \\
&= \sum_{k=1}^{n} \mathbb{P}(|X_{n,k}| > b_n) \to 0
\end{aligned}$$

by condition (1). Note that $a_n = \mathbb{E}[\overline{S_n}]$, and so

$$\mathbb{E}\left[\left(\frac{\overline{S_n} - a_n}{b_n}\right)^2\right] = b_n^{-2} \operatorname{Var}(\overline{S_n}) = b_n^{-2} \sum_{k=1}^{n} \operatorname{Var}(\overline{X_{n,k}}) \leq b_n^{-2} \sum_{k=1}^{n} \mathbb{E}[\overline{X_{n,k}}^2] \to 0$$

by condition (2). Thus, we have that

$$\frac{\overline{S_n} - a_n}{b_n} \xrightarrow{\mathcal{L}^2} 0 \qquad \text{and so} \qquad \frac{\overline{S_n} - a_n}{b_n} \xrightarrow{\mathbb{P}} 0$$

This gives

$$\mathbb{P}\left(\left|\frac{\overline{S_n} - a_n}{b_n}\right| > \varepsilon\right) \to 0 \qquad \text{and so} \qquad \mathbb{P}\left(\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right) \to 0,$$

as desired. □

Lemma 58: Let $Y \geq 0$ be a non-negative random variable. Then

$$\mathbb{E}[Y^2] = \int_0^\infty 2y\mathbb{P}(Y > y)dy$$

Proof. Note that $\mathbb{P}(Y > y) = \mathbb{E}[\mathbb{1}_{\{Y>y\}}]$. Thus,

$$\begin{aligned}
\int_0^\infty 2y\mathbb{P}(Y > y)dy &= \int_0^\infty \mathbb{E}[2y\mathbb{1}_{\{Y>y\}}]dy \\
&= \mathbb{E}\left[\int_0^\infty 2y\mathbb{1}_{\{Y>y\}}dy\right] && \text{by Fubini's} \\
&= \mathbb{E}\left[\int_0^Y 2y\, dy\right] \\
&= \mathbb{E}[Y^2]
\end{aligned}$$

□

Theorem 59: Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables with $x\mathbb{P}(|X_1| > x) \to 0$ as $x \to \infty$. Define $S_n = X_1 + \cdots + X_n$ and $\mu_n \mathbb{E}[X_1 \mathbb{1}_{\{|X_1| \le n\}})$. Then $\frac{S_n}{n} - \mu_n \xrightarrow{\mathbb{P}} 0$

Proof. Take $X_{n,k} = X_k$ and $b_n = n$. Then $\sum_{k=1}^n \mathbb{P}(|X_{n,k}| > b_n) = n \cdot \mathbb{P}(|X_1| > n) \to 0$, satisfying condition (1) of the previous theorem. Now notice that

$$b_n^{-2} \sum_{k=1}^n \mathbb{E}[\overline{X_{n,k}}^2] = n^{-2} \cdot n \cdot \mathbb{E}[\overline{X_{n,1}}^2] = \frac{1}{n}\mathbb{E}[\overline{X_{n,1}}^2]$$

From the above lemma, we have that

$$\mathbb{E}[\overline{X_{n,1}}^2] = \int_0^\infty 2y\mathbb{P}(|\overline{X_{n,1}}| > y)dy \le \int_0^n 2y\mathbb{P}(|X_1| > y)dy$$

since whenever $|\overline{X_{n,1}}| > y$, then also $|X_{n,1}| > y$ and whenever $y > n$, this is trivially false (since $\overline{X_{n,1}}$ is truncated). Now since $y\mathbb{P}(|X_1| > y) \to 0$, we have that

$$\frac{1}{n}\int_0^n 2y\mathbb{P}(|X_1| > y)dy \to 0$$

(since the running average converges to the limit). This implies that $b_n^{-2} \sum_{k=1}^n \mathbb{E}[\overline{X_{n,k}}^2] \to 0$, thereby satisfying condition (2) of the previous theorem. Now by the previous theorem,

$$\frac{S_n - a_n}{b_n} \xrightarrow{\mathbb{P}} 0 \quad \iff \quad \frac{S_n - n\mu_n}{n} = \frac{S_n}{n} - \mu_n \xrightarrow{\mathbb{P}} 0$$

□

Remark: Now we'll look at a weak law of large numbers that does not require finite second moments. Note also that if $p \le q$ and $\mathbb{E}[|X|^q] < \infty$ and $\mathbb{E}[|X|^p] < \infty$. In particular, if $0 < x \le 1$ then $x^q \le x^p \le 1$ and if $x > 1$ then $x^p \le x^q$, which shows the result. Thus having finite means is a weaker condition than having finite second moments.

Theorem 60. Weak Law of Large Numbers: Let $X_1, X_2, \dots$ be an i.i.d. sequence of random variables with $\mathbb{E}[|X_1|] < \infty$. Let $S_n = X_1 + \cdots + X_n$ and $\mu = \mathbb{E}[X_1]$. Then $\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$.

Proof.

$$\begin{aligned} x\mathbb{P}(|X_1| > x) &= \mathbb{E}[x\mathbb{1}_{\{|X_1|>x\}}] \\ &\le \mathbb{E}[|X_1|\mathbb{1}_{\{|X_1|>x\}}] \to 0 \end{aligned}$$

To see this, note that by the dominated convergence theorem

$$\mathbb{E}[|X_1|\mathbb{1}_{\{|X_1|>x\}} \to 0$$

since $|X_1|\mathbb{1}_{\{|X_1|>x\}} \le |X_1|$ and $|X_1|\mathbb{1}_{\{|X_1|>x\}} \to 0$ as $x \to \infty$. Note also that by the dominated convergence theorem,

$$\mu_n = \mathbb{E}[X_1\mathbb{1}_{\{|X_1|\le n\}}] \to \mathbb{E}[X_1] = \mu$$

since $|X_1 \mathbb{1}_{\{|X_1| \leq n\}}| \leq |X_1|$ and $X_1 \mathbb{1}_{\{|X_1| \leq n\}} \to X_1$ as $n \to \infty$.. Thus, we see that

$$\frac{S_n}{n} - \mu = \left(\frac{S_n}{n} - \mu_n \right) + (\mu_n - \mu) \xrightarrow{\mathbb{P}} 0$$

which in turn shows that $\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu$. To see this, we would need to show that if $X_n \xrightarrow{\mathbb{P}} X$ and $Y_n \xrightarrow{\mathbb{P}} Y$ then $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$, though this is straightforward. □

Remark: Note that many of the previous theorems are forms of weak laws of large numbers, but the above theorem is the most classical statement.

Lemma 61: If $y \geq 0$, then

$$2y \sum_{k \in \mathbb{Z} \,:\, k > y} k^{-2} \leq 4$$

Proof. For $y \geq 1$, note that

$$\sum_{k \in \mathbb{Z} \,:\, k > y} k^{-2} = \sum_{k = \lfloor y \rfloor + 1}^{\infty} \frac{1}{k^2} \leq \sum_{k = \lfloor y \rfloor + 1}^{\infty} \left(\frac{1}{k-1} - \frac{1}{k} \right) = \frac{1}{\lfloor y \rfloor}$$

Since $y \geq 1$, then $\frac{y}{\lfloor y \rfloor} \leq 2$ and so $2y \sum_{k \in \mathbb{Z} \,:\, k > y} k^{-2} \leq 4$. If instead $0 \leq y < 1$, then

$$\sum_{k \in \mathbb{Z} \,:\, k > y} k^{-2} = 1 + \underbrace{\sum_{k=2}^{\infty} k^{-2}}_{\leq 1} \leq 2$$

and so $2y \sum_{k \in \mathbb{Z} \,:\, k > y} k^{-2} \leq 4$. □

Lemma 62: Let $X_1, X_2, \ldots$ be a sequence of random variables have identical distributions and pairwise independent. Define $Y_k = X_k \mathbb{1}_{\{|X_k| \leq k\}}$. Then

$$\sum_{k=1}^{\infty} \frac{\operatorname{Var}(Y_k)}{k^2} \leq 4\mathbb{E}[|X_1|] < \infty$$

Proof. Note

$$\operatorname{Var}(Y_k) = \mathbb{E}[Y_k^2] - \mathbb{E}[Y_k]^2 \leq \mathbb{E}[Y_k^2] = \int_0^{\infty} 2y\mathbb{P}(|Y_k| > y)dy \leq \int_0^{k} 2y\mathbb{P}(|X_1| > y)dy$$

Thus, we see that

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\text{Var}(Y_k)}{k^2} &\leq \sum_{k=1}^{\infty} \frac{\mathbb{E}[Y_k^2]}{k^2} \\
&\leq \sum_{k=1}^{\infty} \frac{1}{k^2} \int_0^{\infty} \mathbb{1}_{\{y<k\}} 2y \mathbb{P}(|X_1| > y) dy \\
&= \int_0^{\infty} \underbrace{\sum_{k=1}^{\infty} \frac{1}{k^2} \mathbb{1}_{\{y<k\}} 2y}_{\leq 4} \mathbb{P}(|X_1| > y) dy && \text{by Fubini's} \\
&\leq 4 \int_0^{\infty} \mathbb{P}(|X_1| > y) dy && \text{By prev. lemma} \\
&= 4\mathbb{E}[|X_1|] < \infty
\end{aligned}$$

□

Lecture 18 11/21

Theorem 63. Strong Law of Large Numbers: Let $X_1, X_2, \ldots$ be a sequence of random variables that have identical distributions and are pairwise independent. Assume $\mathbb{E}[|X_1|] < \infty$ and let $\mathbb{E}[X_1] = \mu$ and $S_n = X_1 + \cdots + X_n$. Then, $\frac{S_n}{n} \overset{\text{a.s.}}{\to} \mu$.

Proof. Let $Y_k = X_k \mathbb{1}_{\{|X_k| \leq k|\}}$. Let $T_n = Y_1 + Y_2 + \cdots + Y_n$. We argue that it suffices to prove that $\frac{T_n}{n} \overset{\text{a.s.}}{\to} \mu$. To see this, note that

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_k| > k) \leq \int_0^{\infty} \mathbb{P}(|X_1| > t) dt = \mathbb{E}[|X_1|] < \infty$$

where $E[|X_1|] < \infty$ since $\mathbb{E}[X_1]$ exists. Now since $X_k \neq Y_k$ only when $|X_k| > k$, then by the first Borel-Contelli lemma, the probability $\mathbb{P}(X_k \neq Y_k \text{ i.o.}) = 0$. So almost surely, $X_k \neq Y_k$ only happens for finitely many k. This implies then that $\frac{S_n}{n} - \frac{T_n}{n} \overset{\text{a.s.}}{\to} 0$ since S_n and T_n only differ by finitely many terms, and $\frac{1}{n} \to 0$. Thus, $\frac{S_n}{n} \overset{\text{a.s.}}{\to} \mu$ if and only if $\frac{T_n}{n} \overset{\text{a.s.}}{\to} \mu$. Also note that since we can always separate $X = X^+ - X^-$ (where X^+ is the positive part and X^- is the negative part), it suffices to prove the result for non-negative $X \geq 0$ (as we can analyze each part separately).

The idea of the proof is to prove the convergence for a subsequence and then use the fact that $X \geq 0$ to get the general convergence through inequalities.

Let $\alpha > 1$ and let $k(n) = \lfloor \alpha^n \rfloor$. We will show that $\frac{T_{k(n)}}{k(n)}$ converges. Note by Chebyshev's

inequality, for any $\varepsilon > 0$

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|T_{k(n)} - \mathbb{E}[T_{k(n)}]| > \varepsilon k(n)) &\leq \varepsilon^{-2} \sum_{n=1}^{\infty} \frac{\operatorname{Var}(T_{k(n)})}{k(n)^2} \\ &= \varepsilon^{-2} \sum_{n=1}^{\infty} \left(\frac{1}{k(n)^2} \sum_{m=1}^{k(n)} \operatorname{Var}(Y_m) \right) \\ &= \varepsilon^{-2} \sum_{m=1}^{\infty} \operatorname{Var}(Y_m) \sum_{n \,:\, k(n) \geq m} \frac{1}{k(n)^2} \qquad \text{by Fubini's}\end{aligned}$$

where we can add the variances to get Y_m since our X_n are pairwise independent. Note that

$$\sum_{n \,:\, k(n) \geq m} \frac{1}{k(n)^2} = \sum_{n \,:\, k(n) \geq m} \frac{1}{\lfloor \alpha^n \rfloor^2} \leq 4 \sum_{n \,:\, \alpha^n \geq m} \alpha^{-2n} \leq 4 \cdot \frac{m^{-2}}{1 - \alpha^{-2}}$$

since $\lfloor \alpha^n \rfloor \geq \frac{\alpha^n}{2}$ for all $n \geq 1$ and $\alpha > 1$ and by bounding the sum by the geometric series $\sum_{n=m}^{\infty} \alpha^{-2n}$. Continuing our main proof, we have

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(|T_{k(n)} - \mathbb{E}[T_{k(n)}]| > \varepsilon k(n)) &\leq \varepsilon^{-2} \sum_{m=1}^{\infty} \operatorname{Var}(Y_m) \sum_{n \,:\, k(n) \geq m} \frac{1}{k(n)^2} \\ &\leq \varepsilon^{-2} 4 (1 - \alpha^{-2})^{-1} \underbrace{\sum_{m=1}^{\infty} m^{-2} \operatorname{Var}(Y_m)}_{< \infty} \\ &< \infty\end{aligned}$$

where we finiteness of the sum follows by our previous lemma. Now, by the first Borel-Contelli lemma,

$$\mathbb{P}\left(\frac{|T_{k(n)} - \mathbb{E}[T_{k(n)}]|}{k(n)} > \varepsilon \text{ i.o.} \right) = 0$$

Therefore, almost surely,

$$\limsup_{n \to \infty} \frac{|T_{k(n)} - \mathbb{E}[T_{k(n)}]|}{k(n)} \leq \varepsilon$$

Since this holds for all $\varepsilon > 0$, we conclude that $\frac{T_{k(n)} - \mathbb{E}[T_{k(n)}]}{k(n)} \overset{\text{a.s.}}{\to} 0$. Now since $Y_k \overset{\text{D}}{=} X_1 \mathbb{1}_{\{|X_1| \leq k\}} \overset{\text{a.s.}}{\to} X_1$ and $\mathbb{E}[|X_1|] < \infty$, then by dominated convergence theorem $\mathbb{E}[Y_k] \to \mathbb{E}[X_1] = \mu$. Thus, we see that $\frac{\mathbb{E}[T_{k(n)}]}{n} \to \mathbb{E}[X_1]$. Hence,

$$\frac{T_{k(n)}}{k(n)} - \mathbb{E}[X_1] = \underbrace{\left(\frac{T_{k(n)}}{k(n)} - \frac{\mathbb{E}[T_{k(n)}]}{k(n)} \right)}_{\overset{\text{a.s.}}{\to} 0} + \underbrace{\left(\frac{\mathbb{E}[T_{k(n)}]}{k(n)} - \mathbb{E}[X_1] \right)}_{\to 0} \overset{\text{a.s.}}{\to} 0$$

and so our result holds for the subsequence $\{Y_{k(n)}\}$, it remains only to show it for the general case.

For the general form, note that for any m, there is some n such that $k(n) \leq m \leq k(n+1)$. Now note that

$$\frac{T_{k(n)}}{k(n+1)} \leq \frac{T_m}{m} \leq \frac{T_{k(n+1)}}{k(n)} \tag{*}$$

since T_n is a partial sum of non-negative values, and is therefore increasing. Now note that $\frac{k(n+1)}{k(n)} \to \alpha$ and so

$$\frac{T_{k(n)}}{k(n+1)} = \underbrace{\frac{T_{k(n)}}{k(n)}}_{\overset{\text{a.s.}}{\to} \mathbb{E}[X_1]} \cdot \underbrace{\frac{k(n)}{k(n+1)}}_{\to \frac{1}{\alpha}} \overset{\text{a.s.}}{\to} \frac{1}{\alpha}\mathbb{E}[X_1]$$

Thus, taking limits in (*) we have

$$\frac{1}{\alpha}\mathbb{E}[X_1] \leq \liminf_{n\to\infty} \frac{T_m}{m} \leq \limsup_{n\to\infty} \frac{T_m}{m} \leq \alpha\mathbb{E}[X_1]$$

Then, taking $\alpha \to 1$, we have that $\frac{T_m}{m} \overset{\text{a.s.}}{\to} \mathbb{E}[X_1]$. □

Remark: Suppose that $X_1, X_2, \ldots$ are i.i.d. and that $\mathbb{E}[X_1^+] = \infty$ but $\mathbb{E}[X_1^-] < \infty$, then we have that $\frac{S_n}{n} \to \infty$. We would show this by a truncation argument showing that $\liminf_{n\to\infty} \frac{S_n}{n} > x$ for all x.

Example: In renewal theory, interval times for the (continuous) arrival of customers, or service times of a server, or lifespan of light bulbs, etc. are are all i.i.d. positive random variables. We can then define $T(n) = X_1 + \cdots + X_n$ is the time until the nth occurrence of the event. From this, we can define $N_t = \sup\{n : T(n) \leq t\}$ to be the number of occurrences by time t.

Theorem 64: Let $X_1, X_2, \ldots$ be i.i.d. positive random variables. Define $T(n) = X_1 + \cdots + X_n$ and $N_t = \sup\{n : T(n) \leq t\}$ as in the above example. If $\mathbb{E}[X_1] = \mu < \infty$, then $\frac{N_t}{t} \overset{\text{a.s.}}{\to} \frac{1}{\mu}$ as $t \to \infty$.

Proof. By the strong law of large numbers, $\frac{T(n)}{n} \overset{\text{a.s.}}{\to} \mu$. Note that $T(N_t) \leq t < T(N_t + 1)$ (i.e., in particular $T(N_t)$ is the time of the last event before time t, and $T(N_t+1)$ is the time of the next event after time t). We see then that

$$\frac{T(N_t)}{N_t} \leq \frac{t}{N_t} < \frac{T(N_t+1)}{N_t} = \frac{T(N_t+1)}{N_t+1} \cdot \frac{N_t+1}{N_t}.$$

Now since $\lim_{t\to\infty} N_t \to \infty$ and $\lim_{t\to\infty} \frac{N_t+1}{N_t} \to 1$, we have

$$\lim_{t\to\infty} \frac{t}{N_t} = \lim_{t\to\infty} \frac{T(N_t)}{N_t} = \lim_{n\to\infty} \frac{T(n)}{n} = \mu \text{ a.s.}$$

Hence,

$$\lim_{t\to\infty} \frac{N_t}{t} = \frac{1}{\mu} \text{ a.s.}$$

□

Lecture 19 11/26

Theorem 65. Glivenko-Contelli Theorem: Suppose $X_1, X_2, \ldots$ are i.i.d. samples form a known distribution (function) F. Let $F_n(x) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{\{X_n \leq x\}}$ be the "empirical distribution function." (Intuitively, $F_n(x)$ is estimating $\mathbb{P}(X_1 \leq x)$ by the fraction of our X_i's such that $X_i \leq x$.) Then $\sup_x |F_n(x) - F(x)| \to 0$ almost surely as $n \to \infty$. (Note this uniform convergence implies that note only do we have pointwise convergence, but also the shapes will converge.)

Proof. Note that almost sure pointwise convergence holds trivially as a result of apply strong law of large numbers to the indicator functions $I_n = \mathbb{1}_{\{X_n \leq x\}}$ for a fixed x. Note also that $F_n(x-) = \lim_{y \to x^-} F_n(y) = \frac{1}{n}\sum_{m=1}^{n} \mathbb{1}_{\{x_m < x\}}$ will converge almost surely to $\mathbb{P}(X_n < x) = F(x-)$. However, this does not directly imply uniform convergence (especially for non-continuous distribution functions). For $k = 1, 2, \ldots$ and $1 \leq j \leq k-1$, define $x_{j,k} := \inf\{y : F(y) \geq \frac{j}{k}\}$. Effectively, $x_{j,k}$ is the $\frac{j}{k}$th quantile. Then by the definition of $x_{j,k}$, we have that $F(x_{j,k}-) - F(x_{j-1,k}) \leq \frac{1}{k}$. Due to almost sure pointwise convergence, then almost surely there exists $N_k(\omega)$ such that $|F_n(x_{j,k}) - F(x_{j,k})| < \frac{1}{k}$ and $|F_n(x_{j,k}-) - F(x_{j,k}-)| < \frac{1}{k}$ for all $1 \leq j \leq k-1$ and $n \geq N_k(\omega)$. Thus, for $x \in [x_{j-1,k}, x_{j,k})$ and since both F_n and F are non-decreasing, we have that

$$F_n(x) \leq F_n(x_{j,k}-) \leq F(x_{j,k}-) + \frac{1}{k} \leq F(x) + \frac{2}{k}$$

On the other hand,

$$F_n(x) \geq F_n(x_{j-1}, k) \geq F(x_{j-1,k}) - \frac{1}{k} \geq F(x) - \frac{2}{k}$$

We see then that $|F_n(x) - F(x)| \leq \frac{2}{k}$ for all $n \geq N_k(\omega)$. Thus since this holds for all x, we have that $\sup_x |F_n(x) - F(x)| \leq \frac{2}{k}$. Moreover, this holds for all k, thus taking the limit as $n \to \infty$, we have that $\sup_x |F_n(x) - F(x)| \overset{a.s.}{\to} 0$. □

Lemma 66: For any $x \in \mathbb{R}$.

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^n}{m!} \right| \leq \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}$$

Proof. Use Taylor's Theorem. □

Theorem 67: If $\mathbb{E}[X^2] < \infty$ then

$$\varphi(t) = 1 + it\mathbb{E}[X] - \frac{t^2\mathbb{E}[X^2]}{2} + o(t^2)$$

Proof. Use the above lemma with $n = 2$. The error term becomes $\leq t^2\mathbb{E}[\frac{|t|\cdot|X|^3}{6} \wedge |X|^2]$. This converges to 0 as $t \to 0$ and is bounded by $|X^2|$, which is integrable (since $\mathbb{E}[X^2] < \infty$). Then by dominated convergence, the expectation inside of the expectation goes to 0 as $t \to 0$ (and is therefore $o(1)$). Therefore, multiplying this term by t^2 we get that the error term is $o(t^2)$. □

Theorem 68. Central Limit Theorem: Let $X_1, X_2, \dots$ be are i.i.d. random variables with $\mathbb{E}[X_1] = \mu < \infty$ and $\text{Var}(X_1) = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1) \qquad \iff \qquad \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$$

Proof. Since we can always take $X_i' = X_i - \mathbb{E}[X_i]$, it suffices to prove the result for $\mu = 0$. From the above theorem, $\varphi(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$ (since we have that $\mathbb{E}[X_i'] = 0$). Therefore, we see that

$$\varphi_{\frac{S_n}{\sigma\sqrt{n}}}(t) = \mathbb{E}[e^{it\frac{S_n}{\sigma\sqrt{n}}}] = \varphi_{S_n}(\tfrac{t}{\sigma\sqrt{n}}) = (\varphi(\tfrac{t}{\sigma\sqrt{n}}))^n = (1 - \tfrac{t^2}{2n} + o(\tfrac{t^3}{n}))^n \to e^{-t^2/2}$$

However, this is the characteristic function of a $\mathcal{N}(0,1)$! Thus we see that $\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$ by the continuity theorem. □

Example: Note that if $X_1, X_2, \dots \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$ random variables, then we know that $S_n = X_1 + \cdots + X_n \sim Bin(n,p)$. Note we know that $\mathbb{E}[X_1] = p$ and $\text{Var}(X_1) = p(1-p)$. Then, by the central limit theorem we know that $\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$. So, $S_n - np \approx \mathcal{N}(0, n\sigma^2)$ and $S_n \approx \mathcal{N}(np, n\sigma^2)$ (though this isn't a rigorous argument). Thus, we can approximate a $Binomial(n,p)$ distribution by an $\mathcal{N}(np, np(1-p))$ distribution for large n. For example, when $p = \frac{1}{2}$ and so $\sigma^2 = \frac{1}{4}$, we have

$$\mathbb{P}\left(\frac{S_n - \frac{n}{2}}{\frac{1}{2}\sqrt{n}} \in [a,b]\right) \approx F_Z(b) - F_Z(a) \quad \text{and} \quad \mathbb{P}(S_n \in [c,d]) \approx F_Z\left(\frac{d - \frac{n}{2}}{\frac{1}{2}\sqrt{n}}\right) - F_Z\left(\frac{c - \frac{n}{2}}{\frac{1}{2}\sqrt{n}}\right)$$

where $Z \sim \mathcal{N}(0,1)$. If you for instance take $n = 10000$, $c = 4900$ and $d = 5100$, then

$$\mathbb{P}(S_n \in [4900, 5100]) \approx F_Z(2) - F_Z(-2) \approx 0.95$$

Theorem 69. Lindeberg-Feller CLT: Consider a triangular array of random variables

$$\begin{array}{ccccc} X_{1,1} & & & & \\ X_{2,1} & X_{2,2} & & & \\ X_{3,1} & X_{3,2} & X_{3,3} & & \\ \vdots & \vdots & \vdots & \ddots & \end{array}$$

Where for each $n = 1, 2, \dots$ we have that $X_{n,m}$ for $m = 1, 2, \dots, n$ are independent random variables with $\mathbb{E}[X_{n,m}] = 0$. If (1) $\sum_{m=1}^n \mathbb{E}[X_{n,m}^2] \to \sigma^2 > 0$ for all $n = 1, 2, \dots$ and (2) for any fixed $\varepsilon > 0$, $\lim_{n\to\infty} \sum_{m=1}^n \mathbb{E}[|X_{n,m}|^2 \cdot \mathbb{1}_{\{|X_{n,m} > \varepsilon\}}] = 0$, then $S_n = X_{n,1} + \cdots + X_{n,n}$ has $S_n \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2)$. (1) and (2) are sometimes called the Lindeberg-Feller conditions. Intuitively, the result of a large number of independent random effects is approximately normal. Note also that we don't need independence between rows, and we don't need all random variables to have a shared distribution.

Proof. The result is omitted for brevity since it is quite technical. □

Chapter 6 Conditional Expectation

Lecture 20 11/28

Remark: In basic probability courses, we defined conditional expectation

$$\mathbb{E}[Y|X=x] = \begin{cases} \sum_y y \cdot \mathbb{P}(Y=y|X=x) & \text{when } X,Y \text{ are discrete} \\ \int y f_{Y|X}(y|x)dy & \text{when } X,Y \text{ are absolutely continuous} \end{cases}$$

where $\mathbb{P}(Y=y|X=x) = \frac{\mathbb{P}(Y=y,X=x)}{\mathbb{P}(X=x)}$ and $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$. It's unclear why these two should match, since $\mathbb{P}(X=x)=0$ for continuous X. However, how should we define conditional expectation in the general case (e.g., when X and Y are singular or not-absolutely continuous)?

Idea 1: Define the "conditional distribution" rigorously first, and then define the conditional expectation as the expectation of w.r.t. this conditional distribution. Turns out, this is very difficult to do rigorously since finding the distribution is not straightforward.

Idea 2: Define the conditional expectation from a different perspective, without conditional distributions.

Note: Note that $\mathbb{E}[Y|X=x]$ (using the non-rigorous definition of condiotional expectation we usually work with) is a number for each x. Therefore, $\mathbb{E}[Y|X=x]$ is a sort of mapping, sending each possible value of X to a number. Thus, we consider $\mathbb{E}[Y|X]$ as a random variable such that its value at any point ω is $E[Y|X=X(\omega)]$. Moreover, since $\mathbb{E}[Y|X]$ is determined by the value of X, it is a function of X. I.e., $\mathbb{E}[Y|X]=g(X)$ for some function g. In other words, $\mathbb{E}[Y|X]$ is a $\sigma(X)$-measurable random variable. Formalizing this intuition yields the following definition.

Definition. Conditional Expectation Given Sub-σ-field: Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space and X be a random variable on it, with $\mathbb{E}[|X|]<\infty$. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Then the conditional expectation of X given $\mathcal{G}$ is denoted by $\mathbb{E}[X|\mathcal{G}]$ and is a random variable Y such that

(1) Y is $\mathcal{G}$-measurable.

(2) For any $A\in\mathcal{G}$, $\int_A X d\mathbb{P} = \int_A Y d\mathbb{P}$.

Intuitively, $\mathbb{E}[X|\mathcal{G}]$ is the best estimate of X when we only have the resolution of $\mathcal{G}$. It's like looking through mosaic glass, we flatten the depth within each part of the mosaic to a single color. Similarly, Y lacks the resolution of X and is constant on portions where X is not constant.

Definition. Absolutely Continuous Measure: Let μ and ν be two measures defined on the same measurable space $(\Omega,\mathcal{F})$. Then ν is said to be absolutely continuous with respect to μ, denoted by $\nu \ll \mu$, if for any $A\in\mathcal{F}$ then whenever $\mu(A)=0$ we also have $\nu(A)=0$.

Theorem 70. Radon-Nikodym Theorem: Let μ and ν be σ-finite measure defined on the same measurable space $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, then there exists an $\mathcal{F}$-measurable function f such that for any measurable set $A \in \mathcal{F}$, $\int_A f d\mu = \nu(A)$. This f is often denoted $\frac{d\nu}{d\mu}$ and is called the Radon-Nikodym derivative of ν with respect to μ or density of ν with respect to μ.

Proof. This proof is from real analysis, and thus out-of-scope for this course. □

Proposition 71: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable on it, with $\mathbb{E}[|X|] < \infty$. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The conditional expectation $Y = \mathbb{E}[X|\mathcal{G}]$ exists.

Proof. (1) When $X \geq 0$. Let $\mu = \mathbb{P}|_{\mathcal{G}}$ (the restriction of $\mathbb{P}$ to $\mathcal{G}$). Note then that μ is a probability measure on $\mathcal{G}$ with $\mu(A) = \mathbb{P}(A)$ for all $A \in \mathcal{G}$. Now define $\nu(A) = \int_A X d\mu$ for any $A \in \mathcal{G}$. One can show that ν defined as such is a measure on $(\Omega, \mathcal{G})$. Moreover, since whenever $\mu(A) = 0$ then clearly $\nu(A) = \int_A X d\mu = 0$ so that $\nu \ll \mu$. Then, by Radon-Nikodym theorem, there is a $Y = \frac{d\nu}{d\mu} \in \mathcal{G}$ such that $\int_A X d\mathbb{P} = \nu(A) = \int_A Y d\mu = \int_A Y d\mathbb{P}$ which follows since Y is $\mathcal{G}$-measurable.

(2) In the general case, we can write $X = X^+ - X^-$. Let $Y_1 = \mathbb{E}[X^+|\mathcal{G}]$ and $Y_2 = \mathbb{E}[X^-|\mathcal{G}]$ so that $Y = Y_1 - Y_2 \in \mathcal{G}$. Then we can write

$$\int_A X d\mathbb{P} = \int_A X^+ d\mathbb{P} - \int_A X^- d\mathbb{P} = \int_A Y_1 d\mathbb{P} - \int_A Y_2 d\mathbb{P} = \int_A Y d\mathbb{P}$$

for all $A \in \mathcal{G}$.

□

Proposition 72: The conditional expectation is unique almost surely. That is, the conditional expectation is unique up to difference on sets of measure of 0.

Proof. Assume both Y and Y' satisfy the conditions in the definition of conditional expectation. By way of contradiction, suppose that $\mathbb{P}(Y \neq Y') > 0$. Without loss of generality, suppose that $\mathbb{P}(Y > Y') > 0$. Let $A := \{Y > Y'\}$ with $\mathbb{P}(A) > 0$. Since Y and Y' are $\mathcal{G}$-measurable, we have that $A \in \mathcal{G}$. Therefore, $\int_A X d\mathbb{P} = \int_A Y d\mathbb{P} = \int_A Y' d\mathbb{P}$. It follows then that $\int_A (Y - Y') d\mathbb{P}$. However, since $Y - Y' > 0$ on A and $\mathbb{P}(A) > 0$, it must be that $\int_A (Y - Y') d\mathbb{P} > 0$, a contradiction. (Note this follows since if $\mathbb{P}(Y - Y' > 0) > 0$, there must be an $\varepsilon > 0 > 0$ such that $\mathbb{P}(Y - Y' > \varepsilon) > 0$, and integrating over this area we get > 0 difference.) □

Definition. Conditional Expectation: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X, Y be random variables on it, with $\mathbb{E}[|Y|] < \infty$. Define $\mathbb{E}[Y|X] := \mathbb{E}[Y|\sigma(X)] = \mathbb{E}[Y|\mathcal{G}]$ where $\mathcal{G} = \{\{\omega \in \mathcal{F} : X \in A\} : A \in \mathscr{B}\}$.

Proposition 73. Properties of Conditional Expectation:

(1) Linearity: For $a \in \mathbb{R}$, $\mathbb{E}[aX + Y|\mathcal{F}] = a\mathbb{E}[X|\mathcal{F}] + \mathbb{E}[Y|\mathcal{F}]$.

(2) Monotonicity: If $X \leq Y$ then $\mathbb{E}[X|\mathcal{F}] \leq \mathbb{E}[Y|\mathcal{F}]$.

(3) Continuity: If $X_n \geq 0$ and $X_n \to X$ from below and $\mathbb{E}[|X| < \infty]$, then $\mathbb{E}[X_n|\mathcal{F}] \to \mathbb{E}[X|\mathcal{F}]$ from below.

Proof. (1) Follows trivially from the definition of conditional expectation and the linearity of integration.

(2) Can be proved using the same as idea as we used to show the uniqueness of conditional expectation. More precisely, define $A := \{\mathbb{E}[X|\mathcal{F}] > \mathbb{E}[Y|\mathcal{F}]\}$. Then $A \in \mathcal{F}$ and

$$\int_A \mathbb{E}[X|\mathcal{F}]d\mathbb{P} = \int_A X d\mathbb{P} \leq \int_A \leq Y d\mathbb{P} = \int_A \mathbb{E}[Y|\mathcal{F}]d\mathbb{P}$$

This implies that it must be that $\mathbb{P}(A) = 0$.

(3) Define $Y_n := X - X_n$ and $Z_n := \mathbb{E}[Y_n|\mathcal{F}] = \mathbb{E}[X|\mathcal{F}] - \mathbb{E}[X_n|\mathcal{F}]$ by linearity. Then we clearly have that $Y_n \to 0$ from above. Since $|Y_n| \leq |X|$ (and $|X|$ is integrable), then by the dominated convergence theorem, we have that $\int_A Y_n d\mathbb{P} \to 0$ from above for any $A \in \mathcal{F}$. But then, $\int_A Y_n d\mathbb{P} = \int_A Z_n d\mathbb{P}$ and so this implies that $\int_A \mathbb{Z}_n d\mathbb{P} \to 0$ from above. On the other hand, by property (2) we know that Z_n is decreasing. Denote the limit $Z_\infty := \lim_{n\to\infty} Z_n$. Then by monotone convergence theorem we know that $\int_A Z_n d\mathbb{P} \to \int_A Z_\infty d\mathbb{P}$. However, we showed that $\int_A Z_n d\mathbb{P} \to 0$, which implies that we must have $\int_A Z_\infty d\mathbb{P} = 0$ for all $A \in \mathcal{F}$. Now since Z_∞ is non-negative, it must be that $Z_\infty = 0$ almost surely. Therefore, $\mathbb{E}[X_n|\mathcal{F}] \to \mathbb{E}[X|\mathcal{F}]$ from below.

□

Note: If $X \in \mathcal{F}$ is $\mathcal{F}$-measurable, then $\mathbb{E}[X|\mathcal{F}] = X$ almost surely.

Theorem 74: If $\mathcal{F}_1 \subseteq \mathcal{F}_2$ are σ-fields, 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. (1) This holds trivially since $\mathbb{E}[X|\mathcal{F}_1] \in \mathcal{F}_1 \subseteq \mathcal{F}_2$.

(2) For any $A \in \mathcal{F}_1 \subseteq \mathcal{F}_2$, then

$$\int_A \underbrace{\mathbb{E}[X|\mathcal{F}_1]}_{\mathcal{F}_1\text{-measurable}} d\mathbb{P} = \int_A X d\mathbb{P} = \int_A \underbrace{\mathbb{E}[X|\mathcal{F}_2]}_{\mathcal{F}_2\text{-measurable}} d\mathbb{P}$$

This implies that $\mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1] = \mathbb{E}[X|\mathcal{F}_1]$.

□

Remark: Note that by the above result, we have $\mathbb{E}[\mathbb{E}[X|\mathcal{F}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}]|\{\emptyset, \Omega\}] = \mathbb{E}[X]$.

Theorem 75: Suppose X and Y are random variables. If X is $\mathcal{F}$-measurable and $\mathbb{E}[|X|] < \infty$, and $\mathbb{E}[|XY|] < \infty$, then $\mathbb{E}[XY|\mathcal{F}] = X\mathbb{E}[Y|\mathcal{F}]$.

Proof. We use the definition of conditional expectation and check conditions (1) and (2). First, we check that $X\mathbb{E}[Y|\mathcal{F}]$ is $\mathcal{F}$ measurable. But since X is $\mathcal{F}$ measurable and $\mathbb{E}[Y|\mathcal{F}]$ is necessarily $\mathcal{F}$ measurable this holds trivially. First, for $X = \mathbb{1}_B$ with $B \in \mathcal{F}$, and $A \in \mathcal{F}$ has

$$\int_A X\mathbb{E}[Y|\mathcal{F}]d\mathbb{P} = \int_A \mathbb{1}_B\mathbb{E}[Y|\mathcal{F}]d\mathbb{P} = \int_{A\cap B} \mathbb{E}[Y|\mathcal{F}]d\mathbb{P} = \int_{\underbrace{A\cap B}_{\in\mathcal{F}}} Yd\mathbb{P} = \int_A \mathbb{1}_B Yd\mathbb{P} = \int_A XYd\mathbb{P}$$

so we see that this holds for $X = \mathbb{1}_B$. By the linearity of integration, this also holds for any simple function. Now, by taking an increasing limit $X_n \to X$, which is non-negative and integrable, the result also holds. Finally, we can generalize to any random variable by decomposing it into its positive and negative parts. We're basically following the same method we used for Lebesgue integration. □

Theorem 76: Let X and Y be independent random variables with $\mathbb{E}[|Y|] < \infty$. Then $\mathbb{E}[Y|X] = \mathbb{E}[Y]$.

Proof. We know that any set in $\sigma(X)$ has the form $\{X \in B\}$ for some $B \in \mathscr{B}$. Then, for any $A \in \sigma(X)$ we have

$$\int_A Yd\mathbb{P} = \int Y \cdot \mathbb{1}_A d\mathbb{P}$$

Since $A = \{X \in B\}$ for some B. Thus, $\mathbb{1}_A$ is a function of X, and so $\mathbb{1}_A$ and Y are independent. Then by independence, we have

$$\mathbb{E}[Y \cdot \mathbb{1}_A] = \mathbb{E}[Y] \cdot \mathbb{E}[\mathbb{1}_A] = \int_A \mathbb{E}[Y]d\mathbb{P}$$

We see then that $\int_A Yd\mathbb{P} = \int_A \mathbb{E}[Y]d\mathbb{P}$ for all $A \in X$. This, in turn, implies that $\mathbb{E}[Y|X] = \mathbb{E}[Y]$, as desired. □

Theorem 77. Conditional Jensen's inequality: Let φ be a convex function, and X be a random variable such that $\mathbb{E}[|X|] < \infty$ and $\mathbb{E}[|\varphi(X)|] < \infty$. Then $\varphi(\mathbb{E}[X|\mathbf{f}]) \leq \mathbb{E}[\varphi(X)|\mathcal{F}]$

Proof. Follows by applying Jensen's inequality to the conditional distribution. □

Corollary 78: $\mathbb{E}[X|\mathcal{F}]$ has a small or equal L^p norm than X for any $p \geq 1$.

Proof. Note that $|x|^p$ is a convex function. Hence, by Jensen's inequality we have that

$$|\mathbb{E}[X|\mathcal{F}]|^p \leq \mathbb{E}[|X|^p|\mathcal{F}]$$

Thus, taking the expectation again on both sides we get that

$$\|\mathbb{E}[X|\mathcal{F}]\|_p^p = \mathbb{E}[|\mathbb{E}[X|\mathcal{F}]|^p] \leq \mathbb{E}[\mathbb{E}[|X|^p|\mathcal{F}]] = \mathbb{E}[|X|^p] = \|X\|_p^p$$

□

Proposition 79. Orthogonality: Let X be a random variable such that $\mathbb{E}[X^2] < \infty$ and $\mathcal{F}$ be a σ-field. Let Y be a random variable such that $\mathbb{E}[Y^2] < \infty$ and Y is $\mathcal{F}$-measurable. Then $X - \mathbb{E}[X|\mathcal{F}]$ and Y are uncorrelated (i.e., $\text{Cov}(X - \mathbb{E}[X|\mathcal{F}], Y) = 0$). This in particular implies that $\mathbb{E}[X|\mathcal{F}]$ and $X - \mathbb{E}[X|\mathcal{F}]$ are uncorrelated.

Proof. Note that $\mathbb{E}[X - \mathbb{E}[X|\mathcal{F}]] = 0$. Hence, it suffices to show that $\mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}])Y] = 0$ (since $\text{Cov}(X, y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$). We see that by adding another layer of conditional expectation gives us

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}]) \cdot Y] = \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}]) \cdot Y]|\mathcal{F}] = \mathbb{E}[Y \cdot \underbrace{\mathbb{E}[X - \mathbb{E}[X|\mathcal{F}]|\mathcal{F}]}_{=0}] = \mathbb{E}[Y \cdot 0] = 0$$

□

Theorem 80. Minimal Distance: Let X be a random variable with $\mathbb{E}[X^2] < \infty$. Then for any $\mathcal{F}$-measurable random variable Z with $\mathbb{E}[Z^2] < \infty$ we have that

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}])^2] \leq \mathbb{E}[(X - Z)^2].$$

Proof. Note that

$$\mathbb{E}[(X - Z)^2] = \mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}] + \mathbb{E}[X|\mathcal{F}] - Z)^2]$$

We have seen that $X - \mathbb{E}[X|\mathcal{F}]$ and $E[X|\mathcal{F}] - Z$ are uncorrelated, however, and so

$$\mathbb{E}[(X-Z)^2] = \mathbb{E}[(X-\mathbb{E}[X|\mathcal{F}]+\mathbb{E}[X|\mathcal{F}]-Z)^2] = \mathbb{E}[(X-\mathbb{E}[X|\mathcal{F}])^2]+\mathbb{E}[(\mathbb{E}[X|\mathcal{F}]-Z)^2] \geq \mathbb{E}[(X-\mathbb{E}[X|\mathcal{F}])^2]$$

□

Example. Wald's Identity: Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbb{E}[|X_1|] < \infty$. Let N be a non-negative, integer-valued random variable with with $\mathbb{E}[|N|] < \infty$ and independent of $X_1, X_2, \ldots$. Then,

$$\mathbb{E}\left[\sum_{n=1}^{N} X_n\right] = \mathbb{E}[N] \cdot \mathbb{E}[X_1]$$

Proof.

$$\mathbb{E}\left[\sum_{n=1}^{N} X_n\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{n=1}^{N} X_n|N\right]\right] = \mathbb{E}[N \cdot \mathbb{E}[X_1]] = \mathbb{E}[N] \cdot \mathbb{E}[X_1]$$

since for $N = k$ the inner expectation is

$$= \mathbb{E}\left[\sum_{n=1}^{N} X_n|N = k\right] = \mathbb{E}\left[\sum_{n=1}^{k} X_n|N = k\right] = \mathbb{E}\left[\sum_{n=1}^{k} X_n\right] = k\mathbb{E}[X_1]$$

□

Definition. Conditional Variance: Let X, Y be random variables such that $\mathbb{E}[X^2] < \infty$. Define the conditional variance as $\text{Var}(X|Y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y$.

Example. EVVE's Law: Let X, Y be random variables such that $\mathbb{E}[X^2] < \infty$. Then, $\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y])$.

Proof.

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[((X - \mathbb{E}[X|Y]) + (\mathbb{E}[X|Y] - \mathbb{E}[X]))^2] \\
&= \mathbb{E}[(X - \mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y] - \mathbb{E}[X])^2] \\
&= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X - Y])^2|Y]] + \mathbb{E}[(\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Y]])^2] \\
&= \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y])
\end{aligned}$$

This can be interpreted as grouping our observations according to the value of Y. Then the overall variance is the sum of the intra-group variance $\mathbb{E}[\text{Var}(Y|X)]$ and inter-group variance $\text{Var}(\mathbb{E}[X|Y])$. □

Note: In STAT 902 (Theory of Probability 2) is the continuation of this course but towards stochastic calculus. Particularly, learning about stochastic processes X_t which change according to time t. You will learn about stochastic differential equations $X_t d\mathsf{B}_t$ where B_t is the Brownian motion. This allows us to integrate w.r.t. other stochastic calculus $\int X_t d\mathsf{B}_t$ via Ito integration. This will also involve learning about Martingales, and particular stochastic calculus with continuous semi-martingales. In order the topics are

- Discrete and continuous martingale theory. A martingale is a stochastic process $\mathbb{E}[X_t|\mathcal{F}_s] = X_s$.
- Brownian motion.
- Itô integration w.r.t. Brownian motion.
- Semi-martingales and Itô integration w.r.t continuous semi-martingales.
- SDEs, local times, etc.

Index