DISTRIBUTION THEORY

ABSTRACT. Distribution theory is a class by professor Sun. It is an introductory level probabilistic course without heavy inclination to measure theory based proofs.

I shall here present the class material by courses, and might re-organize them into topics.

CONTENTS

1. 9/27 : Probability Spaces and Random Variables

1.1. **Probability Spaces.**

As a start, let's just say that

- Ω - sample space.
- Subsets of Ω - events.

Def 1.1. *A set of subsets $\mathcal{F}$ of Ω is called an algebra if*

- $\Omega \in \mathcal{F}$.
- *if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.*
- *if $A, B \in \mathcal{F}$, so is $A \cap B$.*

Example 1.1. *Let $\Omega = (0, 1]$, then*

- $\mathcal{F} = \{\emptyset, (0, 1], (0, \frac{1}{2}], (\frac{1}{2}, 1]\}$ *is an algebra*
- $\mathcal{F} = \{\emptyset, (0, 1], (0, \frac{1}{2}), (\frac{1}{2}, 1]\}$ *is* ***not*** *an algebra*

Def 1.2. *A σ-algebra $\mathcal{F}$ is an algebra that is closed under countable intersection (and union).*

Def 1.3. *A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:*

- *Ω is the sample space.*
- *$\mathcal{F}$ is a σ-algebra.*
- *$\mathbb{P}$ is a probability measure.*

where a probability measure is just a function from $\mathcal{F}$ into $[0, 1]$ that is defined below:

Def 1.4. *A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a function such that:*

- $\mathbb{P}(\Omega) = 1$.
- $\mathbb{P}(A) \geq 0$.

- $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(A_i\right)$ *for distinct* A_i.

Example 1.2.

Let

- $\Omega = [0, 1]$.
- $\mathcal{F}$ = the set of all Borel subsets of $[0, 1]$.
- $\mathbb{P}$ = the Lebesgue measure on $[0, 1]$.

Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Remark 1.1. *Why do we post countable additivity in the definition of σ-algebra?*

Well, the answer is that if we want to impose as much as possible, but not too much such that something goes wrong. And taking uncountable additivity will make things go wrong, as illustrated in the following example.

Example 1.3.

If uncountable union is allowed, then we can get

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{x\in[0,1]} \{x\}\right) = \sum_{x\in[0,1]} \mathbb{P}\left(\{x\}\right)$$

Now the problem is that there is no value for $\mathbb{P}(\{x\})$. Out of consistency reason we really want to have the same measure for all singletons, which means that

- if $\mathbb{P}(\{x\}) = 0$, then the right hand side of our expression will equal to 0, not 1.
- if $\mathbb{P}(\{x\}) \neq 0$, then the right hand side of our expression will be ∞, not 1.

which means that this definition is inconsistent.

1.2. **Properties of Probability Measures.**

Let's now look more of probability measures.

Proposition 1.1.

(*a*) *if $A \subset B \in \mathcal{F}$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.*
(*b*) *if $A_n \uparrow A$, $A_1 \subseteq A_2 \subseteq \ldots$, $\lim_{n\to\infty} A_n = A$, then $\mathbb{P}(A_n) \to \mathbb{P}(A)$ as $n \to \infty$.*

The proof is straight forward since there really isn't much tool to use, so we just use the only non-trivial axiom.

Proof. (a):

$$\mathbb{P}(B) = \mathbb{P}\left(A \cup (B - A)\right) = \mathbb{P}(A) + \mathbb{P}(B - A)$$

since $\mathbb{P}(B_A) \geq 0$, we get what we want.

(b):

$$A = \bigcup_{n=1}^{\infty} (A_n - A_{n-1})$$

where $A_0 = \emptyset$, and we get

$$\begin{aligned}\mathbb{P}(A) &= \sum_{n=1}^{\infty} \mathbb{P}(A_n - A_{n-1}) = \sum_{n=1}^{\infty} \Big(\mathbb{P}(A_n) - \mathbb{P}(A_{n-1})\Big) \\ &= \lim_{N\to\infty} \sum_{n=1}^{N} \Big(\mathbb{P}(A_n) - \mathbb{P}(A_{n-1})\Big) = \lim_{n\to\infty} \mathbb{P}(A_n)\end{aligned}$$

□

But wait! Before we go to the next topic, what does $\lim_{n\to\infty} A_n = A$ even mean?

Def 1.5. *(Set limit)*

$\underline{\lim_{n\to\infty} A_n = A}$ *if* $\forall x \in A, \exists N$ *s.t.* $x \in A_n$ *for all* $n \geq N$. *And if* $x \notin A$, $\exists N$ *s.t.* $x \notin A_n$ *for all* $n \geq N$.

1.3. **Random Variables.**

Def 1.6. *A* $\underline{\text{random variable}}$ *is a map* $X : \Omega \to \mathbb{R}$ *such that* $\forall t \in \mathbb{R}$, *we have*

$$\{\omega \in \Omega | X(\omega) \leq t\} \in \mathcal{F}$$

i.e. X *is a measurable function from* $(\Omega, \mathcal{F})$ *to* $(\mathbb{R}, \mathcal{B})$.

Def 1.7. *The* $\underline{\text{cumulative distribution function } F_x(t)}$ *is defined by*

$$F_x(t) = \mathbb{P}(\{\omega | X(\omega) \leq t\})$$

Well, the definition is obviously well-defined since we kind of "cheated" to make things easier when defining random variables. For simplification, we write

$$\{X \leq t\} := \{\omega | X(\omega) \leq t\}$$

Example 1.4. *Let's consider two coin tosses:*

- $\Omega = \{HH, HT, TH, TT\}$.
- $\mathcal{F} = P(\Omega)$*, the power set of* Ω.
- $X =$ *number of heads.*

$$\textit{Then } F_x(t) = \begin{cases} 0 & t \le 0 \\ \frac{1}{4} & t \in [0,1) \\ \frac{3}{4} & t \in [1,2) \\ 1 & t \ge 2 \end{cases}$$

From the graph(maybe) of the function above, we can guess that it is a Càdlàg function.

Def 1.8. *A function on reals is called Càdlàg (continue à droite, limite à gauche), or RCLL (Right continuous with Left Limits) if for all* $t \in \mathbb{R}$, $\lim_{x \to t^-} f(t)$ *exists and* $\lim_{x \to t^+} f(x) = f(t)$.

Proposition 1.2. *The distribution function* F *satisfies*

(a) F *is non-decreasing.*
(b) $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$.
(c) F *is Càdlàg.*

Still, we use what we only have at hand.

Proof.

(a)For $x < y$, we want $F(x) \le F(y)$. But $F(x) = \mathbb{P}(\{X \le x\})$, $F(x) = \mathbb{P}(\{X \le y\})$, and $\{X \le x\} \subseteq \{X \le y\}$. By proposition 1.1 (a) we are done.

(b) Since $\lim_{x \to \infty} F(x) = \lim_{x \to \infty} \mathbb{P}(\{X \le x\})$ and $\{X \le n\} \uparrow \Omega$, by proposition 1.1 (b) we have

$$\lim_{n \to \infty} \mathbb{P}(\{X \le n\}) = \mathbb{P}(\Omega) = 1$$

and the other side is proven in the exact same manor.

(c)Left limits: we want to prove that if $x_n \uparrow t^-$, then $\lim_{x_n \to t^-} F(x_n)$ exists.

Note that $\lim_{x_n \to t^-} F(x_n) = \lim_{x_n \to t^-} \mathbb{P}(\{X \le x_n\})$ and $\mathbb{P}(\{X \le x_n\})$ is non-decreasing.

Since $\{X \le x_n\} \subseteq \{X \le t\}$, we also know that $\mathbb{P}(\{X \le x_n\})$ is bounded above by $\mathbb{P}(\{X \le t\})$.

Therefore, since a bounded monotone sequences has a limit in $\mathbb{R}$, there is a left limit.

Right continuous: We want to show that if $x_n \downarrow t^+$, then $\lim_{x_n \to t^+} F(x_n) = F(t)$.

Now $\lim_{x_n \to t^+} F(x_n) = \lim_{x_n \to t^+} \mathbb{P}(\{X \le x_n\})$. Since $\{X \le x_1\} \supseteq \{X \le x_2\} \supseteq \dots$ implies

$$\{X \le x_1\} = \{X \in (x_2, x_1]\} \cup \{X \in (x_3, x_2]\} \cup \cdots \cup \{X \in (-\infty, t]\}$$

which implies

$$F(x_1) = F(t) + \sum_{n=1}^{\infty} \left(F(x_n) - F(x_{n+1}) \right)$$

$$\Rightarrow F(x_1) - F(t) = \lim_{N\to\infty} \sum_{n=1}^{N} \Big(F(x_n) - F(x_{n+1}) \Big) = F(x_1) - \lim_{N\to\infty} F(x_N)$$

$$\Rightarrow F(t) = \lim_{n\to\infty} F(x_n)$$

□

2. 9/29 Representing function of a random variable

We have proven in proposition 1.2 some properties of the distribution function. But it is natural to ask whether the proposition can be inverted, i.e., suppose a function has those properties, would it be the distribution of some random variable?

The answer to this question is yes, which is what the following theorem says.

Proposition 2.1. *If a function G satisfies*

- *F is non-decreasing*
- $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$
- *F is Càdlàg.*

Then G is the distribution function of some random variable X.

Before proving the proposition, let's look at an easy case when besides the three properties, G is also continuous and strictly increasing:

But this means that G has an inverse $G^{-1} : (0,1) \to \mathbb{R}$. Take $X = G^{-1}(U)$ where U is the uniform distribution on (0,1), then we claim that X has distribution function equal to G.

This is because

$$\mathbb{P}(x \le t) = \mathbb{P}(G^{-1}(U) \le t) = \mathbb{P}(U \le G(t)) = \mathbb{P}(\{\omega | U(\omega) \le G(t)\}) = G(t).$$

Moving on to the more general case, we define a analogous function to G^{-1} (which doesn't exist since G might not be invertible).

Def 2.1. $G^-(u) = \inf\{x | G(x) \ge u\}$

Our goal, of course, is to show that this to prove that $X(u) = G^-(u)$ has distribution function G. But first let us see an example of the function G^- and get the idea of how G^- can be obtained.

Example 2.1.

Suppose $G = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$, then for example $G^-\left(\frac{1}{2}\right) = \inf_x\{x | G(x) \ge \frac{1}{2}\} = 0$ and $G^-\left(\frac{1}{2} + \varepsilon\right) = \inf_x\{x | G(x) \ge \frac{1}{2} + \varepsilon\} = 1$.

Anyway, after a few tryouts we can see that $G^-(u) = \begin{cases} 0 & 0 < u \le \frac{1}{2} \\ 1 & \frac{1}{2} < u < 1 \end{cases}$.

Therefore we can see from the expression of x(which is G^-!) that there's $\frac{1}{2}$ probability of x being 0 and $\frac{1}{2}$ being 1.

This process of generating G^- can be done by flipping the function with respect to the line $x = y$, deleting all vertical part and adding all missing horizontal part in a left-continuous pattern. Personally, this pattern makes sense because the point of continuous is the exact point that is continuous (to the right here).

Now in order to prove proposition 2.1, we first do a close observation on G^-.

Proposition 2.2. *If G is a function satisfies*

- *F is non-decreasing.*
- $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.
- *F is Càdlàg.*

then G^- satisfies

(1) For $0 < u < 1$, we have

$$u > G(x) \iff G^-(u) > x \quad \textit{and} \quad u \le G(x) \iff G^-(u) \le x\ .$$

(2) $G^-(G(x)) \le x$.
(3) $G(G^-(u)) \ge u$.
(4) G^- is non-decreasing.
(5) G^- has right limits and is left continuous.

Proof.

(1) We only prove the first inequality since they are the same. If $u > G(x)$, then

$$G^-(u) = \inf\{y|G(y) \ge u\} > x.$$

(2) Plugging in $u = G(x)$ we get

$$G^-(G(x)) = \inf\{y|G(y) \ge G(x)\} \le x$$

(3) Plugging in $x = G^-(u)$ we get

$$G(G^-(u)) = G(\inf\{x|G(x) \ge u\}) \ge u$$

(4) For $u < v$, we want $G^-(u) \le G^-(v)$. We prove it with (1) and (3) above:

$$u < v \le G(G^-(v)) \Rightarrow G^-(u) \le G^-(v)$$

(5) Right limit: by bounded monotone convergence (very similar to proof of proposition 1.2).

Left continuous: Suppose $u_n \to u-$ and since by the same argument as right limit we know the left limit exists, i.e. $\lim_{n\to\infty}(G^-(u_n)) = L$.

We know that $L \le G^-(u)$, so we only need to show $L \ge G^-(u)$. But by (1) we only need to show $G(L) \ge u$. But this is true since $G(G^-(u_n)) \ge u_n$ and thus $G(L) \ge u_n$ for all n.

□

Now we prove proposition 2.1, which is fairly straight forward using (1).

Proof. (proposition 2.1) Let $X = G^-(U)$ for uniform U, then

$$F_X(t) = \mathbb{P}(G^-(U) \leq t) = \mathbb{P}(U \leq G(t)) = G(t)$$

□

Now we look at the inequality $F(F^-(u)) \geq u$ and wonder when the equality hold.

Proposition 2.3. *$F(X) \sim$ Uniform $[0,1]$ ($X = F^-(U)$) iff F is continuous.*

Proof. $\Leftarrow$: Suppose F is continuous, then since we know $u \leq F(F^-(u))$, we only want to show $u \geq F(F^-(u))$. Since F is continuous it is left continuous, so $F(F^-(u)) = \lim_{x \to F^-(u)^-} F(x)$. Now by property (1) we have $x < F^-(u) \iff F(x) < u$, and since each x in the limit is less than $F^-(u)$, the limit is less or equal to u, which means $F(F^-(u)) \leq u$, and we are done.

$\Rightarrow$: If F is not continuous, Then $\exists x_0$ such that $F(x_0) > \lim_{x \to x_0^-} F(x)$ which means

$$\mathbb{P}(X = x_0) = \mathbb{P}(X \leq x_0) - \mathbb{P}(X < x_0) = F(x_0) - \lim_{x \to x_0^-} F(x) > 0.$$

Note that if $F(X)$ is uniform then $\mathbb{P}(F(X) = t) = 0$ for any $t \in [0,1]$. Yet now

$$\mathbb{P}(F(X) = F(x_0)) = \mathbb{P}(\{\omega | F(X(\omega)) = F(x_0)\}) \geq \mathbb{P}(x = x_0) > 0$$

contradiction! □

Def 2.2. *A function $R : (0,1) \to \mathbb{R}$ is a representing function for a random variable X if*

- *R is non-decreasing;*
- *$R(U)$ has the same distribution function as X for uniform U.*

For example, if X has distribution function F, then F^- is a representing function for X.

Theorem 2.4. *If X has distribution function F, then $R : (0,1) \to \mathbb{R}$ is a representing function for X iff $F^-(u) \leq R(u) \leq F^-(u+)$ where $F^-(u+) = \lim_{v \to u^+} F^-(v)$.*

Proof. $\Leftarrow$: Suppose $F^-(u) \leq R(u) \leq F^-(u+)$. We need to show the two properties in the definition of a representing function.

Non-decreasing: For $u < v$,

$$R(u) \leq F^-(u+) \leq F^-(v) \leq R(v).$$

$R(U)$ has the same distribution function as X: Since X has the same distribution function as F^-, we only need to show that $R(U)$ has the same distribution as F^-. To show this we only need to notice

$$\mathbb{P}(R(U) \neq F^-(U)) = \mathbb{P}(\{\omega | F^- \text{ is discontinous at } \omega\}) = 0$$

since the set of discontinuities is countable, and thus has measure 0.

$\Rightarrow$: Suppose R is a representing function of X.

$R(u) \geq F^-(u)$: We note that

$$R(u) \geq F^-(u) \iff F(R(u)) \geq u$$

and since

$$F(R(u)) = \mathbb{P}(X \leq R(u)) = \mathbb{P}(R(U) \leq R(u)) \geq u$$

we are done.

$R(u) \leq F^-(u+)$: We want to show that $\forall v > u, F^-(v) \geq R(u)$, which can be deduced from $\forall w < R(u), \forall v > u, w < F^-(v)$. The trick here is to get a strictly less than instead of a less or equal to. Now we can use

$$F(w) = \mathbb{P}(X \leq w) \leq \mathbb{P}(X < R(u)) = \mathbb{P}\left(R(U) < R(u)\right) \leq u$$

where we see that the last inequality comes exactly from the strictly less sign.

Now, if $v > u$, then $v > u \geq F(w)$, which means $F^-(v) > w$, and thus clocludes the last bit. $\square$

3. 10/4: Quantile, QQ-plots, Independence of R.V., Some discrete R.V.

3.1. Quantiles.

Def 3.1. *For $0 < \alpha < 1$, x is an α-quantile for X if*

$$\begin{cases} \mathbb{P}(X \le x) \ge \alpha \\ \mathbb{P}(X \ge x) \ge 1 - \alpha \quad (\iff \mathbb{P}(X < x) \le \alpha) \end{cases}$$

Example 3.1.

- If F is continuous and strictly increasing, then $x_\alpha = F^{-1}(\alpha)$ is unique.
- If $X \sim Binomial(2, \frac{1}{2}) = \begin{cases} 0 & p = \frac{1}{4} \\ 1 & p = \frac{1}{2} \\ 2 & p = \frac{1}{4} \end{cases}$, then $\begin{cases} \text{the 0.3-quantile is 1.} \\ \text{anything in [0,1] is a 0.25-quantile.} \end{cases}$

A claim that we will make here and not proof (since it's easy) is the following.

Proposition 3.1. *If $a < b$ are α-quantiles, then $\forall c \in [a, b]$ is an α-quantile.*

Def 3.2. *A function $Q : (0, 1) \to \mathbb{R}$ is a quantile function if $Q(\alpha)$ is an α-quantile for all α.*

Theorem 3.2. *A function $Q(\alpha)$ is a quantile function for X iff it is a representing function for X.*

Proof. By definition,

$$Q(\alpha) \text{ is } \alpha\text{-quantile} \iff \begin{cases} F(Q(\alpha)) \ge \alpha \\ F(Q(\alpha)-) \le \alpha \end{cases} \iff \begin{cases} Q(\alpha) \ge F^-(\alpha) \\ ? \end{cases}$$

and we need $Q(\alpha) \le F^-(\alpha+)$ at the position of ? by theorem 2.4.

So really, what we need to do is to prove

$$F(Q(\alpha)-) \le \alpha \iff Q(\alpha) \le F^-(\alpha+)$$

$\Rightarrow$: Since $\forall w \le Q(\alpha)$ we have $F(w) \le \alpha$, so for $\beta > \alpha$,

$$\beta > \alpha \ge F(w) \quad \Rightarrow \quad F^-(\beta) > w \quad \Rightarrow \quad F^-(\alpha+) \ge Q(\alpha).$$

$\Leftarrow$: We need to show $F(Q(\alpha)-) \le \alpha$, but note that for $\forall w < Q(\alpha), \forall \beta > \alpha$,

$$F(Q(\alpha)-) \le \alpha \iff F(w) \le \alpha \iff F(w) < \beta \iff w < F^-(\beta)$$

where the last inequality holds because

$$Q(\alpha) \le F^-(\alpha+) \Rightarrow F^-(\beta) \ge Q(\alpha) \Rightarrow F^-(\beta) > w.$$

□

3.2. **QQ-Plots (Quantile-Quantie plot).**

Suppose we have random variables x_1, x_2 with quantile function $F_1^- = Q_1, F_2^- = Q_2$.

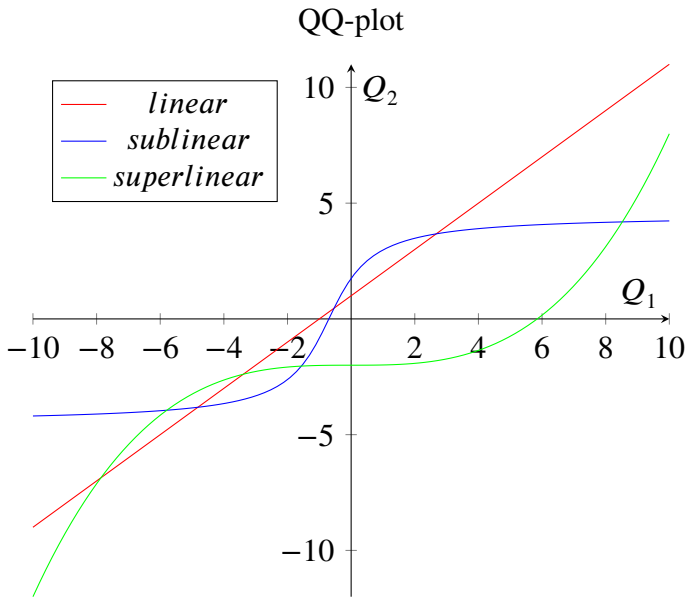
Def 3.3. *The QQ-plot of* (X_1, X_2) *is the parametric plot of* $(Q_1(u), Q_2(u))$ *for* $u \in (0, 1)$.

Proposition 3.3. *If* T *is non decreasing and* $X_2 = T(X_1)$*, then* $Q_2 = T(Q_1)$.

A special case would be the linear case:

$$T(x) = mx + b \quad \Rightarrow \quad Q_2(u) = mQ_1(u) + b \quad \Rightarrow \quad \text{The QQ-plot is linear.}$$

This means that for X_1 fixed an known, for instance $X_1 \sim N(0, 1)$, and X_2 is just some empirical sampling distribution. For example:



- Since the red line is linear, X_2 is roughly Gaussian.
- Since the blue line is sublinear, X_2 is roughly less than Gaussian at the end points, that is, it might be Gaussian with light tails.
- Since the blue line is superlinear, X_2 is roughly less than Gaussian at the end points, that is, it might be Gaussian with light tails.

3.3. **Independence of random variables.**

Def 3.4. *For random variables* $X_1, \ldots, X^n$ *on* $(\Omega, \mathcal{F}, \mathbb{P})$*, then they are independent iff*

$$\mathbb{P}(X_1 \in A_1, \ldots, X^n \in A^n) = \mathbb{P}(X_1 \in A_1) \cdots \mathbb{P}(X_n \in A_n)$$

for all sets of Borel sets $\{A_1, \ldots, A_n\}$.

This definition is really saying that knowing the behavior of one of the variables does not help much about knowing others. Other than that...well, the definition is really hard to check

since we need to go over all sets of Borel sets. Therefore we give a easier to check description of the idea of independence.

Proposition 3.4. *For independence, we only need*

$$\mathbb{P}(X_1 \leq c_1, \dots, X^n \leq c_n) = \mathbb{P}(X_1 \leq c_1) \cdots \mathbb{P}(X_n \leq c_n). \tag{3.1}$$

What we did is to pick a particular set of Borel sets such that checking this is sufficient for independence.

Proof. We only prove the case where $n = 2$. For larger n, the procedure is the same.

We only prove that we can do for squares (with different boundary conditions), and by Darboux partition we know we'll be done in the plane. A generalization of this is not simple but easy to get the idea.

Maybe we want to choose measurable set $(b_1, c_1]$ and $(b_2, c_2]$. The independence property also holds here since

$$\begin{aligned}\mathbb{P}(X_1 \in (b_1, c_1], X_2 \in (b_2, c_2]) &= \mathbb{P}(X_1 \leq c_1, X_2 \leq c_2) - \mathbb{P}(X_1 \leq b_1, X_2 \leq c_2) \\ &\quad - \mathbb{P}(X_1 \leq c_1, X_2 \leq b_2) + \mathbb{P}(X_1 \leq b_1, X_2 \leq b_2) \\ &= \big(\mathbb{P}(X_1 \leq c_1) - \mathbb{P}(X_1 \leq b_1)\big)\big(\mathbb{P}(X_2 \leq c_2) - \mathbb{P}(X_2 \leq b_2)\big) \\ &= \mathbb{P}(X_1 \in (b_1, c_1])\mathbb{P}(X_2 \in (b_2, c_2])\end{aligned}$$

This deals with all squares with upper and right boundary closed and the other two sides open. We will show that an open square has also this property and thus be done with the proof.

$$\begin{aligned}\mathbb{P}(X_1 \in (b_1, c_1), X_2 \in (b_2, c_2)) &= \lim_{x \to c_1^-, y \to c_2^-} \mathbb{P}(X_1 \in (b_1, x])\mathbb{P}(X_2 \in (b_2, y]) \\ &= \lim_{x \to c_1^-, y \to c_2^-} \mathbb{P}(X_1 \in (b_1, x])\mathbb{P}(X_2 \in (b_2, y]) \\ &= \mathbb{P}(X_1 \in (b_1, c_1))\mathbb{P}(X_2 \in (b_2, c_2))\end{aligned}$$

□

3.4. **Discrete Random Variables.**

Def 3.5. *A random variable X is discrete if it takes only countably many values. The distribution of a discrete random variable is determined by a probability mass function (pmf)*

$$\mathbb{P}(X = x_k) = F(x_k) - \lim_{x \to x_k^-} F(x).$$

Below are some common discrete random variables:

- Bernoulli(p) is a random variable with probability p being 1 and $(1-p)$ being 0. Therefore, the pmf is
$$\begin{cases} \mathbb{P}(X=0) = 1-p \\ \mathbb{P}(X=1) = p \end{cases}$$
- Binomial(n,p) $= X_1 + \cdots + X_n$, with each X_i being an independent Bernoulli(p). The pmf is
$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$
- Poisson(λ) for $\lambda > 0$ with (for non-positive interger k)
$$\mathbb{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

We will end this part with a quick proof of the fact that Poisson distribution is the limit of binomial distribution.

Proposition 3.5. *If $X_n \sim$ Binomial(n, p_n) such that $np_n \to \lambda$, then $X_n \to$ Poisson(λ).*

Proof.
$$\mathbb{P}(X_n = k) = \binom{n}{k} p_n^k (1-p_n)^{n-k} = \binom{n}{k} n^{-k} (np_n)^k \left(1 - \frac{np_n}{n}\right)^{n-k}.$$
Now, note that $(n-k) = n\frac{n-k}{n} \to n$ as $n \to \infty$, and $\left(1 - \frac{c_n}{n}\right)^n \to e^{-c}$ when $c_n \to c$, we get
$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{n^k} \frac{1}{k!} (np_n)^k \left(1 - \frac{np_n}{n}\right)^{n-k} = \frac{1}{k!}\lambda^k e^{-\lambda}$$
and thus we're done. □

4. 10/6 Continuous Random Variable, Change of Variables

4.1. **Continuous Random Variable.**

Def 4.1. *A random variable X is continuous if it has a density, a Borel-measurable function $f : \mathbb{R} \to [0, \infty)$, such that for any Borel set $A \in \mathbb{R}$,*

$$\mathbb{P}(X \in A) = \int_A f(x)dx.$$

In this course, we will only take f to be piece-wise continuous, and A to be a countable disjoint union of intervals.

Proposition 4.1. *The density satisfies*

(a) $\int_{-\infty}^{\infty} f(x)dx = 1.$

(b) If f is a density for X and $g(x) = f(x)$ for all x outside of a countable set, then so is $g(x)$.

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

The proof is simple. So let's look at a few examples of density functions.

Example 4.1.

- *The Uniform distribution $U(0,1)$ has density function* $f(x) = \begin{cases} 0 & x \notin (0,1) \\ 1 & x \in (0,1) \end{cases}$, *note that the endpoints does not matter by property (b).*
- *The normal distribution $N(\mu, \sigma^2)$ has density function* $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- *The exponential distribution Exp(λ) has density function* $f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$

Proposition 4.2. *Let F_X be the distribution function X, Let a set $A = \{x_1, x_2 \dots\}$ of isolated points (a set such that all points are separated by a distance larger than a uniform constant ε, or that there's no limiting points). Then*

$$X \text{ has density continuous on } \mathbb{R} - A \iff F \text{ is continuous and is } C^1 \text{ on } \mathbb{R} - A.$$

Before the proof, let's see an example. Consider the Bernoulli$\left(\frac{1}{2}\right)$ distribution, which has distribution function $F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} & 1 \leq x < 2 \\ 1 & x \leq 2 \end{cases}$, then since F is not continuous, we can say that there's no density for X. Indeed, it is not a continuous distribution.

Proof. ($\Rightarrow$): Suppose $f(x)$ is continuous on $\mathbb{R} - A$ and is a density for X.

Now, $F(x) = \int_{-\infty}^{x} f(y)dy$ is continuous by definition of integration (and by property above the density is indeed integrable).

Then, for $x \notin A$,

$$\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h\to 0} \frac{1}{h} \int_{x}^{x+h} f(y)dy = \lim_{h\to 0} \frac{1}{h} f(x) = f(x)$$

where the last equality holds since f is continuous on the integral domain. This computation shows that F is C^1 (indeed since the derivative is just f!) on $\mathbb{R} - A$.

($\Leftarrow$): Suppose F is continuous and is C^1 on $\mathbb{R} - A$. Take $f(x) = F'(x)$ on $\mathbb{R} - A$ we have that for intervals $(a, b] \subset \mathbb{R} - A$,

$$\mathbb{P}(X \in (a, b]) = F(b) - F(a) = \int_{a}^{b} F'(x)dx$$

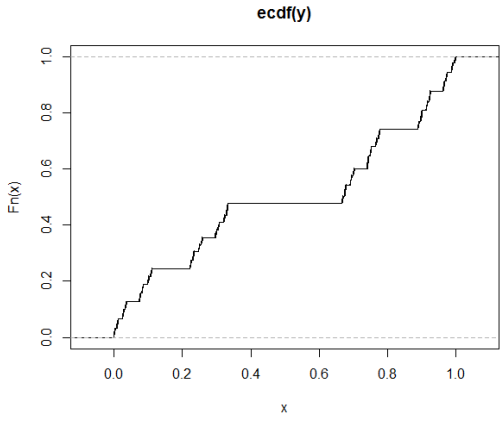
and since it is enough to check for all intervals of the type $(a, b]$ we can show for all Borel sets, we can say that f is a density for X. □

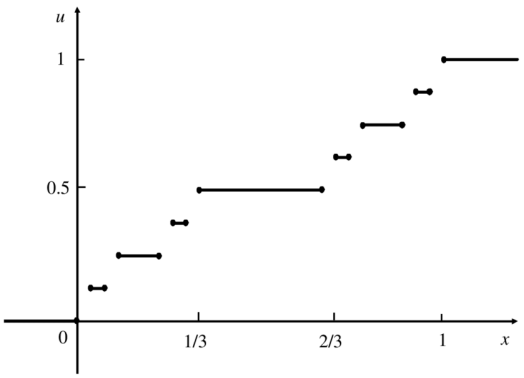
Now we will briefly go over something that we will not cover nor concern in the later part of this course.

Def 4.2. *A Random Variable is singular continuous if it takes values in a set of Lebesgue measure 0, but each point has probability zero.*

Example 4.2. *The Cantor distribution is the distribution of* $X = 2\sum_{n=1}^{\infty} \frac{A_n}{3^n}$ *where* $A_n \sim$ *Bernoulli*$\left(\frac{1}{2}\right)$ *that is defined in the cantor set.*

The cumulative distribution function F of the Cantor distribution is continuous, non-decreasing, but never differentiable on the Cantor set. The graph of F is attached below. As we can see, it really is composed of "lines" of little pieces, yet continuous.





Theorem 4.3. *(Lebesgue decomposition Theorem): For any distribution F, we can decompose F into*

$$F(x) = \alpha_1 F_1(x) + \alpha_2 F_2(x) + \alpha_3 F_3(x)$$

where

- *F_1 is a the distribution of a discrete random variable,*
- *F_2 is a the distribution of a continuous random variable,*
- *F_3 is a the distribution of a singular continuous random variable.*

4.2. **Change of Variables.** If x is a Random Variable, then so is any function $y = g(x)$. Our goal here is to find the density of y given the density of x. We will not cover every case possible (since some pathological case concerns measure theory), but will go over most cases:

Case 1: When g is 1-1, differentiable, and strictly increasing (since 1-1, it is sufficient to only claim for increasing. But this is clearer.), then we have

$$F_y(y) = \mathbb{P}(g(x) \leq y) = \mathbb{P}(x \leq g^{-1}(y)) = F_x(g^{-1}(y)).$$

Further, if X has a continuous density f on $\mathbb{R} - A$ for some A of discrete points (which by proposition 4.2 implies F continuous on A and continuously differentiable everywhere else), then the density of y is defined since

$$f_y(y) = F_y'(y) = f_x(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y).$$

Example 4.3. *For $X \sim$Uniform$\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, $y = \tan(x)$, i.e. the graph of y is a straight line passing through $(0, 1)$ with angle between the vertical axis x. Then*

$$f_y(y) = \frac{1}{\pi} \frac{d}{dy}(\arctan(y)) = \frac{1}{\pi} \cdot \frac{1}{1 + y^2} \sim \textit{Cauchy}$$

Case 2: When g is 1-1, differentiable, and strictly decreasing, we have

$$F_y(y) = \mathbb{P}(g(x) \leq y) = \mathbb{P}(x \geq g^{-1}(y)) = \mathbb{P}(x > g^{-1}(y)) = 1 - F_x(g^{-1}(y))$$

where the third equality is because the measure of a single set is 0. By the same argument as above we have

$$f_y(y) = -f_x(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \geq 0$$

since $\frac{d}{dy} g^{-1}(y) \leq 0$.

Combining the above two cases we can get the result that

$$f_y(y) = f_x(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|$$

when g is strictly monotone.

The general case: Even though we say it's the general case, we want to exclude the case when g either has a horizontal piece, or g is oscillating very badly. We characterize the above condition by

- There is no set A with $\mu(A) > 0$ such that $g(A) = \{y\}$ for some y;
- $\exists$ partition $\mathbb{R} = H_0 \cup \left(\bigcup_{i=1}^{\infty} H_i \right)$ where H_0 is countable and g is 1-1 and differentiable on H_i for $i > 0$.

In this case we have

$$f_y(y) = \sum_{i=1}^{\infty} f_y^{(i)}(y)$$

where

$$f_y^{(i)} = \begin{cases} f_x(g_i^{-1}) \left| \frac{d}{dy} g_i^{-1}(y) \right| & y \in g(H_i) \\ 0 & \text{Otherwise} \end{cases}$$

for $g_i = g|_{H_i}$.

Example 4.4. *For $X \sim N(0,1), Y = X^2$, we let $H_0 = \{0\}, H_1 = (-\infty, 0), H_2 = (0, \infty)$ and thus on H_i, $g_i^{-1}(y) = \pm\sqrt{y}$ accordingly. By the above result we have*

$$f_Y(y) = \phi(\sqrt{y}) \frac{1}{2\sqrt{y}} + \phi(-\sqrt{y}) \frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{\sqrt{y}}$$

where $\phi(x) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}}$.

5. 10/11: Joint distribution, change of variable, and examples.

5.1. Joint distribution.

Def 5.1. *Suppose that $X_1, \dots, X_n$ is defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then,*

- *Their joint distribution is a random vector* $\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$.
- *Their joint distribution function is*
$$F_X(x_1, \dots, x_n) = \mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n).$$
- *The marginal distribution function is*
$$F_{X_1}(x_1) = \lim_{x_2 \to \infty, \dots, x_n \to \infty} F_X(x_1, x_2, \dots, x_n).$$
- *A joint distribution function is continuous if there exists a joint density function*
$$f_X(x_1, \dots, x_n)$$
such that
$$\mathbb{P}(x \in A) = \int_A f_X(x_1, \dots, x_n) dx.$$

Proposition 5.1. *If X is continuous with cdf F and pdf f, then*
$$f_X(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_X(x_1, \dots, x_n).$$

Example 5.1.

- *If $X_1 \sim U[0,1]$, $X_2 \sim U[0,1]$ and they are independent. Then $f_X = 1$ and $F_X(x, y) = x + y$ is a slope plane in the space.*
- *If $X \in \mathbb{R}^d$ with each single random variable $\sim N[0,1]$, then*
$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{||x||_2^2}{2}\right).$$

5.2. Change of variables.

We will use $X \in \mathbb{R}^2$ and $Y = g(X) \in R^2$ in the following parts, i.e.
$$Y = g(X) = \begin{pmatrix} g_1(x_1, x_2) \\ g_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$
as for higher dimension case, the core is the same.

We are concerned about the question that if we have input f_X and g, how can we find $f_Y(y)$.

As we've seen last class, we will need the derivative of g. In the $2d$ case, the derivative is the Jacobian

$$J = \begin{pmatrix} \frac{\partial g_1(x_1,x_2)}{\partial x_1} & \frac{\partial g_1(x_1,x_2)}{\partial x_2} \\ \frac{\partial g_2(x_1,x_2)}{\partial x_1} & \frac{\partial g_2(x_1,x_2)}{\partial x_2} \end{pmatrix}.$$

We will claim that (as a analogy to the 1d case):

Proposition 5.2. *Suppose g is 1-1 and the Jacobian of g is invertible. Then*

$$f_Y(y) = f_X([x]) \frac{1}{|\det J[x]|}\Big|_{g(x)=y}$$

where the expression really means that $[x] = \sum_{y=g(x)} g^{-1}(y)$ when suitable y is discrete and $[x] = \int_{\{y=g(x)\}} g^{-1}(s)ds$ when not. It is true that there are cases when the sum is an infinite sum and the convergence is not guaranteed. However, we will not consider those cases.

Note that in 1-D, the Jacobian is simply the derivative of g.

Remark 5.1. *There is a simple way to remember this formula, especially whether the determinant is on top or bottom.*

Take a (undefined) look at the expectation of $h(Y)$, we have

$$\mathbb{E}(h(Y)) = \int h(y) f_Y(y) dx = \int h(g(x)) f_X(x) dx$$

Then we can easily see that

$$dy = |\det J([x])| dx.$$

5.3. **Examples.**

Example 5.2. *Let $X_1, X_2 \sim U[0,1]$ and are iid, let* $\begin{cases} Y_1 = \sqrt{-2\log X_1}\cos(2\pi X_2) \\ Y_2 = \sqrt{-2\log X_1}\sin(2\pi X_2). \end{cases}$

Then, Y is the normal distribution, and this is a fast way to generate normal distribution on computers.

To verify, we compute the Jacobian first:

$$J = \begin{pmatrix} \frac{1}{2}(-2\log X_1)^{-\frac{1}{2}} \cdot (-2) \cdot \frac{1}{X_1} \cdot \cos(2\pi X_2) & \sqrt{-2\log X_1} \cdot (-2\pi) \cdot \sin(2\pi X_2) \\ \frac{1}{2}(-2\log X_1)^{-\frac{1}{2}} \cdot (-2) \cdot \frac{1}{X_1} \cdot \sin(2\pi X_2) & \sqrt{-2\log X_1} \cdot (2\pi) \cdot \sin(2\pi X_2) \end{pmatrix}$$

and thus

$$\det J = -\frac{2\pi}{X_1}\cos^2(2\pi X_2) - \frac{2\pi}{X_1}\sin^2(2\pi X_2) = -\frac{2\pi}{X_1}$$

and

$$f_{Y_1,Y_2}(y_1, y_2) = f_X(x_1, x_2) \cdot \frac{|X_1|}{2\pi} = \frac{|X_1|}{2\pi}$$

since the density function of X is always 1. Notice that $Y_1^2 + Y_2^2 = -2\log X_1$, we can simply plug back and get the result.

Example 5.3. *Let* $X_1, X_2 \sim U[0,1]$ *and are iid, and let* $\begin{cases} Y_1 = \min(X_1, X_2) \\ Y_2 = \max(X_1, X_2). \end{cases}$

We can of course compute through the formula from here, in which process we will use the $[x]$ as sum of two parts of the function. However, it is also geometrically easy to get the solution: if $y_1 > y_2$ it is impossible, and for each $y_1 \le y_2$ (except when they are equal, but measure 0 there) there are two choices of X, and we add 2 to that situation. It's geometric image is simply to "fold" the unit box into it's upper triangular.

So (as one can check with the formula),

$$f_Y(y_1, y_2) = \begin{cases} 2 & y_1 \le y_2 \\ 0 & y_1 > y_2. \end{cases}$$

Example 5.4. *(Order statistics) Let* $X_1, \dots, X_n \sim U[0,1]$ *and are iid, and let*

$$(Y_1, \dots, Y_n) = (X_{\sigma(1)}, \dots, X_{\sigma(n)}) = g(X)$$

where $X_{\sigma(1)} \le \cdots \le X_{\sigma(n)}$. *This means nothing more than* Y *is a reordering of* X*, from small to large.*

Following the same idea in last example, we get

$$f_Y(y_1, y_2) = \begin{cases} n! & y_1 \le \cdots \le y_n \\ 0 & \text{Otherwise.} \end{cases}$$

Since we are mapping $n!$ kind of permutations of X into 1.

Example 5.5. *(Student's t distribution) Let* $X_1 \sim U[0,1]$ *and* $X_2 \sim \chi_n^2 = Z_1^2 + \cdots + Z_n^2$. *We want to derive for the density of* $Y = \dfrac{X_1}{\sqrt{X_2/n}}$.

Well our method is to first manually add $Y_2 = X_2$ and make a joint distribution out of Y. Then, we can get $\int f_Y(y_1, y_2)dy_2 = f_{Y_1}(y)$ which is its marginal distribution. Note that our choice of Y_2 is because of $f_Y(y_1, y_2) = f_{Y_2}(y_2) \cdot f_{Y_1}(Y_1|Y_2)$, which makes it significantly easier to compute $f_Y(y_1, y_2)$.

Example 5.6. *Let* $X_1, \dots, X_n \sim U[0,1]$ *and are iid, and let*

$$(Y_1, \dots, Y_n) = (X_{\sigma(1)}, \dots, X_{\sigma(n)}) = g(X)$$

where $X_{\sigma(1)} \le \cdots \le X_{\sigma(n)}$, so the same situation as the order statistics (example 5.4).

What's different is we claim that, for $V_i \sim \exp(X_i)$ and partial sum $S_i = \sum_{j=1}^{i} V_i$. Our clain is that $Y = g(X) \overset{d}{=} \left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \dots, \frac{S_n}{S_{n+1}}, \right)$.

One can view V_i as the time we need to wait before the i-th bus came, and S_i is the total time to wait before the $i+1$-th time.

Proof. (claim in example 5.6)

Let $g : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ be

$$g(V_1, \dots, V_{n+1}) = \left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \dots, \frac{S_n}{S_{n+1}}, S_{n+1} \right) := Y$$

to write V in terms of Y, we notice that $\begin{cases} V_1 = y_1 \cdot y_{n+1} \\ V_2 = (y_2 - y_1) \cdot y_{n+1} \\ \vdots \\ V_n = (y_n - y_{n-1}) \cdot y_{n+1} \\ V_{n+1} = y_{n+1} - y_n \cdot y_{n+1} \end{cases}$ which is the same as $V = AY$ whose Jacobian is

$$J = \begin{pmatrix} y_{n+1} & & & & & y_1 \\ -y_{n+1} & y_{n+1} & & & & y_2 - y_1 \\ & -y_{n+1} & y_{n+1} & & & y_3 - y_2 \\ & & & \ddots & & \vdots \\ & & & -y_{n+1} & y_{n+1} & y_n - y_{n-1} \\ & & & & -y_{n+1} & 1 - y_n \end{pmatrix}$$

and to compute the determinant we add each row to the row below and get

$$\det J \overset{\substack{r_2 = r_2 + r_1 \\ r_3 = r_3 + r_2 \\ \vdots \\ r_{n+1} = r_{n+1} + r_n}}{=} \begin{vmatrix} y_{n+1} & & & & & y_1 \\ & y_{n+1} & & & & y_2 \\ & & y_{n+1} & & & y_3 \\ & & & \ddots & & \vdots \\ & & & & y_{n+1} & y_n \\ & & & & & 1 \end{vmatrix} = (y_{n+1})^n.$$

Now we only need to plug in the formula:

$$f_Y(y_1, \dots, y_{n+1}) = f_V(v) |\det J[v]|_{V=AY} = e^{-y_1 - y_2 - \cdots - y_n} (y_{n+1})^n = e^{-y_{n+1}} (y_{n+1})^n$$

which, by integrating over the last term (marginal), we get $n!$ by computation. □

6. 10/13: Expectation, properties, and expected value of transformations.

Def 6.1. *For a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we use*

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega)$$

to be its expectation.

All look good expect that we don't even know whether the integral is well defined. Now there's more common way to define this using the limit of simple functions. Yet there's too much measure theory involved and we'll not take that route. Rather, we verify that it is well defined in the following parts.

Case 1: If $X \geq 0$, then

$$\mathbb{E}[X] = \int_0^\infty X\,dF(X) := \lim_{n\to\infty} \sum_{i=1}^{\infty} \frac{i}{2^n}\left[F\left(\frac{i}{2^n}\right) - F\left(\frac{i-1}{2^n}\right)\right]$$

where we show that the sum is well defined (possibly ∞).

As it is common practice, we explicitly write out the limiting sequence as

$$a_n = \sum_{i=1}^{\infty} \frac{i}{2^n}\left[F\left(\frac{i}{2^n}\right) - F\left(\frac{i-1}{2^n}\right)\right]$$

now since we are in the extended real line, we only want to prove that a_n is decreasing, which will then imply that the limit exists.

To see this, we split cases when i is even and i is odd in the computation below:

$$\begin{aligned} a_n - a_{n+1} = &\sum_{i=1}^{\infty} \frac{1}{2^n}\Bigg\{ i\left[F\left(\frac{i}{2^n}\right) - F\left(\frac{i-1}{2^n}\right)\right] \\ &- \frac{2i}{2}\left[F\left(\frac{i}{2^n}\right) - F\left(\frac{2i-1}{2^{n+1}}\right)\right] - \frac{2i-1}{2}\left[F\left(\frac{2i-1}{2^{n+1}}\right) - F\left(\frac{i-1}{2^n}\right)\right]\Bigg\} \\ = &\sum_{i=1}^{\infty} \frac{1}{2^n+1}\left[F\left(\frac{2i-1}{2^{n+1}}\right) - F\left(\frac{i-1}{2^n}\right)\right] \geq 0 \end{aligned}$$

Case 2: In general case, we can split $X = X^+ - X^-$ and let

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$$

then if at least one of X^+ or X^- is well defined, then X is well defined, and we call X quasi-integrable. If both are well defined then we call it integrable, which also means $\mathbb{E}[|X|] < \infty$.

Let's see some examples on this.

Example 6.1. *Let $\mathbb{P}(X = 2^k) = \frac{1}{2^k}$, then $\mathbb{E}[X^-] = 0$ and $\mathbb{E}[X^+] = \infty$, so X is quasi-integrable.*

$\mathbb{P}(X = 2^k) = \frac{1}{2^{k+1}}$ and $\mathbb{P}(X = -2^k) = \frac{1}{2^{k+1}}$, then both the positive part and the negative part is not finite, so X is not integrable.

Example 6.2. *If X has density function f_X, then*

$$\mathbb{E}[X^+] = \int_0^\infty x f_X(x) dx \quad \textit{and} \quad \mathbb{E}[X^-] = \int_{-\infty}^0 x f_X(x) dx$$

Example 6.3. *If $X \sim Exp(\lambda)$, $\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$*

Also, Cauchy decays like $\frac{1}{x^2}$ so it is not integrable.

Proposition 6.1. *For some set A, let $\mathbb{1}_A$ be the indicator function of the set, then $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$.*

This is usually the definition of expectation in the first place, but just a property here. Anyway what we've written down are all self-consistent.

6.1. **Properties of Expectation.** : We list some properties of expectations here, and give explanation to some of them.

(1) If X, Y are integrable, then so is $X + Y$.
(2) If X, Y quasi-integrable, we cannot say anything about $X + Y$.
    A counter example is when $X = X^+ + X^-$ for X Cauchy, i.e. $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
(3) Linearity: If X is integrable, then cX also is and $\mathbb{E}[cX] = c\mathbb{E}[X]$.
(4) $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$.
(5) If X, Y are independent and both integrable, then $\mathbb{E}[XY]$ is also integrable and $\mathbb{E}[XY] = \mathbb{E}[Y]\mathbb{E}[X]$.

Proof. We only prove in the case when X, Y has density.

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{X\times Y} |x||y| f_X(x) f_Y(y) dx dy \\
&= \int_Y \int_X \left(|x| f_X(x)\right) dx |y| f_Y(y) dy u \\
&= \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}$$

where really we used independent to get the fact that the integral is over a product space. □

6.2. **Expected value of transformations.** :

Proposition 6.2. *If $Y = g(X)$, X has cdf $F_X(x)$, then $\mathbb{E}[Y] = \int g(x) dF(x)$.*

Proof. We prove in the case where X has density and g is 1-1. Of the two conditions the second can be dealt by a similar argument from last week, but there's no way to fix the first on with our current learning.

We just use the result from lat time, i.e. $f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)$,

$$\mathbb{E}[Y] = \int y f_Y(y)dy = \int y f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)dy = \int g(x)f_X(x)dx.$$

□

Proposition 6.3. *For $X \geq 0$, if Q is a quantile function of X, then $\mathbb{E}[X] = \int_0^1 Q(u)du$.*

Proof. The key observation is that $du = \mathbb{1}du$ where $\mathbb{1}$ is uniform on $[0, 1]$. Now since $Q(U) \sim X$ since it is a Quantile function, plugging in proposition 6.2 and we are done. □

Proposition 6.4. *For X, Y, and $X \overset{d}{=} Y$ means that they have the same distribution function. Then*

$$X \overset{d}{=} Y \iff \mathbb{E}[g(X)] = \mathbb{E}[g(Y)]$$

for all bounded and continuous g.

Proof.

($\Rightarrow$): By (a), they have the same cdf.

($\Leftarrow$): What we need to show is that $F_X(z) = F_Y(z)$ for all z, but this is equivalent to $\mathbb{E}[\mathbb{1}\{X \leq z\}] = \mathbb{E}[\mathbb{1}\{X \leq z\}]$.

What is stopping us from just letting $g = \mathbb{1}$? Well, it is not continuous. But since we can easily approximate continuous function with step functions, we are done. This is indeed rigorous since step functions are dense in continuous functions in some domain. □

7. 10/18: Conditional probability, Conditional distribution, Conditional Expectation

7.1. **Conditional probability.**

Remember that we've defined the probability space as a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra. Now, we want to use a sub σ-algebra $\mathcal{G} \subset \mathcal{F}$ to define the conditional probability on it. Some preliminaries first.

For $A \in \mathcal{F}$, we define a measure on $\mathcal{G}$, $\nu(B) = \mathbb{P}(A \cap B)$ for $B \in \mathcal{G}$.

Def 7.1. *We say that a measure ν is absolutely continuous with respect to $\mathbb{P}$, denoted by $\nu \ll \mathbb{P}$, if $\mathbb{P}(B) = 0$, then $\nu(B) = 0$.*

For $A \in \mathcal{F}$, we define a measure on a sub σ-algebra $\mathcal{G}$, $\nu(B) = \mathbb{P}(A \cap B)$ for $B \in \mathcal{G}$. It is absolutely continuous with respect to $\mathbb{P}$ since if $\mathbb{P}(B) = 0$, then

$$\nu(B) = \mathbb{P}(A \cap B) \leq \mathbb{P}(B) = 0.$$

We will also need the following theorem to make our later definitions valid, but our goal here is to make some connection between the above definition and the term $\mathbb{P}(A|\mathcal{G})$, whatever that meant.

Theorem 7.1. *(Radon-Nikodym Theorem) Suppose ν, μ are two finite non-negative measures on the same domain with $\nu \ll \mu$, then, there exists a measurable function denoted by $f : \frac{d\nu}{d\mu}$ such that*

$$\nu(A) = \int_A f \, d\mu$$

that is unique up to μ measure 0 values.

The thing about this theorem is that we don't really have a unique function, so when we define things later we are defining for a class of functions.

Def 7.2. *A conditional probability of A given $\mathcal{G}$ is a random variable*

$$\mathbb{P}(A|\mathcal{G}) : \Omega \to \mathbb{R}$$

that is

(a) $\mathcal{G}$-measurable, integrable,
(b) and $\forall B \in \mathcal{G}$ we have

$$\mathbb{P}(A \cap B) = \nu(B) = \int_B \mathbb{P}(A|\mathcal{G}) d\mathbb{P}. \tag{7.1}$$

(Recall that a function $f : \Omega \to \mathbb{R}$ is $\mathcal{G}$-measurable if $\{\omega | f(\omega) \leq t\} \in \mathcal{G}$ for all t.)

Also, by Radon-Nikodym theorem the conditional probability exists and is unique up to a set of $\mathbb{P}$-measurable 0.

A few examples (especially the extreme ones) may help us understand the above notions.

Example 7.1. *If $\mathcal{G} = \{\emptyset, \Omega\}$, then $\mathbb{P}(A|\mathcal{G})$ is constant since it is $\mathcal{G}$-measurable.*

By (7.1) we compute the value with $B = \Omega$:

$$\mathbb{P}(A) = \int_\Omega \mathbb{P}(A|\mathcal{G}) d\mathbb{P} = \mathbb{P}(A|\mathcal{G}).$$

Example 7.2. *If $\mathcal{G} = \mathcal{F}$, we claim that $\mathbb{P}(A|\mathcal{G}) = \mathbb{1}_A$.*

To see why, we can simply check by (7.1)

$$\int_B \mathbb{1}_A d\mathbb{P} = \mathbb{P}(A \cap B).$$

But what does conditional probability really mean?

Proposition 7.2. *Any conditional probability satisfies the following:*

(a) $\forall A \in \mathcal{F}$, $0 \le \mathbb{P}(A|\mathcal{G}) \le 1$ has probability 1. This saying really makes sense since $\mathbb{P}(A|\mathcal{G})$ is a random variable.

(b) If A_i are countable disjoint sets, then

$$\mathbb{P}\left(\left(\bigcup_i A_i\right)\Big|\mathcal{G}\right) = \sum_i \mathbb{P}(A_i|\mathcal{G})$$

with probability 1.

Proof.

(a):

Since $\int_B \mathbb{P}(A|\mathcal{G}) d\mathbb{P} \ge 0$, we know that $\mathbb{P}(A|\mathcal{G}) > 0$ with probability 1.

If $\mathbb{P}(A|\mathcal{G}) > 1$ in a set of positive measure, then

$$\mathbb{P}(A \cap B) = \int_B \mathbb{P}(A|\mathcal{G}) d\mathbb{P} > \int_B 1 d\mathbb{P} = \mathbb{P}(B)$$

which is impossible.

(b): We can simply check that the RHS of the claim with probability properties.

□

A word of caution is that, suppose $\mathcal{G}$ is generated by disjoint sets $\mathcal{G}_1, \dots, \mathcal{G}_k$ such that

$$\Omega = \mathcal{G}_1 \uplus \cdots \uplus \mathcal{G}_k$$

then for $\mathcal{G}_i$ with $\mathbb{P}(\mathcal{G}_i) = 0$, $\mathbb{P}(A|\mathcal{G})$ can be of any value on $\mathcal{G}_i$.

7.2. **Conditional Distributions.**

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

Def 7.3. *The Conditional distribution of X given $\mathcal{G}$ is a function $q(A, \omega)$, where $A \in \mathbb{R}$ and $\omega \in \Omega$, such that*

- *For fixed ω, $q(\cdot, \omega)$ is a probability distribution on $\mathbb{R}$.*
- *For fixed A, $q(A, \cdot)$ is a conditional probability $\mathbb{P}(X = A|\mathcal{G})$.*

This definition is related to the usual understanding in that:

- For X, Y jointly defined, if $\mathcal{G}$ is generated by $y^{-1}(B)$ for Borel sets B, then, this "conditional distribution of X given Y" is denoted as $\mathbb{P}(X \in A|Y)$.
- If $\mathcal{G}$ is generated by $\{\omega|Y(\omega) = y\}$, then $\mathbb{P}(X \in A|Y = y)$ is the conditional distribution that satisfies

$$\mathbb{P}(X \in A, Y \in B) = \int_B \mathbb{P}(X \in A|Y = y) dF_Y(y)$$

Note that this is really how this definition coincides with the familiar one.

Some applications of this are:

Example 7.3. *If we take $A = (-\infty, t]$, then we get the conditional cdf*

$$F_{X|Y=y}(t) = \mathbb{P}(X \leq t|Y = y).$$

Def 7.4. *The conditional quantile function is*

$$Q_{X|Y=y}(u) = \inf\{t|F_{X|Y=y}(t) \geq u\}$$

which implies

$$Q_{X|Y=y} \overset{d}{=} \{X|Y = y\}.$$

The notion above also extends to a joint equality

$$\left(Q_{X|y=F_Y^-(U')}(U), F_Y^-(U')\right) \overset{d}{=} \{X, Y\}$$

which sometimes we call the Rosenblatt transformation.

7.3. **Conditional Expectations.** Still, we define this notion by properties since the satisfying results are not unique.

Def 7.5. *A random variable $\mathbb{E}[X|\mathcal{G}]$ is a conditional expectation if it satisfies*

(a) $\mathcal{G}$-measurable, integrable,
(b) and for $A \in \mathcal{G}$,

$$\int_A \mathbb{E}[X|\mathcal{G}] d\mathbb{P} = \int_A X d\mathbb{P}.$$

A question that I had learning this is: why not just take $\mathbb{E}[X|\mathcal{G}] = X$? Turns out that this is, in most cases, prohibited by the condition (a).

Maybe let's see some not-so-trivial example:

Example 7.4. *Let* $\Omega = (0, 1]$ *with Lebesgue measure (which functions just like* $U(0, 1]$*),* $\mathcal{F}$ *be the Borel* σ*-algebra, and* $\mathcal{G} = \left\{\emptyset, \Omega, \left(0, \frac{1}{2}\right], \left(\frac{1}{2}, 1\right]\right\}$.

Let's see what kind of function f can satisfy the defintion above and can be labeled $\mathbb{E}[X|\mathcal{G}]$.

First, f has to be $\mathcal{G}$-measurable, which means that the pre-image of any $(-\infty, t] \in \mathcal{G}$, i.e.

$$\{\omega | f(\omega) \leq t\} \in \mathcal{G}$$

which implies that $f\left(\left(0, \frac{1}{2}\right]\right)$ is a constant, and so is $f\left(\left(\frac{1}{2}, 1\right]\right)$.

Then we want to satisfy the second property. Let $A = \left(0, \frac{1}{2}\right]$, we get that

$$\begin{cases} \int_A \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_0^{\frac{1}{2}} f(x)dx = \frac{1}{2}f(w) & \text{for } w \in \left(0, \frac{1}{2}\right]; \\ \int_A x d\mathbb{P} = \int_0^{\frac{1}{2}} X(w)dw \end{cases}$$

which tells us that for $w \in \left(0, \frac{1}{2}\right]$, $\mathbb{E}[X|\mathcal{G}](w) = 2\int_0^{\frac{1}{2}} X(s)ds$.

Similarly, for $w \in \left(\frac{1}{2}, 1\right]$, we have $\mathbb{E}[X|\mathcal{G}](w) = 2\int_{\frac{1}{2}}^1 X(s)ds$.

More specifically, if $\mathcal{G} = \sigma(y)$, which says that $\mathcal{G}$ is generated by the pre-images of Borel sets under $Y(w) = \mathbb{1}_{w \leq \frac{1}{2}}$, then $\mathcal{G} = \left\{\emptyset, \Omega, \left(0, \frac{1}{2}\right], \left(\frac{1}{2}, 1\right]\right\}$ and

$$\mathbb{E}[X|Y](y) = \begin{cases} 2\int_0^{\frac{1}{2}} X(w)dw & y = 1; \\ 2\int_{\frac{1}{2}}^1 X(w)dw & y = 0. \end{cases}$$

Proposition 7.3.

(a) $\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}[X]$,
(b) $\mathbb{E}\left[X \cdot h(y)|Y\right] = h(y)\mathbb{E}[X|Y]$ *if* h *is* Y*-measurable.*

Proof.

(a):

Take $A = \Omega$, then

$$\begin{cases} \int_A \mathbb{E}[X|\mathcal{G}]d\mathbb{P} = \int_\Omega \mathbb{E}[X|Y]d\mathbb{P} = \mathbb{E}\left[\mathbb{E}[X|Y]\right] \\ \int_A x d\mathbb{P} = \int_\Omega x d\mathbb{P} = \mathbb{E}[X] \end{cases}$$

but by (b) in definition 7.5, they are really the same.

(b):

We need to show that $h(y)\mathbb{E}[X|Y]$ has properties of $\mathbb{E}[X \cdot h(y)|Y]$, especially that

$$\int_A h(y)\mathbb{E}[X|Y]d\mathbb{P} = \int_A h(y)xd\mathbb{P}$$

but we will not show this strictly since it's more of a measure-based proof. The stategy to approach this is to first prove for h is indicator functions of $A \in \sigma(Y)$, then we use the simple function convergence method. □

A final word on all these is that what "conditional" really means is that it takes out all randomness unrelated to $\mathcal{G}$ or Y (which is captured by the measurable property) and it needs to to it in a proper way (the other property).

8. 10/20: Variance and Covariance, Multivariate Gaussian, Copulas

8.1. Variance and Covariance.

Def 8.1. *The variance of a random variable X (when exists) is*

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

where the second equation is just by linearity of expectations.

Proposition 8.1. *(Law of total variance) When things are well-defined, we have*

$$\mathrm{Var}(y) = \mathbb{E}[\mathrm{Var}(y|x)] + \mathrm{Var}(\mathbb{E}[y|x])$$

where

$$\mathrm{Var}(y|x) = \mathbb{E}[y^2|x] - \mathbb{E}[y|x]^2.$$

Proof.

$$\begin{aligned}
\mathrm{Var}(y) &= \mathbb{E}[y^2] - \mathbb{E}[y]^2 \\
&= \mathbb{E}[\mathbb{E}[y^2|x]] - \mathbb{E}[\mathbb{E}[y|x]^2] + \mathbb{E}[\mathbb{E}[y|x]^2] - \mathbb{E}[\mathbb{E}[y|x]]^2 \\
&= \mathbb{E}\left[\mathbb{E}[y^2|x] - \mathbb{E}[y|x]^2\right] + \mathrm{Var}(\mathbb{E}[y|x]) \\
&= \mathbb{E}[\mathrm{Var}(y|x)] + \mathrm{Var}(\mathbb{E}[y|x])
\end{aligned}$$

□

Def 8.2. *The covariance of joint distribution X, Y is*

$$\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Note that by simply computing everything we have

- $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$,
- $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$.

Now it happens that we want something to estimate how much two random variable is correlated. Why can't we just use covariance? It's just because we haven't normalize each random variable, so we have:

Def 8.3. *The correlation is*

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \in [-1, 1]$$

where the range is something we will prove later.

Proposition 8.2. *(Stigler's Formula) Suppose that X, Y, XY are all integrable. Then*

$$\mathrm{Cov}(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[F_{X,Y}(t, s) - F_X(t)F_Y(s)\right] dtds.$$

Proof. Let's consider 2 iid copies of random variables (X, Y) and (X', Y'). Then we have

$$\mathbb{E}\big[(X-X')(Y-Y')\big] = \mathbb{E}[XY+X'Y'-XY'-X'Y] = 2\mathbb{E}[XY]-2\mathbb{E}[X]\mathbb{E}[Y] = 2\,\mathrm{Cov}(X,Y)$$

where the second equality is because X and Y' are independent, while X and Y are not.

Now we write things in a fancy way to get:

$$X - X' = \int_{-\infty}^{\infty} \left(\mathbb{1}_{X'\le t} - \mathbb{1}_{X\le t}\right) dt$$

and

$$Y - Y' = \int_{-\infty}^{\infty} \left(\mathbb{1}_{Y'\le t} - \mathbb{1}_{Y\le t}\right) dt.$$

We can do this because for suitable ω,

$$(X - X')(\omega) = \int_0^{(X-X')(\omega)} 1dt = \int_0^{\infty} \mathbb{1}_{\{t\le (X-X')(\omega)\}} dt$$

but the last expression is really nothing but the length of the set $(X'(\omega), X(\omega))$ when $X'(\omega) \le X(\omega)$ and the negative of it when their order is reversed. So since the integral of the indicator function is just the measure of the indicated set, we have

$$(X - X')(\omega) = \int_0^{\infty} \mathbb{1}_{\{X'(\omega)\le t\le X(\omega)\}} dt = \int_{-\infty}^{\infty} \left(\mathbb{1}_{X'(\omega)\le t} - \mathbb{1}_{X(\omega)\le t}\right) dt$$

which justifies our writing above since ω is arbitrary.

So back to the proof, we can then write

$$\begin{aligned}
\mathbb{E}\big[(X - X')(Y - Y')\big] &= \mathbb{E}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\mathbb{1}_{X'\le t} - \mathbb{1}_{X\le t}\right)\left(\mathbb{1}_{Y'\le t} - \mathbb{1}_{Y\le t}\right) dtds\right] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}\left[\left(\mathbb{1}_{X'\le t} - \mathbb{1}_{X\le t}\right)\left(\mathbb{1}_{Y'\le s} - \mathbb{1}_{Y\le s}\right)\right] dtds \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}\left[\mathbb{1}_{X'\le t, Y'\le s} + \mathbb{1}_{X\le t, Y\le s} - \mathbb{1}_{X'\le t, Y\le s} - \mathbb{1}_{X\le t, Y'\le s}\right] dtds \\
&= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \big(F_{X,Y}(t,s) - F_X(t)F_Y(s)\big)\, dtds
\end{aligned}$$

where the second equality is because the integrand is the product of 2 well-behaved functions, and the last equality is due to the fact that $\mathbb{E}[\mathbb{1}_{X\in A}] = \int_A 1d\mathbb{P} = \mathbb{P}(X \in A)$. □

8.2. **Multivariate Gaussian.**

Def 8.4. *For $\mu \in \mathbb{R}^p$ and positive semi-definite non-singular $\Sigma \in \mathbb{R}^{p\times p}$, the multivariate Gaussian $X \sim N(\mu, \Sigma)$ is the continuous distribution with density*

$$P_X(x) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right).$$

There are some properties of this random variable:

- $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$.
- The marginal distribution of any k coordinates is also multivariate Gaussian with mean and covariance given as subsets (submatrices and subterms) of (μ, Σ) with the corresponding indexes.
- One way to sample from $N(\mu, \Sigma)$ is by the following:

$$X \overset{d}{=} \mu + \Sigma^{\frac{1}{2}}\eta$$

where $\eta \sim N(0, I)$. Since Σ is positive definite, its square root is well-defined.

8.3. **Copulas.**

We introduce this concept of Copulas as a tool to measure the dependence of two random variables. It tells us more about the dependence of two random variables then the correspondence since the latter is just a number.

Def 8.5. *A copula is a distribution function of a joint random variable (U, V) such that $C(U, V)$ (the copula) satisfies $C(u, 1) = u$ and $C(1, v) = v$ on $[0, 1]$, i.e. the marginal distribution of u and v is uniform on $[0, 1]$. Note that this condition is really is the normalization of two random variables.*

Well, this definition do seem a little off-topic since how the hell does a out of thin air distribution measures correspondence? To see why we see the following procedure to generate a copula from two known random variables.

General Procedure:

Step 1: We start with any jointly defined (X, Y),

Step 2: Let $(U, V) := (F_X(x), F_Y(y))$.

Then, the joint distribution function of U, V is a copula.

So we see what kind of colupas are related.

Def 8.6. *The Gaussian copula is the result of the above procedure if $(X, Y) \sim N(0, \Sigma_{2x2})$.*

More concretely, it is given by $C_\Sigma(u, v) = \Phi_\Sigma\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$.

Now, let's see why this measures correlation. For some extreme correlation vase we have:

- If U, V are independent, then

$$C(u, v) = \mathbb{P}(U \leq u, V \leq v) = uv.$$

- If $U = V$, or that they are related as good as it can be,

$$C(u, v) = \mathbb{P}(U \leq u, V \leq v) = \min\{u, v\}.$$

- If $U = 1 - V$, or that they are as negatively related as it can be,

$$C(u, v) = \mathbb{P}(U \leq u, 1 - U \leq v) = \begin{cases} 0 & v \leq 1 - u \\ u + v - 1 & \text{otherwise.} \end{cases}$$

And the range of correlation is indeed measured by what one will guess:

Theorem 8.3. *(Fréchet-Hoeffding) Let $W(u, v)$ be the copula when $u = 1 - v$ and $M(u, v)$ be when $u = v$. Then, for any copula $C(u, v)$ we have*

$$W(u, v) \leq C(u, v) \leq M(u, v).$$

Let's see a little teaser in the end. It says that copulas are indeed as common as you might think they are.

Theorem 8.4. *(Sklar, 1956) Suppose X, Y have joint distribution F and marginal distribution $X \sim H$, $Y \sim G$. Then, there always exists some copula*

$$F(x, y) = C(H(x), G(y)).$$

In particular, if H, G are strictly increasing, then C is unique and

$$C(u, v) = F(H^{-1}(u), G^{-1}(u)).$$

9. 10/27: Moment inequalities

Today is going to be quite chill. We will only look at a lot of inequalities that we will use in the later half of the course. Many of them occur in different contexts, so one might already have some familiarity with them.

9.1. **Markov and Chebychev.**

Theorem 9.1. *(Markov's inequality) Let $X \geq 0$. Then $\forall c > 0$ we have*

$$\mathbb{P}(X \geq c) \leq \frac{1}{c}\mathbb{E}[X].$$

It might look daunting at first since it's on general random variables. But just because of that it must be trivial...there's always a balance between generality and profoundness. What it really says, intuitively, is that if X has a lot of large values, then the expectation of X will be large, which is as expected.

Proof.

$$\mathbb{E}[X] \geq \mathbb{E}[\min\{X, c\}]$$

but $\min\{X, c\} = \begin{cases} c & X \geq c \\ x & \text{otherwise} \end{cases}$ so really

$$\mathbb{E}[X] \geq \mathbb{E}[\min\{X, c\}] = \mathbb{P}(X \geq c) \cdot c + p \geq \mathbb{P}(X \geq c) \cdot c$$

where $p \geq 0$ since it's just the expectation of a part of a non-negative random variable. □

And as we can see from the proof the idea is that something smaller than a part of the contribution to the expectation is smaller than the whole expectation. But it is indeed very useful, as we'll see below.

Theorem 9.2. *(Chebyshev's inequality) If* $\mathrm{Var}(X)$ *exists, the for any $c > 0$*

$$\mathbb{P}\big(\,|X - \mathbb{E}[X]| \geq c\big) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

Proof. We simply plug in Markov's inequality.

Let $Y = (X - \mathbb{E}[X])^2$, then $\mathbb{E}[Y] = \mathrm{Var}(X)$ and $Y \geq 0$, $\tilde{c} = c^2$. Then, by Markov's we get

$$\mathbb{P}\big(\,|X - \mathbb{E}[X]| \geq c\big) = \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq \tilde{c}\right) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\tilde{c}} = \frac{\mathrm{Var}(X)}{c^2}.$$

□

Theorem 9.3. *(Chernoff) For $t > 0, c > 0$, we have*

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tc}}.$$

Proof. Let $Y = e^{tX}$, $\tilde{c} = e^{tc}$. By Markov's we get

$$\mathbb{P}(X \geq c) = \mathbb{P}(e^{tX} \geq e^{tc}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{tc}}.$$

□

9.2. **Jensen's Inequality.**

Def 9.1. *A function $f : I \to \mathbb{R}$ is convex if for $0 \leq \alpha \leq 1$*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

which, in the case of f'' exists, is the same as $f'' \geq 0$.

Theorem 9.4. *(Jensen's Inequality) If $f : I \to \mathbb{R}$ is convex, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

where as if f is concave down(flipping the inequality in def of convexity), we have

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)].$$

Proof. We can always find this line that is always below or equal to f, and coincides with f at a particular point x_0. This can be done by using the line connected with a point on right of x_0, call it x, and letting $x \to x_0^+$. After this construction we can say that $\exists\beta$ (the slope) such that

$$f(x) \geq f(x_0) + \beta(x - x_0).$$

Since $f(x_0)$ is fixed and x is changing, for any random variable we can choose $x_0 = \mathbb{E}[X]$ and $x = X$. Then plugging in we have

$$f(X) \geq f(\mathbb{E}[X]) + \beta(X - \mathbb{E}[X])$$

and by taking expectation on both sides

$$\mathbb{E}[f(X)] \geq \mathbb{E}[f(\mathbb{E}[X])] + \beta(\mathbb{E}[X] - \mathbb{E}[X]) = f(\mathbb{E}[X]).$$

□

There are a lot of applications of Jensen's inequality, we list a few here. They all source from this version of the inequality:

$$f\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{f(x_1) + \cdots + f(x_n)}{n}$$

- For $f(x) = e^x$:

$$\exp\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{e^{x_1} + \cdots + e^{x_n}}{n}.$$

- For $f(x) = \frac{1}{x}$:

$$\frac{n}{x_1 + \cdots + x_n} \leq \frac{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}{n} \iff \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \leq \frac{x_1 + \cdots + x_n}{n}.$$

- For $f(x) = x^2$:

$$\left(\frac{x_1 + \cdots + x_n}{n}\right)^2 \leq \frac{x_1^2 + \cdots + x_n^2}{n}.$$

9.3. **Hölder's inequality.**

Theorem 9.5. *(Hölder's inequality) For $a, b \geq 1$ such that $\frac{1}{a} + \frac{1}{b} = 1$, we have*

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^a]^{\frac{1}{a}} \cdot \mathbb{E}[|Y|^b]^{\frac{1}{b}}.$$

One thing to note is that when the right hand side exists, the left hand side exists. This gives a way to see the integrability of a joint random variable only based on the two random variables.

Another thing to notice is when $a = b = 2$, it is exactly the Cauchy-Schwartz inequality:

$$\mathbb{E}[|XY|]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

To prove it we need the following:

Lemma 9.6. *(Young's inequality) If $a, b > 0$, $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$, then*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

(I think this use of characters (uses a, b for totally different things in the above two statements) is like shit. Prof Sun probably agrees, but he realized it too late.)

Proof. (of Young's Inequality) Let $f(x) = \log x$. Then one way to do it is just by concavity of f. Our way below is essentially the same, but wrapped up with Jensen's.

Let

$$X = \begin{cases} a^p & \text{with probability} \frac{1}{p} \\ b^q & \text{with probability} \frac{1}{q} \end{cases}$$

then by Jensen's, since log is concave down we have

$$\begin{cases} f(\mathbb{E}[X]) = \log(\frac{a^p}{p} + \frac{b^q}{q}) \\ \mathbb{E}[f(X)] = \frac{1}{p}\log(a^p) + \frac{1}{q}\log(b^q) = \log(ab) \end{cases}$$

$$\Rightarrow \log(\frac{a^p}{p} + \frac{b^q}{q}) \geq \log(ab) \Rightarrow ab \quad \leq \quad \frac{a^p}{p} + \frac{b^q}{q}.$$

□

Now, the p, q becomes a, b in the following proof.

Proof. (Hölder's)

Try 1: Just using Young's inequality we have

$$|XY| \leq \frac{|X|^a}{a} + \frac{|Y|^b}{b}$$

and taking the expectation we get

$$\mathbb{E}[|XY|] \leq \frac{1}{a}\mathbb{E}[|X|^a] + \frac{1}{b}\mathbb{E}[|X|^b]$$

so not exactly correct. But at least we get something the form we want on the right.

Try 2: let our new a, b(as in the lemma) be $\frac{|X|}{\mathbb{E}[|X|^a]^{\frac{1}{a}}}, \frac{|Y|}{\mathbb{E}[|Y|^b]^{\frac{1}{b}}}$, then we have

$$\frac{|XY|}{\mathbb{E}[|X|^a]^{\frac{1}{a}} E[|Y|^b]^{\frac{1}{b}}} \leq \frac{|X|^a}{a\mathbb{E}[|X|^a]} + \frac{|Y|^b}{bE[|Y|^b]}$$

and this time by taking the derivative we get

$$\frac{\mathbb{E}[|XY|]}{\mathbb{E}[|X|^a]^{\frac{1}{a}} E[|Y|^b]^{\frac{1}{b}}} \leq \frac{1}{a} + \frac{1}{b} = 1$$

and we are done. □

Finally, let's give some comments on the right hand side of Hölder's.

Def 9.2. *The p-mean of X is*

$$||x||_p = \mathbb{E}[|X|^p]^{\frac{1}{p}}.$$

Proposition 9.7. *$||X||_p$ is non-decreasing in p.*

Proof. Given $p < q$, we need to show $||X||_p \leq ||X||_q$, but we know

$$||X||_p \leq ||X||_q \iff \mathbb{E}[|X|^p]^{\frac{1}{p}} \leq \mathbb{E}[|X|^q]^{\frac{1}{q}}$$

$$\iff \mathbb{E}[|X|^p]^{\frac{q}{p}} \leq \mathbb{E}[|X|^q] = \mathbb{E}[(|X|^p)^{\frac{q}{p}}].$$

But now let $f(x) = x^{\frac{q}{p}}$, since $\frac{q}{p} > 1$, f is convex and by Yensen's we get the above inequality, which is equal to what we need. □

10. 11/1: Modes of convergence; Fubini's theorem

10.1. **Modes of convergence.**

Today we talk about some different modes of convergence.

Def 10.1. *If $\{x_n\}_{n\geq 1}$, X are random variables with*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all continuous points x of F, we say that $\{X_n\}$ converges in distribution to X, for which we denote $X_n \Rightarrow X$ or $X_n \xrightarrow{d} X$. We also call this weak convergence.

Def 10.2. *If $\{x_n\}_{n\geq 1}$, X are jointly defined such that for any $\varepsilon > 0$ and*

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

we say that $\{X_n\} \to X$ convergence in probability, denoted as $X_n \xrightarrow{p} X$.

Def 10.3. *If $\{x_n\}_{n\geq 1}$, X are jointly defined such that*

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

we say that $\{X_n\} \to X$ almost surely, denoted as $X_n \xrightarrow{as} X$ or $X_n \to X$ a.s.. We also call this strong convergence.

So if you just write "$X_n \to X$", it doesn't make sense.

Proposition 10.1.

$$X_n \xrightarrow{as} X \underset{*}{\Rightarrow} X_n \xrightarrow{p} X \underset{\dagger}{\Rightarrow} X_n \xrightarrow{d} X$$

Proof. $(*)$: $X_n \xrightarrow{as} X$ means that

$$\mathbb{P}\left(\forall \varepsilon > 0, \exists N s.t. \forall n > N, |X_n - X| < \varepsilon\right) = 1$$

and we can move the universal quantifier outside to get

$$\forall \varepsilon > 0, \mathbb{P}\left(\exists N s.t. \forall n > N, |X_n - X| < \varepsilon\right) = 1$$

which then implies

$$\forall \varepsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = 1$$

which is what we want.

$(\dagger)$: $\forall \varepsilon > 0$,

$$F_n(x) = \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)$$

because in case of $\mathbb{P}(X_n \leq x)$, either $\mathbb{P}(X \leq x + \varepsilon)$ is true or $\mathbb{P}(|X_n - X| > \varepsilon)$ is true. And by plugging back

$$F_n(x) \leq \mathbb{P}(X \leq x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon) = F(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)$$

and thus

$$\limsup_{n\to\infty} F_n(x) \leq F(x + \varepsilon).$$

Now we need to prove the other direction, but the idea is similar. Note that for the same reason as the above, we have

$$\mathbb{P}(X \leq x - \varepsilon) \leq \mathbb{P}(X_n \leq x) + \mathbb{P}(|X - X_n| > 0)$$

which implies

$$F(x - \varepsilon) \leq F_n(x) + \mathbb{P}(|X - X_n| > 0).$$

Taking liminf this time on both sides we have

$$F(x - \varepsilon) \leq \liminf_{n\to\infty} F_n(x).$$

Then, we have

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

which exactly at the continuity points of F, we have the good definition of convergence in distribution. □

Here is another reason why we call the convergence in distribution the "weak convergence".

Proposition 10.2. *We have $X_n \xrightarrow{d} X$ iff $\forall$ continuous bounded function f we have*

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)] \tag{10.1}$$

The idea of this proof is that to view f as "illegally" $f(y) = \mathbb{1}_{y\leq x}$. But this is too not concrete, so we do the proof.

Proof. ($\Leftarrow$): Suppose (10.1) holds, then we define

$$f_t(x) = \begin{cases} 1 & x < 0 \\ 1 - tx & 0 < x < \frac{1}{t} \\ 0 & \frac{1}{t} < x \end{cases}$$

where we have by definition

$$\mathbb{1}_{u\leq x} \leq f_t(u - x) \leq \mathbb{1}_{u\leq x+1/t}$$

$$\Rightarrow F_n(x) = \mathbb{P}(X_n \leq x) \leq \mathbb{E}[f_t(X_n - x)]$$

$$\Rightarrow \limsup_{n\to\infty} F_n(x) \leq \mathbb{E}[f_t(X-x)] \leq F\left(X+\frac{1}{t}\right)$$

by (10.1). So we can now guess that we'll do the opposite direction:

$$F_n(x) \geq \mathbb{E}[f_t(X_n - x + 1/t)]$$

$$\Rightarrow \liminf_{n\to\infty} F_n(x) \geq \mathbb{E}[f_t(X_n - x + 1/t)] \geq F(x - 1/t)$$

$$\Rightarrow F(x-1/t) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F(X+\frac{1}{t})$$

and hence we are done with proving that $X_n \xrightarrow{d} X$.

($\Rightarrow$): Suppose $X_n \xrightarrow{d} X$, the idea to do this is through weighted sum.

For $\varepsilon > 0$, choose $M \geq 0$ such that $M, -M$ are continuity points of $F(x)$ and

$$1 - F(M) + F(-M) = \mathbb{P}(|x| > M) \leq \varepsilon.$$

Moreover, since $F_n(M) \to F(M)$ and $F_n(-M) \to F(-M)$ for large n, we get

$$\mathbb{P}(|X_n| > M) \leq 2\varepsilon$$

with one additional ε comes from the convergence argument.

Now, we deal with the middle part of the integral. Choose m large enough and $-M = x_1 < x_2 < \dots x_{m+1} = M$ such that the following holds:

$$\max_i \sup_{x\in[x_i, x_i+1]} \left|f(x_i) - f(x)\right| < \varepsilon$$

which is always possible since f is uniformly continuous on a compact set.

Now we just mimic the Riemann sum with left endpoints and get this construction: $f_m(x) = f(x_1)$ for $x \in [x_i, x_{i+1})$. We note that f_m is a simple step function with left endpoints.

Then we have

$$\begin{aligned}
\int_{-M}^{M} f_m(x) d F_n(x) &= \sum_{i=1}^{m} f(x_i)\left(F_n(x_{i+1}) - F_n(x_i)\right) \\
&\underset{n\to\infty}{\to} \sum_{i=1}^{m} f(x_i)\left(F(x_{i+1}) - F(x_i)\right) \\
&= \int_{-M}^{M} f_m(x) d F(x)
\end{aligned}$$

which implies

$$\begin{aligned}
&\left|\int_{-M}^{M} f(x)\left(dF(x)-dF_n(x)\right)\right| \\
(\text{trig}) \le &\left|\int_{-M}^{M}\left(f(x)-f_m(x)\right)dF_n(x)\right| + \left|\int_{-M}^{M} f_m(x)\left(dF(x)-dF_n(x)\right)\right| \\
&+\left|\int_{-M}^{M}\left(f(x)-f_m(x)\right)dF(x)\right| \\
=&\varepsilon+\varepsilon+\varepsilon = 3\varepsilon
\end{aligned}$$

where the first and third ε is because of the integrabilty of f on a compact support, and the second is by the inequality above.

Combining all them together, we have

$$\begin{aligned}
&\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)] \\
&\le \left|\int_{-M}^{M} f(x)\left(dF(x)-dF_n(x)\right)\right| + \left|\int_{|x|>M} f(x)dF_n(x)\right| + \left|\int_{-M}^{M} f(x)dF(x)\right| \\
&\le \varepsilon(3+3\sup_x|f(x)|) = c\varepsilon
\end{aligned}$$

and hence we are done for sufficiently large n. □

10.2. **Fubini's theorem.**

We also review some Fubini's theorem and see the probabilistic version of it.

Theorem 10.3. *(Fubini's theorem)*

Let $f : \mathbb{R}^2 \to \mathbb{R}$ such that one of the following holds

- $f(x,y) \ge 0$;
- $\int\left(\int |f(x,y)|dx\right)dy < \infty$;
- $\int\left(\int |f(x,y)|dy\right)dx < \infty$;

Then

$$\iint f(x,y)dxdy = \int\left(\int f(x,y)dx\right)dy = \int\left(\int f(x,y)dy\right)dx.$$

Theorem 10.4. *(Fubini's theorem for summation)*

Let $\{a_{n,m}\}$ be a sequence. If one of the following holds:

- $a_{n,m} \ge 0$;
- $\sum_{n\ge1}\left(\sum_{m\ge1}|a_{n,m}|\right) < \infty$;

- $\sum_{m\geq 1}\left(\sum_{n\geq 1}|a_{n,m}|\right)<\infty$;

Then

$$\sum_{n,m\geq 1} a_{n,m}=\sum_{n\geq 1}\sum_{m\geq 1} a_{n,m}=\sum_{m\geq 1}\sum_{n\geq 1} a_{n,m}.$$

Theorem 10.5. *(Fubini's theorem for Expectation)*

Let X be a random variable and $\{f_n\}$ be functions. If one of the following holds:

- $f_n(X)\geq 0$;
- $\mathbb{E}[\sum_{n=1}^{\infty}|f_n(x)|]<\infty$;
- $\sum_{n=1}^{\infty}\mathbb{E}[|f_n(x)|]<\infty$;

Then

$$\mathbb{E}[\sum_{n=1}^{\infty} f_n(x)]=\sum_{n=1}^{\infty}\mathbb{E}[f_n(x)].$$

11. 11/3: Limits and integrals; Uniform Integrability; Moment generating function

11.1. **Limits and integrals.**

We start by a simple question from analysis: Suppose $f_n(x) \to f(x)$ pointwise, then when is $\int f_n(x)dx \to \int f(x)dx$?

From the simple non-uniform convergence example below we can tell that

$$\lim \int f_n(x)dx \neq \int \lim f_n(x)dx.$$

Example 11.1. $f_n(x) = \begin{cases} n & 0 < x < \frac{1}{n} \\ 0 & otherwise \end{cases}$

Theorem 11.1. *(Dominated convergence theorem) If $f_n \to f$ pointwise, and $|f_n(x)| \leq g(x)$ such that $\int g(x)dx < \infty$ then*

$$\int \lim_{n\to\infty} f_n(x)dx = \lim_{n\to\infty} \int f_n(x)dx$$

and

$$\int \left|f(x) - f_n(x)\right| dx \to 0.$$

Note that the second condition above is a stronger condition.

Theorem 11.2. *(Fatou's Lemma) Suppose $f_n(x) \geq l(x)$ with $\int |l(x)|dx < \infty$, then*

$$\int \liminf_{n\to\infty} f_n(x)dx \leq \liminf_{n\to\infty} \int f_n(x)dx.$$

Note here that the lim inf already guarantees integrabily on the positive part. So we should really view l as a lower bound.

In particular, if $f_n(x) \to f(x)$, then

$$\int f(x)dx \leq \liminf \int f_n(x)dx.$$

Now we see some generalizations that we will use.

Theorem 11.3. *(Dominated convergence for random variables) Let $\{X_n\}$ be random variables on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

- *$X_n(\omega) \to X(\omega)$ for $\omega \in \Omega$;*
- *$|X_n(\omega)| \leq y(\omega)$ for y integrable.*

Then, $\mathbb{E}[X_n] \to \mathbb{E}[X]$ and $\mathbb{E}[|X_n - X|] \to 0$.

Theorem 11.4. *(Fatou's Lemma for Random Variables) If X_n are non-negative random variables, then (i.e. $l = 0$)*

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n].$$

11.2. **Uniform Integrability.**

Def 11.1. *A sequence $\{X_n\}$ of random variables is uniformly integrable if*

(a) $\sup_n \mathbb{E}[|X_n|] < \infty$*;*
(b) $\sup_n \mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n}] \to 0$ *as* $\mathbb{P}(A_n) \to 0$.

Remark 11.1. *The (b) in definition above is somewhat not defined. It really means that* $\forall \varepsilon > 0, \exists \delta > 0$ *such that if* $\mathbb{P}(A_n) < \delta$,

$$\sup_n \mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n}] \leq \varepsilon.$$

Also, not that the example 11.1 is not uniformly integrable.

Proposition 11.5. $\{X_n\}$ *is uniformly integrable iff*

$$\lim_{a\to\infty} \sup_n \int_{|X_n|>a} |X_n| d\mathbb{P} = 0 \tag{11.1}$$

Proof. We first notice that the expression $\int_{|X_n|>a} |X_n| d\mathbb{P}$ is nothing but $\mathbb{E}[|X_n| \cdot \mathbb{1}_A]$, so that gives us some grab onto the question. Now we start to do things rigorously.

$(\Rightarrow)$:

If $\{X_n\}$ is uniformly integrable, then by condition (a) we have

$$\sup_n \mathbb{E}[|X_n|] < c.$$

Now, by Markov's inequality,

$$\mathbb{P}\left(|X_n| > a\right) \leq \frac{\mathbb{E}[|X_n|]}{a} < \frac{c}{a}.$$

Also, by (b) we get that $\forall \varepsilon > 0$, we can choose δ so that if $\mathbb{P}(A) < \delta$, then

$$\sup_n \mathbb{E}[|X_n| \cdot \mathbb{1}_A] \leq \varepsilon.$$

Now we can choose a large enough such that $\frac{c}{a} \leq \delta$ and $A_n = \{|X_n| > a\}$. Then

$$\sup_n \mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n}] \leq \varepsilon$$

but this is exactly what we want.

$(\Leftarrow)$:

Check (a):

$$\mathbb{E}[|X_n|] = \mathbb{E}\left[|X_n| \cdot \mathbb{1}_{|X_n| \leq a}\right] + \mathbb{E}\left[|X_n| \cdot \mathbb{1}_{|X_n| > a}\right] \leq a + \mathbb{E}[|X_n| \cdot \mathbb{1}_{|X_n| > a}]$$

by (11.1) we know $\forall \varepsilon > 0, \exists a$ large such that

$$\mathbb{E}[|X_n|] \leq a + \varepsilon.$$

Check (b):

For $\forall A_n$ with $\mathbb{P}(A_n) \leq \delta$,

$$\begin{aligned}\mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n}] &= \mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n} \cdot \mathbb{1}_{|X_n| > a}] + \mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n} \cdot \mathbb{1}_{|X_n| \leq a}] \\ &\leq \mathbb{E}[|X_n| \cdot \mathbb{1}_{|X_n| > a}] + a \cdot \mathbb{P}(A_n)\end{aligned}$$

Hence, for $\varepsilon > 0$, we can choose a large enough such that $\mathbb{E}[|X_n| \cdot \mathbb{1}_{|X_n| > a}] \leq \frac{\varepsilon}{2}$ by (11.1) and we can choose δ small enough such that $a \cdot \delta \leq \frac{1}{2}\varepsilon$.

This, combined with the above observation yields

$$\mathbb{E}[|X_n| \cdot \mathbb{1}_{A_n}] \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

□

11.3. **Moment generating function.**

Def 11.2. *If X is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$, then the moment generating function is*

$$M_X(t) = \mathbb{E}[e^{tX}]$$

where e^{tX} is quasi-integral (since positive) and

$$e^{tX} = 1 + tX + \frac{t^2}{2}X^2 + \dots$$

With this definition, we can say that we are particular in the question of when does

$$M_X(t) = 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \dots$$

holds.

Theorem 11.6. *If $M_X(t)$ is finite on $|t| < \delta$, then $\mathbb{E}[X^n]$ is finite for all n and we can evaluate the moment generating function term-wise, i.e.*

$$M_X(t) = 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \dots$$

and

$$M_X^{(n)}(t) = \mathbb{E}[X^n].$$

Proof. For $|t| \leq \delta$,

$$\mathbb{E}[X^n] \leq \mathbb{E}[|X|^n] \leq \frac{n!}{|t|^n}\mathbb{E}[e^{t|X|}]$$

and

$$\mathbb{E}[e^{t|X|}] \leq \mathbb{E}[e^{t|X|}] + \mathbb{E}[e^{-t|X|}] < \infty.$$

Now, let the partial sum

$$y_n = \sum_{i=0}^{n} \frac{t^i x^i}{i!}; y_n \to e^{tX}$$

then $|y_n| \leq e^{tX}$ is integrable. By dominated convergence theorem,

$$\mathbb{E}[y_n] \to \mathbb{E}[e^{tX}]$$

and what's left is simple. □

Proposition 11.7. *Moment generating function satisfies*

(a) $M(0) = 1$*;*
(b) $M_{(a+bX)}(t) = e^{ta}M_X(bt)$*;*
(c) If X, Y *are independent, then*

$$M_{X+Y}(t) = \mathbb{E}[E^{tX+tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t)M_Y(t).$$

At last, let's just look at some examples:

Example 11.2.

- For $X \sim N(0,1)$,

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-t)^2}e^{\frac{1}{2}t^2}dx = e^{\frac{1}{2}t^2}$$

- For $X \sim \exp(\lambda)$:

$$\mathbb{E}[e^{tX}] = \int_0^{\infty} e^{tx}e^{-\lambda x}\lambda dx = \begin{cases} \infty & t \geq \lambda \\ \frac{\lambda}{\lambda - t} & t < \lambda. \end{cases}$$

- For $X \sim$ Cauchy:

$$\mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx}\frac{1}{\pi}\frac{1}{1+x^2}dx = \begin{cases} 1 & t = 0 \\ \infty & \text{otherwise.} \end{cases}$$

12. 11/8: Law of Large Numbers with extra conditions; Cumulants

Today we prove the strong and weak law of large numbers with extra conditions.

12.1. **Weak LLN.**

Def 12.1. *For iid random variables $X_1, X_2, \dots, X_n$ the sample mean $\overline{X}_n$ is*

$$\overline{X}_n = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right).$$

The question we ask and will answer is how close is $\overline{X}_n$ to $\mathbb{E}[X_i]$.

Theorem 12.1. *(Weak LLN) If $\mathbb{E}[|X_1|] < \infty$ and $\mathrm{Var}(X_i) < \infty$, then*

$$\overline{X}_n \xrightarrow{p} \mathbb{E}[|X_1|].$$

Proof. We want to show that $\forall \varepsilon > 0$,

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X_1]| > \varepsilon\right) \to 0$$

as $n \to 0$. To this end we first use linearity of expectation

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\left(\mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]\right) = \mathbb{E}[X_n]$$

and then use Chebyshev's inequality to get

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X_1]| > \varepsilon\right) = \mathbb{P}\left(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| > \varepsilon\right) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2}$$

which means we only need to show that $\mathrm{Var}(\overline{X}_n) \to 0$ as $n \to \infty$. But this is because

$$\mathrm{Var}(\overline{X}_n) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n) = \frac{1}{n^2} n \,\mathrm{Var}(x_1) = \frac{\mathrm{Var}(X_1)}{n} \to 0$$

and hence

$$\frac{\mathrm{Var}\left(\overline{X}_n\right)}{\varepsilon} = \frac{\mathrm{Var}(X_1)}{n\varepsilon} \to 0$$

which completes the proof. □

This is weak in 2 senses. One is that the conclusion is only convergence in probability, the other is that one of the assumptions, i.e. $\mathrm{Var}(X_1) < \infty$ is actually not needed. We'll come back to the proof after introducing proper tools in the following lectures.

12.2. **Strong LLN.**

Theorem 12.2. *(Strong LLN) Suppose* $\mathbb{E}[|X_i|] < \infty$, $\mathrm{Var}(X_i) < \infty$ *and* $\mathbb{E}[X_i^4] < \infty$, *then*

$$\mathbb{P}\left(\lim \overline{X}_n = \mathbb{E}[X_1]\right) = 1$$

or we can say it converges almost everywhere.

How to deal with this? Well we only have one way to tackle with this kind of things, which is the Borel-Contelli method.

Def 12.2. *Suppose we have events* $A_1, \dots$ *then the limsup is defined as*

$$\limsup\{A_n\} := \{\omega \in \Omega | \omega \text{ lies in infinitely many } A_n\} = \bigcap_{n\geq 1}\left(\bigcup_{m\geq n} A_m\right)$$

The way we will use this notion is through

$$\mathbb{P}\left(\lim_{n\to\infty} y_n = y\right) = 1 \iff \forall \varepsilon > 0,\ \ \mathbb{P}\left(|y_n - y| > \varepsilon \text{ infinitely often }\right) = 0$$

$$\iff \mathbb{P}\left(\limsup\{|y_n - y| > \varepsilon\}\right) = 0$$

and our way to prove something like that is through the following lemma.

Lemma 12.3. *(Borel - Contelli) If* $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, *then* $\mathbb{P}\left(\limsup\{A_n\}\right) = 0$.

Proof. By definition, $\limsup\{A_n\} \subseteq \bigcup_{m\geq n} A_m$

$$\Rightarrow \mathbb{P}\left(\limsup\{A_n\}\right) \leq \mathbb{P}\left(\bigcup_{m>n} A_m\right) \leq \sum_{m=n}^{\infty} \mathbb{P}(A_m).$$

By letting $n \to \infty$ we get $\mathbb{P}\left(\limsup\{A_n\}\right) \to 0$ since the infinite series $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ converges and hence its tail goes to 0. □

We will then prove the Strong version of LLN.

Proof. (of SLLN) We do it by replacing X_i by $X_i' = X_i - \mathbb{E}[X_i]$. This essentially does not change anything, but it simplifies matter later. In other words, we choose $\mathbb{E}[X_i] = 0$.

With the above assumption, what we need to prove is equivalent to

$$\mathbb{P}\left(\lim \overline{X}_n = \mathbb{E}[X_1]\right) = 1 \iff \mathbb{P}\left(\left|\overline{X}_n\right| > \varepsilon \quad \text{infinitely often}\right) = 0$$

for which we can use the Borel-Contelli to convert to proving

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

One way to go is just to try what we did in the proof of WLLN. If we just use Chebyshev, we get

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\overline{X}_n\right| > \varepsilon\right) \leq \sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_i)}{n\varepsilon^2} \sim \sum \frac{1}{n} \to \infty$$

so not very good. :(

So our goal is just to find $O\left(\frac{1}{n^2}\right)$ out of somewhere.

So we modify the Chebyshev's inequality to get this:

$$\mathbb{P}(|\overline{X}_n| > \varepsilon) \leq \frac{\mathbb{E}[|\overline{X_n^4}|]}{\varepsilon^4}$$

whose deduction is simply to use Chebyshev to the random variable $\mathbb{E}[|\overline{X_n^4}|]$.

Now we have

$$\mathbb{E}[\overline{X_n^4}] = \mathbb{E}\left[\frac{1}{n^4}\left(X_1 + \cdots + X_n\right)^4\right] = \frac{1}{n^4}\mathbb{E}\left[\left(X_1 + \cdots + X_n\right)^4\right]$$

and a close examine gives that there are only 2 kinds of terms left in $\mathbb{E}\left[\left(X_1 + \cdots + X_n\right)^4\right]$ due to the fact that $\mathbb{E}[X_i] = 0$:

$$\begin{array}{ccc} X_i^4 & \rightsquigarrow & n\mathbb{E}[X_i^4] \\ X_i^2 X_j^2 & \rightsquigarrow & \binom{4}{2}\binom{n}{2}\mathbb{E}[X_i^2]^2 \end{array}$$

where the rest all are all multiplied by $\mathbb{E}[X_k] = 0$ for some k.

Now the conclusion is:

$$\mathbb{E}[\overline{X_n^4}] \leq \frac{1}{n^4}\left(n\mathbb{E}[X_i^4] + 3n(n-1)\mathbb{E}[X_i^2]^2\right)$$

$$\Rightarrow \mathbb{P}\left(|X_n| > \varepsilon\right) \leq \frac{1}{n^4\varepsilon^4}\left(n\mathbb{E}[X_i^4] + 3n(n-1)\mathbb{E}[X_i^2]^2\right) \leq \frac{c}{n^2\varepsilon}$$

$$\Rightarrow \sum_{n=1}^{\infty} \mathbb{P}\left(\left|\overline{X}_n\right| > \varepsilon\right) < \infty$$

and by Borel Contelli we are done.

□

A final remark on the proof is that everything boils down to finding $\mathbb{P}\left(\left|\overline{X}_n\right| > \varepsilon\right)$.

Again, this is a weak version of the SLLN, since we can take off the constraints on the 4th degree expectation as well.

12.3. **Cumulants.**

Def 12.3. *The Cumulant generating function of a random variable X is*

$$K_X(t) = \log M_x(t) = \log \mathbb{E}[e^{tX}] = \log(1 + t\mathbb{E}[X] + \dots)$$

where we note that since $\log(1+x) \sim x$ *we have* $K_X(t) \sim O(t)$.

Similar to moments, cumulants $K_n(X)$ are the formal coefficients of t in the expression

$$K_X(t) = \sum_{n=1}^{\infty} K_n(t) t^n.$$

Example 12.1. *The best examples are just to see what is exactly the cumulants.*

(1) $K_1(X) = \mathbb{E}[X]$.
(2) $K_2(X) = \mathrm{Var}(X)$.
(3) $K_3(X) = \mathbb{E}[(X - \mathbb{E}[X])^3]$.
(4) $K_4(X) = \mathbb{E}[(X - \mathbb{E}[X])^4] - 3\,\mathrm{Var}(X)^2$.

So we see that cumulants starts to differ from moments beginning from the fourth term, and the difference corresponds to the step in the proof above of SLLN.

Let's end today's lecture with some properties of $K_n(X)$:

(a) for $n \geq 1$, $K_n(X + c) = K_n(X)$;
(b) $K_n(cX) = c^n K_n(X)$;
(c) if $X_1, \dots, x_m$ are independent, then

$$K_n(X_1 + \cdots + X_m) = \sum_{i=1}^{m} K_n(X_i).$$

Proof. We only prove (c) here since the first two are mundane.

And we also only prove for $m = 2$. We can generate to all m since it's finite. For $m = 2$, we have

$$\begin{aligned} K_{X+Y}(t) &= \log \mathbb{E}\left[e^{t(X+Y)}\right] = \log\left(\mathbb{E}\left[e^{tX}\right] \cdot \mathbb{E}\left[e^{tY}\right]\right) \\ &= \log \mathbb{E}\left[e^{tX}\right] + \log \mathbb{E}\left[e^{tY}\right] = K_X(t) + K_Y(t) \end{aligned}$$

□

this gives us an equality of the series. But the cumulants are defined as the coefficients of the series, and hence we are done.

13. 11/10: Character functions

Last time we've proved LLN. To get a stronger version of the two LLNs proven last time, we need to use the tool of character functions of random variables.

13.1. **Definition of character function.**

Def 13.1. *For a random variable X, its characteristic function $\phi_X(t)$ is*

$$\phi_X(t) := \mathbb{E}[e^{itX}].$$

Note that if $t \in \mathbb{R}$, then $\phi_X(t)$ always exists since e^{itX} is bounded, which means integrable.

Proposition 13.1. *The key properties of $\phi_X(t)$ are:*

(a) $X \overset{d}{=} Y$ iff $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$.
(b) $X_n \overset{d}{\to} X$ iff $\phi_{X_n} \to \phi_X(t)$ point wise for $t \in \mathbb{R}$.

Let's look at a few examples before we see the proof of them.

Example 13.1. $X \sim N(0,1)$.

In this case

$$\begin{aligned}\phi_X(t) &= \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{1}{2}(x-it)^2} e^{-\frac{1}{2}t^2} dx \\ &= e^{-\frac{1}{2}t^2} \frac{1}{\sqrt{2\pi}} \int_{\infty-it}^{-\infty-it} e^{-\frac{1}{2}x^2} dx = e^{-\frac{1}{2}t^2}\end{aligned}$$

where the first equality on the second line is through a countour integral with side edges $\to \infty$, and we can flip the upper and lower bound of the integral since the integrand is even in x.

Example 13.2. $X \sim$ *Cauchy.*

By residue theorem we get

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\pi} \frac{1}{1+x^2} dx = \begin{cases} 2\pi i \operatorname{Res}\Big|_i \frac{e^{itx}}{\pi} \frac{1}{1+x^2} = e^{-t} & t < 0 \\ -2\pi i \operatorname{Res}\Big|_{-i} \frac{e^{itx}}{\pi} \frac{1}{1+x^2} = e^{t} & t > 0 \end{cases}$$

where the negative sign is from the direction. Hence our result is

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

Example 13.3. $Y \sim$ *double exponential random variable with density* $f_Y(y) = \frac{1}{2}e^{-|y|}$.

We do it by (implicitly) Fourier inversion:

$$\begin{aligned}\phi_Y(t) &= \int_{-\infty}^{\infty} e^{ity}\frac{1}{2}e^{-|y|}dy = \int_{-\infty}^{0} e^{ity}\frac{1}{2}e^{y}dy + \int_{-\infty}^{0} e^{ity}\frac{1}{2}e^{-y}dy \\ &= \left[\frac{e^{ity+y}}{2(it+1)}\right]\Bigg|_{-\infty}^{0} + \left[\frac{e^{ity-y}}{2(it-1)}\right]\Bigg|_{0}^{\infty} = \frac{1}{1+t^2}.\end{aligned}$$

We say this has something to do with Fourier inversion because really we have

$$\frac{1}{1+t^2} \sim \frac{1}{2}e^{-|y|}$$

by a Fourier inversion. This can be seen from the combination of Example 13.2 and 13.3.

13.2. **Properties of character functions.**

Lemma 13.2. *$\phi_X(t)$ is continuous on $\mathbb{R}$.*

Proof. For $t \in \mathbb{R}$, if $t_n \to t$, we want to show $\phi_X(t) \to \phi_X(t)$, but this is because

$$\lim_{n\to\infty} \phi_X(t_n) = \lim_{n\to\infty} \mathbb{E}[e^{it_nx}] = \mathbb{E}[\lim_{n\to\infty} e^{it_nx}] = \phi_X(t)$$

where the second equality is by dominated convergence theorem since e^{itx} is dominated by any constant > 1. □

Def 13.2. *We define a set of convergence of the moment generating function by*

$$B := \{t | M_X(t) < \infty\}.$$

Def 13.3. *Also, define a supplement function*

$$G_X(z) := \mathbb{E}[e^{zX}]$$

for $z \in \mathbb{C}$.

Note that by our definition

$$\begin{cases} G_X(z) = M_X(z) & z \in \mathbb{R} \\ G_x(z) = G_X(it) = \phi_X(t) & z = it, t \in \mathbb{R}. \end{cases}$$

Lemma 13.3. *$G_X(z)$ is finite on $B^* := \{z | \operatorname{Re}(z) \in B\}$.*

Proof. We have $G_X(z) = \mathbb{E}[e^{zX}]$ is finite if $\mathbb{E}[|e^{zX}|]$ is finite, but that is the same thing as $\operatorname{Re}(z) \in B$. □

Some complex-analysis properties of $G_X(z)$ are:

Proposition 13.4. *If X is a random variable such that B has non-empty interior, then*

(a) $G_X(z)$ is holomorphic in B^*;
(b) for $x \in B^*$, $\mathbb{E}[x^n e^{zX}]$ exists;

(c) $G_X^{(n)}(z) = \mathbb{E}[x^n e^{zX}]$.

They are in general just result from complex analysis, and if the first is proved, the other two are not surprising by a discussion of holomorphic functions in $\mathbb{C}$.

13.3. **A Proof of Proposition 13.1(a).**

Today we use Fourier inversion to prove 13.1 (a). Next time we will use a more probability based method to do so again.

A theorem we use from Fourier analysis is the following:

Theorem 13.5. *If $F_X(x)$ is continuous at a, b, $a < b$, then*

$$F_X(b) - F_X(a) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \phi_X(t) dt$$

where the integral is in fact a principal valued integral.

We first see how this theorem gives us 13.1(a).

Corollary 13.6. $X \overset{d}{=} Y \iff \phi_X(t) = \phi_Y(t)$.

Proof. (of Corollary 13.6)

$(\Rightarrow)$:

This direction uses the Portmanteau Theorem, which almost directly yields the result. Since it wasn't the main focus here I will skip the proof. But really this is the easy direction.

$(\Leftarrow)$: This is direct by theorem 13.5. For any continuity point x of F_X there exists sequence $\{b_n\} \to x$ of continuity points of F_X. Then $F_X(x) = F_Y(x)$ since they are Cadlag and thus they must coincide at every continuous point, but the only points left are actually fixed by Cadlag, so we are done.

□

Proof. (of Theorem 13.5) The proof is easy if the random variables have density, but we need to work a little bit for the case where they don't.

Fix $x \in \mathbb{R}$, and define

$$h(y) := F_X(x + y) - F_X(y)$$

then we only need to do the following

- Show $h(y) \in L^1$.
- Apply Fourier inversion to get the result.

Since the purpose of L^1 is just to apply Fourier inversion, we assume it for now and first see how the result is deduced.

Fourier inversion implies:

$$h(y) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} e^{-ity} \left(\int_{-\infty}^{\infty} h(z) e^{itz} dz \right)$$

holds if h is continuous at y.

Integrating by parts we ahve

$$\begin{aligned}
\int_{-\infty}^{\infty} h(z) e^{itz} dz &= -\frac{1}{it} \int_{-\infty}^{\infty} e^{itz} dh(z) + \left[\frac{e^{itz}}{iz} \right]_{-\infty}^{\infty} \\
&= -\frac{1}{it} \left(\int_{-\infty}^{\infty} e^{itz} d(F(x+z) - F(z)) \right) + \left[\frac{e^{itz}}{iz} \right]_{-\infty}^{\infty} \\
&= -\frac{1}{it} \left(\int_{-\infty}^{\infty} e^{itz} dF(x+z) - \int_{-\infty}^{\infty} e^{itz} dF(z) \right) + 0 \\
&= -\frac{1}{it} \left[e^{-itx} \phi_X(t) - \phi_X(t) \right] = \frac{1 - e^{-itx}}{it} \phi_X(t)
\end{aligned}$$

where the boundary term is 0 by definition of h at ∞, and we convert to ϕ using the definition of ϕ.

Substituting back into the Fourier inversion we have

$$h(y) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-ity} - e^{-it(x+y)}}{it} \phi_X(t) dt$$

from which we can choose $y = a$, $x = b - a$ to get the result.

Now we check $h \in L^1$. For $\forall c < d$ we have

$$\int_c^d |h(y)| dy = \int_c^d h(y) dy = \int_c^d F_X(x+y) dy - \int_c^d F_X(y) dy$$

which is integrating $F_X(y)$ from $c + x$ to $d + x$, then minus c to d, so it is equivalent to integrating d to $d + x$, then minus c to $c + x$:

$$= \int_d^{d+x} F_X(y) dy - \int_c^{c+x} F_X(y) dy \leq x$$

as both integral is less than x. But x is a constant and we are done by letting $c \to -\infty$ and $d \to \infty$.

□

14. 11/15: Smoothing of distribution, uniqueness of moment generating function

14.1. **Smoothing proof of uniqueness of distribution.**

We want to give another proof of the property of characteristic function Corollary 13.6. To do this we notice that if the random variable has a density function, i.e. it is continuous, then it's relatively easy to prove the property.

Lemma 14.1. *If f has density $f_X(x)$, then it is given by*

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)dt.$$

Proof. We simply apply Fourier inversion to

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(t)dx$$

and the result is obvious. We can do this because f_X is L^1. □

But what if we are just given some general random variable without it being continuous? The trick is to blatantly "make it continuous" by adding some perturbation to it. For instance, given random variable X, we will consider

$$X_\sigma = X + \sigma Z$$

for $Z \sim N(0,1)$ that is independent of X. If X is continuous then obviously X_σ, but what about when X is discrete or mixed?

Lemma 14.2. *If $\sigma > 0$, then X_σ is a continuous random variable.*

Proof. Let σZ have density $f_\sigma(w)$, then there exists bounded g such that

$$\begin{aligned}
\mathbb{E}[g(X+\sigma Z)] &= \int_{\mathbb{R}^2} g(x+\sigma w)dF_{X,\sigma Z}(x,w) \\
\text{(Fubini)} \quad &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x+w)dF_{\sigma Z}dF_X \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x+w)f_\sigma(w)dw\right]dF_X \\
&= \mathbb{E}\left[\int_{-\infty}^{\infty} g(X+w)f_\sigma(w)dw\right]
\end{aligned}$$

where I feel like I'd need to justify the use of Fubini's theorem (not in class).

A version of Fubini's not introduced in class is this:

Theorem 14.3. *(Fubini) Suppose $g(x, y)$ is integrable with respect to a product measure $\pi = \mu \times \nu$ on $M \times N$. Then*

$$\int_{M\times N} g(x,y)d\pi = \int_M \left[\int_N g(x,y)d\nu\right] d\mu = \int_M \left[\int_N g(x,y)d\mu\right] d\nu.$$

So we need to check two things to apply it: that $g(x, w)$ is integrable and that the measure is a product measure. That $g(x, w)$ is integrable is because $X + \sigma Z$ is integrable and g is bounded, so it's less than a multiple of the original integral; that the measure is a product measure is because X and Z are independent, i.e. $F_{X,\sigma Z} = F_x F_{\sigma Z}$.

Continue with the proof. By taking $v = X + w$ we have the following:

$$\begin{aligned}
\mathbb{E}[g(X+\sigma Z)] &= \mathbb{E}\left[\int_{-\infty}^{\infty} g(X+w) f_\sigma(w) dw\right] \\
&= \mathbb{E}\left[\int_{-\infty}^{\infty} g(v) f_\sigma(v-X) dv\right] \\
\text{(Fubini as in form of 10.5)} &= \int_{\infty}^{\infty} \mathbb{E}[f_\sigma(v-X)] g(v) dv
\end{aligned}$$

which means that $\mathbb{E}[f_\sigma(v - X)]$ is a density of X_σ. We really only care about the existence of it, so it doesn't matter if it's complicated or not. □

Why can this even happen? Well, the key intuition is to note that by adding another random variable, we are actually doing a convolution of the distributions, which naturally has a smoothen effect.

Theorem 14.4. *If $\phi_X(t) = \phi_y(t)$ for $t \in \mathbb{R}$, then $X \overset{d}{=} Y$.*

Proof. We prove it in 3 steps.

Step 1: For $\sigma \to 0$, by Lemma 14.2 we know that X_σ and Y_σ has densities.

Step 2: We prove it first for X_σ and Y_σ:

$$\phi_{X_\sigma} = \mathbb{E}[e^{it(x+\sigma Z)}] = \mathbb{E}[e^{itX}] \cdot \mathbb{E}[e^{it\sigma Z}] = \phi_X(t) e^{-\frac{1}{2}t^2\sigma^2}$$

by a similar process the expression for Y is the same. Now we have for any positive σ

$$\phi_{X_\sigma}(t) = \phi_{Y_\sigma}(t) \quad \Rightarrow X_\sigma \quad \text{has the same density as} \quad Y_\sigma$$

due to Lemma 14.1. But this means $X_\sigma \overset{d}{=} Y_\sigma$.

step 3: The idea is just to take $\sigma \to 0$, but we have to justify the step of taking the limit. So we use Proposition 10.2, that for all bounded continuous g we have

$$\lim_{n\to\infty} \mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$$

implies convergence in distribution.

So for such g, $g(x_\sigma) \to g(X)$ and since g bounded the left hand side is dominated by $g(X)$, hence by Dominated Convergence theorem we have

$$\mathbb{E}[g(X_\sigma)] \to \mathbb{E}[g(X)] \quad \text{and} \quad \mathbb{E}[g(Y_\sigma)] \to \mathbb{E}[g(Y)]$$

and since for each σ, $X_\sigma \overset{d}{=} Y_\sigma$, we are done. □

14.2. **Uniqueness of moment generating functions.**

Theorem 14.5. *Suppose $M_X(t) = M_y(t)$ are both finite on $t \in (a, b)$ with $a < 0 < b$, then $X \overset{d}{=} Y$.*

Proof. Recall that $B^* = \{z \in \mathbb{C} | \operatorname{Re}(z) \in (a, b)\}$. We consider

$$G_X(z) = \mathbb{E}[e^{Xz}] \text{ and } G_Y(z) = \mathbb{E}[e^{Yz}]$$

which is introduced a few lectures ago (they are nothing but the moment generating function defined on $\mathbb{C}$). By property of complex exponents we know that G_X and G_Y are convergent on B^* and that they are both holomorphic/analytic.

But from complex analysis for analytic functions if they agree on some limit set, then they agree on the whole complex plane. Since $B^* \subset$ imaginary axis, and that set contains a limiting point, so $G_X(z) = G_Y(z)$ on $\mathbb{C}$, which implies $\phi_X(z) = \phi_Y(z)$ on $\mathbb{R}$. Now apply corollary 13.6 we know $X \overset{d}{=} Y$. □

Now a natural question is to ask that, if all the moments are equal, i.e.

$$\mathbb{E}[X^k] = \mathbb{E}[Y^k]$$

for all k, then do we have $X \overset{d}{=} Y$?

The answer is surprisingly no.

One first notice is that even though the moments are the same, the moment generating function might not even converge. Yet our intuition is that we have for all polynomial P, $\mathbb{E}[P(X)] = \mathbb{E}[P(Y)]$. But we're almost one step from proving $X \overset{d}{=} Y$ since for g bounded,

$$\mathbb{E}[g(X)] = \mathbb{E}[g(Y)] \Rightarrow X \overset{d}{=} Y$$

plus that by Stone-weierstrass theorem says that any continuous function can be approximated by polynomials.

But the fact is that we cannot really apply the theorem and the answer is no. A counter example is the following:

Example 14.1. *$X = e^Z$ where $Z \sim N(0, 1)$. We call X the log normal distribution.*

We can compute that

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

And let Y be that

$$f_Y(y) = \begin{cases} f_X(x)[1 + \sin(2\pi \log x)] & x > 0 \\ 0 & x \leq 0 \end{cases}$$

The fact that Y is a random variable is just easy checking by integration. We claim that X and Y have the same moments, i.e. $\mathbb{E}[X^k] = \mathbb{E}[Y^k]$, but $X \overset{d}{\neq} Y$ since they just have a different density function due to their construction.

We show this by direct computation.

$$\begin{aligned} \mathbb{E}[Y^k] &= \int_0^\infty f_Y(x) dx = \int_0^\infty x^k f_X(x)[1 + \sin(2\pi \log x)] dx \\ &= \mathbb{E}[X^k] + \int_0^\infty x^k f_X(x) \sin(2\pi \log x) dx \end{aligned}$$

so we only need to show that the second integral vanish, which is also by direct computation:

$$\begin{aligned} \int_0^\infty x^k f_X(x) \sin(2\pi \log x) dx &= \int_0^\infty x^k \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2}(\log x)^2} \sin(2\pi \log x) dx \\ (t = \log x, dt = \frac{1}{x} dx) &= \int_{-\infty}^\infty e^{tk} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \sin(2\pi t) dt = 0 \end{aligned}$$

since $\sin(2\pi t)$ is odd and $e^{-\frac{1}{2}t^2}$ is even.

One way to explain this phenomena is the following theorem:

Theorem 14.6. *(Carleman's Theorem) Distribution function is uniquely determined by $\mu_k = \mathbb{E}[X^k]$ if*

$$\sum_{k=1}^\infty \mu_{2k}^{-\frac{1}{2k}} \text{ diverges}$$

i.e. μ_{2k} grows way too fast.

Interpretation: Stone Wiestrass theorem only limits the behavior on a compact set where if the distribution is too irregular outside of that compact set, then it cannot account for that situation.

Theorem 14.7. *Let X_n, X have character functions ϕ_n and ϕ. Then*

$$X_n \xrightarrow{d} X \iff \phi_n(t) \to \phi(t) \text{ point wise.}$$

Proof. ($\Rightarrow$:) We can use proposition 10.2 to prove this, which says that $X_n \xrightarrow{d} X$ iff $\forall$ continuous bounded function f we have $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$. Now we can just pick $f(X) = e^{itX}$ and since $|f(x)| \leq 1$ we get

$$\mathbb{E}[e^{itX_n}] \to \mathbb{E}[e^{itX}]$$

which is exactly what we want.

($\Leftarrow$:) Suppose that $\phi_n(t) \to \phi(t)$ and consider the smooth version of the random variables:

$$X_{n,\sigma} = X_n + \sigma Z, \quad X_\sigma = X + \sigma Z$$

for $Z \sim N(0,1)$. Thus in the back of our mind we know that we will eventually take $n \to \infty$ and $\sigma \to 0$, and it's just a matter of how we do it. And we do it in steps.

Step (a): $X_{n,\sigma} \xrightarrow{d} X_\sigma$

let's think about it. The whole reason of using this smoothening is to use the fact that there is a density function, so we should fix σ first and take $n \to \infty$.

Now the character function of $X_{n,\sigma}$ and X_σ is explicit due to independence, which are

$$\begin{cases} \phi_{Y_{n,\sigma}}(t) = \phi_n(t)e^{-\frac{1}{2}t^2\sigma^2} \\ \phi_{Y_\sigma}(t) = \phi(t)e^{-\frac{1}{2}t^2\sigma^2} \end{cases}$$

with densities

$$f_{n,\sigma}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_n(t)e^{-\frac{1}{2}t^2\sigma^2}dt$$

$$f_\sigma(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)e^{-\frac{1}{2}t^2\sigma^2}dt$$

but now if we fix x and take the limit on n, we will see that for $|t|$ large, $e^{-\frac{1}{2}t^2\sigma^2}$ is dominant and thus the two integrals differs minimally, where as for $|t|$ small, the pointwise convergence of ϕ implies that the 2 integrals differ also minimally in the middle.

Thus, after skipping a standard $\varepsilon - \delta$ argument we can conclude that $f_{n,\sigma}(x) \to f_\sigma(x)$ pointwise.

The next step is to prove

$$\int_{-\infty}^{\infty} |f_{n,\sigma}(x) - f_\sigma(x)|dx \to 0$$

for which we use the dominated convergence theorem: note that

$$|f_{n,\sigma}(x) - f_\sigma(x)| \leq |f_{n,\sigma}(x)| + |f_\sigma(x)| = f_{n,\sigma}(x) + f_\sigma(x)$$

and the equality is due to the fact that all density functions are positive. But then the right hand side is integrable as it's the sum of two integrable functions.

Thus, by dominated convergence theorem

$$\begin{aligned} F_{n,\sigma}(x) - F_{\sigma}(x) = \mathbb{P}(X_{n,\sigma} \leq x) - \mathbb{P}(X \leq x) &= \int_{-\infty}^{x} f_{n,\sigma}(s)ds - \int_{-\infty}^{x} f_{\sigma}(s)ds \\ &\leq \left| \int_{-\infty}^{x} f_{n,\sigma}(s) - f_{\sigma}(s)ds \right| \leq \int_{-\infty}^{x} \left| f_{n,\sigma}(s) - f_{\sigma}(s) \right| ds \\ &\leq \int_{-\infty}^{\infty} \left| f_{n,\sigma}(s) - f_{\sigma}(s) \right| ds \to 0 \end{aligned}$$

which means that the smoothened version has $X_{n,\sigma} \xrightarrow{d} X_{\sigma}$.

Step (b): $X_n \xrightarrow{d} X$

What we want to do here is to take $\sigma \to 0$. Let a be a continuous point of $F_X(x)$, then using the exact same reason as we did in proposition 10.1, we get the following inequality: for $\delta > 0$

$$\mathbb{P}(X_{\sigma} \leq a - \delta) - \mathbb{P}(\sigma Z \geq \delta) \leq \mathbb{P}(X \leq a) \leq \mathbb{P}(X_{\sigma} \leq a + \delta) + \mathbb{P}(\sigma Z \geq \delta) \tag{14.1}$$

and

$$\mathbb{P}(X_{n,\sigma} \leq a - \delta) - \mathbb{P}(\sigma Z \geq \delta) \leq \mathbb{P}(X_n \leq a) \leq \mathbb{P}(X_{n,\sigma} \leq a + \delta) + \mathbb{P}(\sigma Z \geq \delta) \tag{14.2}$$

thus we know

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}(X_n \leq a) &\overset{\text{(by 14.2)}}{\leq} \limsup_{n\to\infty} \mathbb{P}(X_{n,\sigma} \leq a + \delta) + \mathbb{P}(\sigma Z \geq \delta) \\ &= \mathbb{P}(X_n \leq a + \delta) + \mathbb{P}(\sigma Z \geq \delta) \end{aligned}$$

(by 14.1 and taking $a - \delta = a + \delta$) $\leq \mathbb{P}(X \leq a + 2\delta) + 2\mathbb{P}(\sigma Z \geq \delta)$

which means that if we take $\sigma \to 0$ first and $\delta \to 0$ second, we will have

$$\lim_{\delta\to 0} \lim_{\sigma\to 0} \left(\limsup_{n\to\infty} \mathbb{P}(X_n \leq a) \right) \leq \lim_{\delta\to 0} \lim_{\sigma\to 0} \left(\mathbb{P}(X \leq a + 2\delta) + 2\mathbb{P}(\sigma Z \geq \delta) \right) = \mathbb{P}(X \leq a)$$

where we know that the left hand side is just $\limsup_{n\to\infty} \mathbb{P}(X_n \leq a)$ and hence

$$\limsup_{n\to\infty} \mathbb{P}(X_n \leq a) \leq \mathbb{P}(X_n \leq a).$$

Now, by a very similar argument but with use of the other side of the inequalities, we get that

$$\liminf_{n\to\infty} \mathbb{P}(X_n \leq a) \geq \mathbb{P}(X_n \leq a)$$

which implies

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t).$$

□

One remark is that the first inequality in 14.1 and 14.2 is actually quite trivial.

15. 11/17: Law of large numbers and Central Limit theorem

15.1. **Characteristic functions and moments.**

Proposition 15.1. *Suppose X has characteristic function $\phi(t)$ and $\mathbb{E}[X^m]$ exists for $m \geq 1$, then*

$$\phi(t) = 1 + \sum_{i=1}^{m} \frac{(it)^j}{j!}\mathbb{E}[X^j] + o(t^m)$$

and

$$\phi^{(j)}(0) = i^j \mathbb{E}[X^j]$$

for $j \leq m$.

As one can guess, this follows from Taylor..

Proof. By Taylor we have

$$e^{(it)} = \sum_{j=0}^{m} \frac{(it)^j}{j!} + R_m(t) \quad \text{where} \quad |R_m(t)| \leq \min\left\{ \frac{|t|^{m+1}}{(m+1)!}, \frac{2|t|^m}{m!} \right\}$$

and hence we get that

$$\phi(t) = \sum_{j=0}^{m} \frac{(it)^j}{j!}\mathbb{E}[X^j] + \mathbb{E}[R_m(tX)]$$

which means that it is sufficient to show $\mathbb{E}[|R_m(tX)|] = o(t)$, in other words, we can show that

$$\lim_{t\to 0} \mathbb{E}\left[\frac{|R_m(tX)|}{|t|^m}\right] = 0.$$

We know that

$$\frac{|R_m(tX)|}{|t|^m} \leq \frac{2|t|^m|x|^m}{m!|t|^m} = \frac{2|x|^m}{m!}$$

which is integrable since we require that the moments exist. Therefore, we can use the Dominated convergence theorem to get

$$\lim_{t\to 0} \mathbb{E}\left[\frac{|R_m(tX)|}{|t|^m}\right] = \mathbb{E}\left[\lim_{t\to 0} \frac{|R_m(tX)|}{|t|^m}\right] \leq \mathbb{E}\left[\lim_{t\to 0} \frac{|t||x|^m}{(m+1)!}\right] \to 0$$

and we are done. □

15.2. **Proof of enhanced LLN.**

Even though we have distinguished weak and strong LLN, the results here is enhanced. See Nov. 8th notes for difference.

Theorem 15.2. *Let $X_1, \ldots, X_n$ be iid random variables with $\mathbb{E}[|x_i|] < \infty$. Then we have*

$$\bar{X}_n : \frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{p} \mathbb{E}[X_i].$$

Proof. Some few observations for this proof first:

(1) It is enough to show convergence in distribution, since convergence in distribution implies convergence in probability when the limiting random variable X is a constant. But that $\mathbb{E}[X_i]$ is nothing but a constant, so we can do this change.
(2) By replacing X_i with $X_i - \mathbb{E}[X_1]$ we can assume $\mathbb{E}[X_i] = 0$.
(3) It suffices to show that $\phi_{\overline{X}_n}(t) \to \phi_0(t) = \mathbb{E}[e^{it0}] = 1$ due to the theorem we've proven twice.

Now let's how this convergence. We can compute that

$$\begin{aligned}\phi_{\overline{X}_n}(t) &= \mathbb{E}\left[e^{it\frac{1}{n}(X_1+\cdots+X_n)}\right] = \mathbb{E}\left[e^{i\frac{t}{n}(X_1+\cdots+X_n)}\right] \\ &= \phi_{X_1+\cdots+X_n}\left(\frac{t}{n}\right) = \phi_{X_1}\left(\frac{t}{n}\right)\cdots\phi_{X_n}\left(\frac{t}{n}\right) = \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n\end{aligned}$$

where we split the characters using independence.

Now, by the proposition above, we can do Taylor to get

$$\phi_{x_1} = 1 + it\mathbb{E}[X_1] + o(t) = 1 + o(t)$$

which implies

$$\left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n = [1 + o\left(\frac{t}{n}\right)]^n = \exp\left(o\left(\frac{t}{n}\right)\cdot n\right) \to 0$$

as $n \to \infty$. □

15.3. **CLT.**

Theorem 15.3. *(Central Limit Theorem): Let X_i be iid random variables with $\mathbb{E}[X_i] = 0$, $\mathrm{Var}(X_i) = 1$. Then we can normalize the sample mean*

$$S_n := \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \xrightarrow{d} N(0,1).$$

As we'll see, this really needs only the same technique.

Proof. We need to show that $\phi_{S_n}(t) \xrightarrow{n\to\infty} e^{-\frac{1}{2}t^2}$ to conclude a convergence in distribution, and we do a very similar expansion as above to get

$$\phi_{S_n}(t) = \mathbb{E}\left[e^{it\frac{1}{\sqrt{n}}(X_1+\cdots+X_n)}\right] = \phi_{X_i}\left(\frac{t}{\sqrt{n}}\right)^n.$$

Now the expansion can be done till the square term and we get

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

which means

$$\phi_{X_i}\left(\frac{t}{\sqrt{n}}\right)^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n = \mathbb{E}[1 + \frac{t^2}{n}\left(-\frac{1}{2} + o(1)\right)]^n$$
$$\xrightarrow{n\to\infty} \exp\left(\frac{t^2}{n}\left[-\frac{1}{2} + o(1)\right] \cdot n\right) \to \exp\left(-\frac{t^2}{2}\right)$$

□

Now we note that the difference between LLN and CLT is just the degree of the denominator, which encourages us to ask what will happen if we take

$$\hat{S}_n = \frac{1}{n^a}\left(X_1 + \cdots + X_n\right)$$

we have

$$\phi_{X_i}\left(\frac{t}{n^a}\right)^n = \left[1 + \frac{t^2}{n^{2a}}\left(-\frac{1}{2} + o(1)\right)\right]^n \to \exp\left(\frac{t^2}{n^{2a}}\left[-\frac{1}{2} + o(1)\right] \cdot n\right) \to \exp\left(-\frac{t^2}{2}n^{1-2a}\right)$$

and which goes to

$$\to \begin{cases} 1 & a > \frac{1}{2} \\ 0 & a < \frac{1}{2} \end{cases}.$$

We now look at an extension of CLT.

Corollary 15.4. *We can get convergence in L^∞:*

$$\sup_x |\mathbb{P}(S_n \le x) - \Phi(x)| \to 0$$

where Φ is the cdf of normal distribution.

Proof. $\forall \varepsilon > 0$, we can choose $x_1 < \cdots < x_n$ such that

$$\Phi(x_1) \le \varepsilon, \quad \Phi(x_{i+1}) - \Phi(x_i) < \varepsilon.$$

Then, for any $x \in [x_i, x_{i+1}]$, we have

$$\mathbb{P}(S_n \le x) - \Phi(x) \;\le\; \mathbb{P}(S_n \le X_{i+1}) - \Phi(x_i) \;\le\; \mathbb{P}(S_n \le X_{i+1}) - \Phi(x_{i+1}) + \varepsilon$$

and similarly

$$\Phi(x) - \mathbb{P}(S_n \le x) \;\le\; \Phi(x_{i+1}) - \mathbb{P}(S_n \le X_i) \;\le\; \Phi(x_i) - \mathbb{P}(S_n \le X_i) + \varepsilon.$$

But simply combining the two inequalities we can get

$$|\Phi(x) - \mathbb{P}(S_n \le x)| \le \max\left\{\left(\Phi(x_i) - \mathbb{P}(S_n \le x_i)\right), \left(\mathbb{P}(S_n \le X_{i+1}) - \Phi(x_{i+1})\right)\right\} + \varepsilon \le 2\varepsilon$$

where the last inequality is due to the fact that $\Phi \to \mathbb{P}$ pointwise by CLT. □

We can continue explore and ask what if we remove the iid condition.

Theorem 15.5. *(Lindeberg-Feller CLT) Suppose X_i are independent, but not identical, with $\mathbb{E}[X_k] = \mu_k$ and $\mathrm{Var}(X_k) = \sigma_k^2$. Let*

$$S_n^2 = \sum_{k=1}^{n} \sigma k^2.$$

If the Lindeberg condition

$$\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{k=1}^{n} \mathbb{E}\left[(X_k - \mu_k)^2 - \mathbb{1}_{|X_k-\mu_k|>\varepsilon S_n}\right] = 0$$

holds, then

$$Z_n = \frac{1}{S_n} \sum_{k=1}^{n} (X_k - \mu_k) \xrightarrow{d} N(0,1).$$

The proof of this theorem is really similar to the above with some extra technical bounds and analysis. More specifically, we start with

$$\phi_{Z_n} = \prod_{i=1}^{n} \left[1 - \frac{t^2\sigma_i^2}{2S_n} + o(t^2)\right]$$

and do some analysis. We don't do it here.

The problem even for now is that we don't know the error's order, but only know the asymptotic trend. We'll see how to some of that next week.

16. Refinements of CLT

16.1. **Barry-Esseen Theorem.**

The following theorem makes CLT non-asymptotic.

Theorem 16.1. *(Barry-Esseen) If X_i are iid random variable with mean 0, variance 1 and finite third moment, then for some $c \leq 0.79$, we have*

$$\sup_x \left| \mathbb{P}\left(\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \leq x \right) - \Phi(x) \right| \leq c \frac{\mathbb{E}[|X_i|^3]}{\sqrt{n}}.$$

We prove a weaker version of this, namely we ignore the finding of the constant c, and that the bound is a little bit weaker.

Proof. The proof uses Stein's method, which uses smoothing and iterative replacement.

Part A: smoothing.

We first define a smooth function

$$g_x(t) = \begin{cases} 0 & t > x + h \\ 1 & t \leq x \\ \text{some intermediate smooth curve in between} & x < t \leq x + h \end{cases}$$

where we don't specify the exact expression. If we do then we can get the bound c, but that's too technical. Now we shift the smooth function using $g_{x-h}(t)$, so that the function is 0 after x and 1 before $x - h$. Note that these functions form a bound for the indicator function:

$$g_{x-h}(t) \leq \mathbb{1}_{t \leq x} \leq g_x(t).$$

Again, we let

$$S_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$$

and we thus have

$$\mathbb{E}[g_{x-h}(S_n)] \leq \mathbb{P}(S_n \leq x) \leq \mathbb{E}[g_x(S_n)]$$

by taking the expectation on all parts.

Now we prove our intermediate goal: for iid Gaussian $Z_1, \ldots, Z_n$ and try to bound the difference

$$\mathbb{E}[g_X(S_n)] - \mathbb{E}\left[g_X\left(\frac{1}{\sqrt{n}}(Z_1 + \cdots + Z_n) \right) \right]$$

How to show this? The trick is to consider the intermediates.

Part B: Let

$$W_k = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_k + Z_{k+1} + \cdots + Z_n)$$

and then we can rewrite

$$\mathbb{E}[g_X(S_n)] - \mathbb{E}\left[g_X\left(\frac{1}{\sqrt{n}}(Z_1 + \cdots + Z_n)\right)\right] = \sum_{k=1}^{n}\left(\mathbb{E}[g_x(W_k)] - \mathbb{E}[g_x(W_{k-1})]\right)$$

So we pick out the common part between W_k and W_{k-1} to get

$$L_k = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_{k-1} + Z_{k+1} + \cdots + Z_n)$$

and thus write

$$W_k = L_k + \frac{X_k}{\sqrt{n}}, \;\; W_{k-1} = L_k + \frac{Z_k}{\sqrt{n}}$$

now we use Taylor to get the order of the difference for each summand

$$\begin{aligned}
&\mathbb{E}\left[g_x\left(L_k + \frac{X_k}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g_x\left(L_k + \frac{Z_k}{\sqrt{n}}\right)\right] \\
=\;& \mathbb{E}\left[g_x(L_k) + \frac{X_k}{\sqrt{n}}g_x'(L_k) + \frac{X_k^2}{2n}g_x''(L_k) + \frac{X_k^3}{6n^{3/2}}g_x'''(A)\right] \\
&+ \mathbb{E}\left[g_x(L_k) + \frac{Z_k}{\sqrt{n}}g_x'(L_k) + \frac{Z_k^2}{2n}g_x''(L_k) + \frac{Z_k^3}{6n^{3/2}}g_x'''(B)\right]
\end{aligned}$$

Here the residue

$$A \in \left[L_k + \frac{1}{\sqrt{n}}X_k\right], B \in \left[L_k + \frac{1}{\sqrt{n}}Z_k\right]$$

But note that many term cancels: the mean and variance are the same and the third moment of normal distribution is 0, so we only need to deal with the residue with A. We can construction g_x such that it's third derivative g_x''' is uniformly bounded, so we can find

$$|g_x'''(t)| \leq \frac{c_3}{h^3}$$

where the denominator comes from the fact that $g_x(t) = \hat{g}_x(t/h)$ and taking 3 derivatives is like finding the third expansion.

Thus, the difference of the above is nothing but

$$\mathbb{E}\left[\frac{X_k^3}{6n^{3/2}}g_x'''(A)\right] \leq \frac{1}{6n^{3/2}}\mathbb{E}[|X_k|^3|g_x'''(A)|] \leq \frac{c_3}{6n^{3/2}h^3}\mathbb{E}[|X_i|^3]$$

and by timing it by n the total difference is

$$\sum_{k=1}^{n}\left(\mathbb{E}[g_x(W_k)] - \mathbb{E}[g_x(W_{k-1})]\right) \leq \frac{c_3}{6h^3}\frac{1}{\sqrt{n}}\mathbb{E}[|X_i|^3]$$

and in conclusion we have

$$\left|\mathbb{E}\left[g_x\left(S_n\right)\right] - \mathbb{E}\left[g_x\left(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\right)\right]\right| \leq \frac{c_3}{6h^3}\frac{1}{\sqrt{n}}\mathbb{E}[|X_i|^3] \tag{16.1}$$

For $N \sim N(0,1)$, by the construction of the smooth function g_x, we have

$$|\Phi(x) - \mathbb{E}[g_x(N)]| \leq Ch$$

and remember we've above shown

$$\mathbb{E}[g_{x-h}(S_n)] \leq \mathbb{P}(S_n \leq x) \leq \mathbb{E}[g_x(S_n)].$$

Note that $\dfrac{Z_1+\cdots+Z_n}{\sqrt{n}} \sim N(0,1)$ since they are scaled down by $\sqrt{n}$, which will result in the same variance. Now by 16.1 we have

$$|\mathbb{E}[g_x(S_n)] - \Phi(x)| \leq Ch + \frac{c_3}{6\sqrt{n}h^3}\mathbb{E}[|X_i|^3]$$

and using the fact that

$$|\Phi(x-h) - \Phi(x)| \leq c_2 h$$

we get

$$\begin{aligned}\left|\mathbb{P}(S_n \leq x) - \Phi(x)]\right| &\leq \left|\mathbb{E}[g_x(S_n)] - \Phi(x)\right| + \left|\mathbb{E}[g_{x-h}(S_n)] - \Phi(x-h)\right| \\ &\leq \frac{2c_3}{6\sqrt{n}h^3}\mathbb{E}[|X_i|^3] + c_1 h\end{aligned}$$

where by taking $h = n^{-1/8}\mathbb{E}[|X_i|^3]^{1/4}$ we get

$$\left|\mathbb{P}(S_n \leq x) - \Phi(x)]\right| \leq \tilde{c}n^{-1/8}\mathbb{E}[|X_i|^3]^{1/4}.$$

So we'll stop here as we already have a bound, just not as good as the one in the more general theorem. □

16.2. **Density and CLT.**

The most obvious problem about investigating a density version of CLT is that S_n might not even have a density. So we want to proof something on that area, i.e. when do we have $f_{S_n} \to f_N$.

For now we only prove a proposition that we'll use.

Proposition 16.2. *If the characteristic function $\phi_X(t)$ satisfies*

$$\int_{-\infty}^{\infty} |\phi_X(t)|dt < \infty$$

then X is a continuous random variable with density

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)dt.$$

We note that the density function is not defined as a principle value integral since by our assumption the function is indeed convergent.

Proof. Remember that for $a < b$ on which F_X is continuous, we had the formula

$$F_X(b) - F_X(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\phi_X(t)dt$$

and note that the nominator can be bounded by

$$|e^{-ita} - e^{-itb}| = |1 - e^{-it(b-a)}| \leq |t| \cdot |b - a|$$

where the last is some trigonometric fact. Hence plugging in we have

$$|F_X(b) - F_X(a)| \leq c \cdot |b - a|.$$

So it looks like that we are only one step away from continuity since we can only do this for continuous points of F_X. But really that requirement is superficial if you think the following way: for any $x < y$, we can find $a < x < y < b$ such that a, b are continuous points of F_X and $|x - y| \geq \frac{1}{2}|a - b|$, which means

$$|F_X(y) - F_X(x)| \leq c \cdot |y - x|.$$

Which implies that F_X is Lipschitz, hence continuous.

Now we just compute the derivative quotient to get that for any x, h

$$\frac{F_X(x+h) - F_X(x)}{h} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-itx} - e^{-it(x+h)}}{ith}\phi_X(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1 - e^{-ith}}{ith}e^{-itx}\phi_X(t)dt$$

which by dominated convergence (if $h \to 0$ the the quotient is bounded, and the whole thing is integrable) we have

$$\begin{aligned}\lim_{h\to 0}\frac{F_X(x+h) - F_X(x)}{h} &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\lim_{h\to 0}\frac{1 - e^{-ith}}{ith}\right)e^{-itx}\phi_X(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)dt.\end{aligned}$$

□

17. CLT AT THE LEVEL OF DENSITY

17.1. **Convergence of Density.**

Theorem 17.1. *If X_n and X are continuous random variable with densities $f_n(x)$, $f(x)$ such that $f_n(x) \to f(x)$ for almost any x, then*

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} |f_n(x) - f(x)| dx = 0$$

and also

$$\lim_{n\to\infty} \sup_{\mathcal{B}-Borel} \left|\mathbb{P}(X_n \in B - \mathbb{P}(X \in B))\right|.$$

The sup in the second equation means that this convergence is at uniform rate, and this is sometimes called "X_n converges strongly to X." Now one might think that the first statement is obvious where as the second a little taunting, but in fact we'll see that the second is an easy corollary of the first.

Proof. Note that

$$|f_n(x) - f(x)| \leq |f_n(x)| + |f(x)|,$$

then by Fatou's lemma we have (with $l = 0$)

$$\begin{aligned}
&\liminf_n \int \left[|f_n(x)| + |f(x)| - |f_n(x) - f(x)|\right] dx \\
\geq &\int \liminf_n \left[|f_n(x)| + |f(x)| - |f_n(x) - f(x)|\right] dx \\
= &\int 2f(x)dx = 2
\end{aligned}$$

On the other hand we have

$$\liminf_n \int \left[|f_n(x)| + |f(x)| - |f_n(x) - f(x)|\right] dx = 2 - \limsup_n \int |f_n(x) - f(x)| dx$$

which, combined with the above gives us

$$\begin{aligned}
&2 - \limsup_n \int |f_n(x) - f(x)| dx \geq 2 \\
\Rightarrow &0 \geq \limsup_n \int |f_n(x) - f(x)| dx \geq 0 \\
\Rightarrow &\limsup_n \int |f_n(x) - f(x)| dx = 0
\end{aligned}$$

which is our first statement. Now for the strong convergence statement, we know that $\forall \mathcal{B}$ Borel

$$|\mathbb{P}(X_n \in \mathcal{B}) - \mathbb{P}(X \in \mathcal{B})| = \left|\int_{\mathcal{B}} [f_n(x) - f(x)]\right| \leq \int_{\mathcal{B}} \left|f_n(x) - f(x)\right| = 0$$

hence we are done. □

Note that this sort of means that pointwise convergence implies L^1 convergence, which is not always true. It is only true when we are dealing with densities, or that $\int f(x)dx = 1$.

Theorem 17.2. *(Local CLT) Let X_i be iid with mean 0 and variance 1. Now suppose*

(a) $|\phi_X(t)|^k$ is integrable for some $k \geq 1$;
(b) $S(\delta) = \sup_{|u|\geq\delta} \{|\phi_X(u)|\} < 1$, $\forall \delta > 0$.

Then $S_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ has bounded continuous density such that $f_n(x) \to \phi$ uniformly.

First a fact is that the condition (b) is actually implied by (a), but we will not prove that here. Then, the statement that we're going to prove can be interpreted by the line: the regularity of f is related to the integrability of its transform ϕ_X, which is as natural as it can get.

Proof. Step 1: show that S_n has a density.

Let ϕ_n be the character function of S_n, then for $\forall n \geq k$ we have

$$\int |\phi_n(t)|dt = \int \left|\phi_X\left(\frac{t}{\sqrt{n}}\right)\right|^n dt$$

as proven in the first line of the proof of theorem 15.3. So we let $S = \dfrac{t}{\sqrt{n}}$ and get

$$\int |\phi_n(t)|dt = \sqrt{n}\int |\phi_X(S)|^n dS \leq \sqrt{n}\int |\phi_X(S)|^k dS < \infty$$

Now by proposition 16.2 we know that S_n is continuous with density

$$f_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_n(t)dt.$$

Step 2: show that $f_n \to \phi$.

By inverse Fourier transform we have

$$\phi(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}e^{-\frac{1}{2}t^2}dt$$

and by subtraction we get

$$\sup_x |f_n(x) - \phi(x)| \leq \frac{1}{2\pi}\int_{-\infty}^{\infty} \sup_x \left|e^{-itx}\left(\phi_n(t) - e^{-\frac{1}{2}t^2}\right)\right| dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left|\phi_n(t) - e^{-\frac{1}{2}t^2}\right| dt$$

now remember before in the proof of 15.3 we have the conclusion

$$\phi_n(t) = \phi_X\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{1}{2}\frac{t^2}{n} + o\left(\frac{t^2}{n}\right)\right)^n \tag{17.1}$$

Step 3: small & large analysis.

This is a typically method we use to deal with these kind of integrals where the bump function decays at infinity and the difference is small when x is small. We'll see what that meant.

Take $\forall \delta > 0$ and $A_n := \left[-\sqrt{n}\delta, \sqrt{n}\delta\right]$. This A_n interval encodes the case where "t is small." Thus the above integral becomes

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\phi_n(t) - e^{-\frac{1}{2}t^2}\right| dt \leq \frac{1}{2\pi}\int_{A_n}\left|\phi_n(t) - e^{-\frac{1}{2}t^2}\right| dt + \frac{1}{2\pi}\int_{A_n^C}\left|\phi_n(t)\right| + \left|e^{-\frac{1}{2}t^2}\right| dt$$

for the $t \in A_n$ part we have $\left|\frac{t}{\sqrt{n}}\right| \leq \delta$ which means that if δ is small enough, then

$$\left|\phi_X\left(\frac{t}{\sqrt{n}}\right)\right| = 1 - \frac{1}{2}\frac{t^2}{n} + o\left(\frac{t^2}{n}\right) \leq 1 - \frac{1}{4}\frac{t^2}{n}$$

by 17.1. Now we use the real number fact

$$1 + a \leq e^a$$

(because $1 + x$ is tangent to e^x from below) to get the following: for $t \in A_n$

$$|\phi_n(t)| = \left|\phi_X\left(\frac{t}{\sqrt{n}}\right)\right|^n \leq \left(1 - \frac{1}{4}\frac{t^2}{n}\right)^n \leq \left(e^{-\frac{1}{4}\frac{t^2}{n}}\right)^n = e^{-\frac{1}{4}t^2}$$

which implies

$$\left|\phi_n(t) - e^{-\frac{1}{2}t^2}\right| \cdot \mathbb{1}_{t\in A_n} \leq 2e^{-\frac{1}{4}t^2}$$

Now by Dominated convergence theorem

$$\lim_{n\to\infty}\int \mathbb{1}_{t\in A_n}\left|\phi_n(t) - e^{-\frac{1}{2}t^2}\right| dt \leq \int_{A_n}\lim_{n\to\infty}\left|\phi_n(t) - e^{-\frac{1}{2}t^2}\right| dt \leq 2\sqrt{n}\delta \cdot e^{-\frac{1}{4}t^2} \to 0$$

as $\delta \to 0$.

Now, for $t \in A_n^C$ we have

$$\int_{A_n^C}|\phi_n(t)|dt = \sqrt{n}\int_{[-\delta,\delta]^C}|\phi_X(s)|^n ds \leq \sqrt{n}S(\delta)^{n-k}\int_{[-\delta,\delta]^C}|\phi_X(s)|^k \to 0$$

as $n \to \infty$ since we know $\int |\phi_X(s)|^k ds < \infty$ is a constant, and by assumption (b) $S(\delta) = c < 1$ thus $\sqrt{n}c^{n-k} \to 0$.

For the other term $\int_{A_n^C} \left|e^{-\frac{1}{2}t^2}\right| dt$, the trick is just plainly that $A_n^C = [-\sqrt{n}\delta, \sqrt{n}\delta]$ and hence the integrand $|e^{-\frac{1}{2}t^2}|$ has upper bound $u_n] \to 0$ as $n \to 0$. Or you can just say that the tail of a normal distribution $\to 0$.

If you followed the whole proof, you'd notice that we're done. Step 1 and 2 are the main thinking part while step 3 is just technical analysis computation. □

The intuition behind this proof is that when we want to show $\int |f(x)|dx \to 0$ when we have $|f(x)| \to 0$ pointwise, we'd often are driven to use dominated convergence theorem. The trick differs in how to use it.

17.2. **Edgeworth Expansion.**

We've proven the uniform convergence of the density for the CLT, i.e. we've proven

$$f_n(x) = \phi(x) + o(1).$$

But what about if we want to make that $o(1)$ more specific? We'll deal with that in the remaining part of the lecture, and of the course.

A baby result is the following theorem.

Theorem 17.3. *For iid X_i with mean 0 and variance 1 with $\mathbb{E}[|X_i|^3] < \infty$, if the character function satisfies*

$$\int |\phi_X(t)|^k < \infty$$

for some $k \geq 1$, then for $n \geq k$, S_n has density

$$f_n(x) = \phi(x) - \frac{1}{6\sqrt{n}}\mathbb{E}[|X_i|^3]\phi'''(x) + o\left(\frac{1}{\sqrt{n}}\right).$$

Proof. Recall that we've introduced the Cumulant generating function (see Def 12.3). By the definition we have

$$\phi_X(t) = \exp\left(\sum_{r=1}^{3} K_r \frac{(it)^r}{r!} + R(t)\right) = \phi_G(t)\exp\left(K_3\frac{(it)^3}{3!} + R(t)\right)$$

where $R(t) = o(t^3)$ and $\phi_G(t)$ encodes the first two terms because the cumulant generating function of normal distribution is

$$K_{normal} = \mu t + \frac{\sigma^2}{2}t^2$$

where the mean and variance are the same with X_i.

Thus for S_n (averaged sum or R.V.), by plugging in $t = \frac{t}{\sqrt{n}}$ we have

$$\phi_n(t) = \phi_X\left(\frac{t}{\sqrt{n}}\right)^n = \phi_G\left(\frac{t}{\sqrt{n}}\right)^n \exp\left(K_3\frac{(it)^3}{3!}n^{-\frac{1}{2}} + nR\left(\frac{t}{\sqrt{n}}\right)\right)$$

where note $\phi_G\left(\frac{t}{\sqrt{n}}\right)^n = \phi_G(t)$.

Now the density inversion formula gives us

$$f_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_G(t)\exp\left(K_3\frac{(it)^3}{3!}n^{-\frac{1}{2}} + nR\left(\frac{t}{\sqrt{n}}\right)\right)dt$$

which by Taylor turns into

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_G(t)\left[1 + K_3\frac{(it)^3}{3!}n^{-\frac{1}{2}} + o\left(\frac{1}{\sqrt{n}}\right)\right]dt$$

where the error term is both the error for Taylor and $R(n)$. Continue disassembling the additions we get

$$= \phi(x) + \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_G(t)K_3\frac{(it)^3}{3!}n^{-\frac{1}{2}}dt + o\left(\frac{1}{\sqrt{n}}\right)$$

now comes the trickiest part of this proof: differentiating with respect to x the expression

$$\int_{-\infty}^{\infty} e^{-itx}\phi_G(t)dt$$

will yield a $(-1)(it)$ term. So we just differentiate it 3 times and we'll get

$$\phi'''(x) = -\frac{1}{2\pi}(it)^3\int_{-\infty}^{\infty} e^{-itx}\phi_G(t)dt$$

and thus plugging back we have that the above middle term

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_G(t)K_3\frac{(it)^3}{3!}n^{-\frac{1}{2}}dt = \frac{K_3}{6\sqrt{n}}(-1)\phi'''(x) = -\frac{\mathbb{E}[|X_i|^3]}{6\sqrt{n}}\phi'''(x)$$

which plugging back we get

$$f_n(x) = \phi(x) - \frac{\mathbb{E}[|X_i|^3]}{6\sqrt{n}}\phi'''(x) + o\left(\frac{1}{\sqrt{n}}\right).$$

□

Now we can do the same process with some more detailed analysis and calculation to get a more specified approximation. We summarize the process below.

Suppose $\mathbb{E}[|X_i|^r] < \infty$ for $r \geq 4$, then we know that

$$\phi_n(t) = \phi_G(t) \exp\left(\sum_{a=3}^{r} K_a \frac{(it)^a}{a!} n^{1-\frac{a}{2}} + nR\left(\frac{t}{\sqrt{n}}\right)\right)$$

where the remainder $nR\left(\frac{t}{\sqrt{n}}\right) = o\left(n^{1-\frac{r}{2}}\right)$. The above can be further expanded to

$$= \phi_G(t)\left(1 + K_3\frac{(it)^3}{3!}n^{-\frac{1}{2}} + \left(\frac{K_4}{4!}(it)^4 + \frac{K_3^2}{72}(it)^6\right)\frac{1}{n} + o\left(\frac{1}{n}\right)\right)$$

where the sixth order term $\frac{1}{n}\frac{K_3^2}{72}(it)^6$ really comes from the term $\frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} = \frac{1}{n}$, which explains not only the K_3^2, but also the $(it)^6$. By the same process we get the approximation

$$f_n(x) = \phi(x) - \frac{\mathbb{E}[|X_i|^3]}{6\sqrt{n}}\phi'''(x) + \left(\frac{K_4}{4!}\phi''''(x) + \frac{K_3^2}{72}\phi^{(6)}(x)\right)\frac{1}{n} + o\left(\frac{1}{n}\right).$$

Acknowledgements.