Probability Notes

Brayden Letwin

Last Updated: September 21, 2025

Contents

Chapter 1

Probability Theory

Throughout these notes we will fix a non-empty set Ω.

Definition 1 (σ-Algebra)**.** Let $A \subset \mathcal{P}(\Omega)$. A is called a σ-algebra if it contains both $\varnothing$ and Ω, and is closed under complements and countable unions.

As a result of this definition, we immediately see that A is closed under countable intersections as well, since

$$\bigcap_{n=1}^{\infty} A_n = \Omega \setminus \bigcup_{n=1}^{\infty} (\Omega \setminus A_n) .$$

Moreover, if $B, C \in A$ then $C \setminus B = C \cap (\Omega \setminus B)$, so this shows that $C \setminus B \in A$.

Definition 2 (Algebra)**.** Let $A \subset \mathcal{P}(\Omega)$. A is called an algebra if it contains both $\varnothing$ and Ω, and is closed under set differences of two elements B, C of A.

It is easy to see that A is closed under finite unions and intersections. Indeed if $B, C \in A$ then $B \cap C = B \setminus (\Omega \setminus C)$ and $B \cup C = \Omega \setminus ((\Omega \setminus B) \cap (\Omega \setminus C))$.

Definition 3 (Semi-Algebra)**.** Let $A \subset \mathcal{P}(\Omega)$. A is called a semi-algebra if it contains both $\varnothing$ and Ω, is closed under intersections of two elements B, C of A, and for any $B, C \in A$ with $B \subset C$ one can write $C \setminus B$ as a union of finitely many disjoint elements of A.

Note that every σ-algebra is an algebra, which is in turn a semi-algebra. Now we discuss rings and semi-rings.

Definition 4 (Ring)**.** Let $A \subset \mathcal{P}(\Omega)$. A is called a ring if it contains $\varnothing$, and is closed under set differences, intersections, and unions of two elements B, C of A.

Definition 5 (Semi-Ring)**.** Let $A \subset \mathcal{P}(\Omega)$. A is called a semi-ring if it contains $\varnothing$, is closed under intersections of two elements B, C of A, and for any $B, C \in A$ with $B \subset C$ one can write $C \setminus B$ as a union of finitely many disjoint elements of A.

If we start with a semi-ring A on Ω then we can complete A to a ring A^* by taking closure under finite unions. Indeed, $\varnothing \in A^*$. Since we take closure under finite unions it follows that we are then closed under set-differences. We are automatically closed under unions too then. Closure under intersections is also just passed along.

The same argument works for the completion of a semi-algebra to an algebra.

Lemma 1. *Let $\mathcal{C}$ be a collection of σ-algebras in some set Ω. Then*

$$X = \bigcap_{A \in \mathcal{C}} A$$

is again a σ-algebra.

Proof. Clear. □

This prompts the following definition.

Definition 6. Let $F \in \mathcal{P}(\Omega)$ be a set. We define $\sigma(F)$ to be the smallest σ-algebra that contains F.

$$\sigma(F) = \bigcap \{A : A \text{ sigma algebra on } \Omega \text{ and } F \subset A.\}$$

Definition 7. A measure μ on a σ-algebra A is a function from A to $[0, \infty]$. so that $\mu(\varnothing) = 0$ and μ is countably additive (note that countable additivity requires the sets to be disjoint).

Lemma 2. *Let $\mu_0 : A \longrightarrow [0, \infty]$ be a countable additive set function such that $\mu_0(\varnothing) = 0$ on an algebra A. Then μ_0 extends to a measure on $\sigma(A)$, whose restriction to A agrees with μ_0.*

Lemma 3. *If μ is a measure on a σ-algebra A then for all $B, C \in A$ with $B \subset C$ one has $\mu(B) \leq \mu(C)$.*

Proof. Write $C = B \cup (C \setminus B)$. Then $\mu(C) = \mu(B) + \mu(C \setminus B) \geq \mu(B)$. □

Now that we have these definition we can talk about measure spaces. A measure space (Ω, A, μ) is a space with a non-empty set Ω, a $\sigma-$algebra A and a measure μ. If $\mu(\Omega) = 1$ then we call (Ω, A, μ) a probability space.

Lemma 4. *If μ is a measure on a σ-algebra A then for $\{A_n\}_{n=1}^{\infty} \subset A$ one has*

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

Proof. Write $G_1 = A_1$, $G_2 = A_2 \setminus A_1$, $G_3 = A_3 \setminus (A_1 \cup A_2)$. Clearly the G_n are disjoint and also each G_n are contained in A_2, so apply countable additivity and monotonicity. □

Example 1. The Borel $\sigma-$algebra on the real line $\mathbb{R}$ is the algebra generated by clopen intervals of the form $(a, b]$ with $a \le b$. (Also it is generated by intervals of the form $(-\infty, t))$.

Lemma 5. *Suppose (Ω, A, μ) is a measure space. If $\{B_n\}$ is an increasing sequence of sets from A then*

$$\lim_{n \longrightarrow \infty} \mu(B_n) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right)$$

Lemma 6. *Suppose (Ω, A, μ) is a measure space. If $\{B_n\}$ is an decreasing sequence of sets from A with $\mu(B_k) < \infty$ for some k. Then*

$$\lim_{n \longrightarrow \infty} \mu(B_n) = \mu\left(\bigcap_{n=1}^{\infty} B_n\right)$$

Given Ω we define the outer-measure of a function $m : B \longrightarrow [0, \infty]$ by

$$m^*(A) = \inf\left\{\sum_{n=1}^{\infty} m(B_n) : B_n \in B : A \subset \bigcup_{n=1}^{\infty} B_n\right\}.$$

Here B is just an arbitrary subset of $\mathcal{P}(\Omega)$, and A is an arbitrary element of $\mathcal{P}(\Omega)$, these need not be associated with a σ-algebra. All we need is

$$\bigcup_{B_n \in B} B_n = \Omega.$$

Definition 8. Let $E \subset \mathbb{R}$. Then E is Lebesgue measurable iff there is a Borel set A such that $E \subset A$ and $\lambda^*(A \backslash E) = 0$. In this case we will have $\lambda(E) = \lambda(A)$.

This shows essentially that Lebesgue measurable sets are essentially Borel measurable sets that differ on a set of measure zero.

The collection of all Lebesgue measurable sets, denoted by $L(\mathbb{R})$ forms a σ-algebra.

Definition 9. Let (X, A) and (Y, B) be two measure spaces. A function

$$f : X \longrightarrow Y$$

is measurable if $f^{-1}(S) \in A$ for all $S \in B$.

Note that every Lebesgue measurable function is Borel measurable, and to check whether a function is Borel measurable is equivalent to checking preimages of a generating set, since preimages play nicely under set operations.

Definition 10. Let (X, A, μ) be a measure space. A simple function $f : X \longrightarrow \mathbb{R}$ is a function for which we can find m disjoint measurable subsets $A_1, \ldots, A_m$ and m real numbers $c_1, \ldots, c_m$ such that

$$f = \sum_{n=1}^{m} c_n \chi_{A_n}.$$

We define the Lebesgue integral of a simple function f by

$$\int_X f(x)\,d\mu = \sum_{n=1}^{m} c_n \mu(A_n).$$

Suppose now that $f : X \longrightarrow \mathbb{R}$ is a bounded μ a.e. function that is Borel measurable and non-negative. We then define

$$\int_X f(x)\,d\mu = \sup\left\{\int_X s(x)\,d\mu : s : X \longrightarrow \mathbb{R} : s = \sum_{n=1}^{N} \inf\{f(x) : x \in A_n\}\chi_{A_n}\right\},$$

where we run over $N \in \mathbb{N}$ and each A_n lies within the domain for which f is bounded. If f is not nessecarily bounded μ a.e. but Borel measurable and non-negative then we define $A_n = f^{-1}([0,n])$, for which $f_n = f\chi_{A_n}$ will be Borel measurable. We define

$$\int_X f(x)\,d\mu = \lim_{n\longrightarrow\infty} \int_X f_n(x)\,d\mu.$$

Finally, for general Borel measurable functions f we define

$$\int_X f(x)\,d\mu = \int_X f_+(x)\,d\mu - \int_X f_-(x)\,d\mu,$$

where $f_+ = f\chi_P$ and $f_- = f\chi_N$, here $P = f^{-1}([0,\infty))$ and $N = f^{-1}((-\infty,0))$. Finally, we set

$$\int_X |f|\,(x)\,d\mu = \int_X f_+(x)\,d\mu + \int_X f_-(x)\,d\mu.$$

We say that f is integrable if

$$\int_X f(x)\,d\mu$$

is finite. Note that f is integrable iff

$$\int_X |f|\,(x)\,d\mu < \infty.$$

Theorem 7 (Monotone Convergence Theorem)**.** *Let (X, A, μ) be a measure space and $\{f_n\}_{n=1}^{\infty}$ be a sequence of Borel or Lebesgue measurable functions such that the sequence is a non-negative sequence, also $\{f_n(x)\}_{n=1}^{\infty}$ is increasing for μ almost every $x \in X$. Then*

$$\lim_{n\longrightarrow\infty} \int_X f_n(x)d\mu = \int_X (\lim_{n\longrightarrow\infty} f_n)(x)\,d\mu.$$

Theorem 8 (Dominated Convergence Theorem I)**.** *Let (X, A, μ) be a measure space and $\{f_n\}_{n=1}^{\infty}$ be a sequence of Borel or Lebesgue measurable functions such that the sequence is a non-negative sequence, also $\lim_{n\longrightarrow\infty} f_n(x)$ exists for μ almost every $x \in X$. Finally suppose for every $n \in \mathbb{N}$ and μ almost every $x \in X$ one has $f_n(x) \leq (\lim_{n\longrightarrow\infty} f_n)(x)$. Then*

$$\lim_{n\longrightarrow\infty} \int_X f_n(x) d\mu = \int_X (\lim_{n\longrightarrow\infty} f_n)(x)\, d\mu.$$

Theorem 9 (Dominated Convergence Theorem II (Negative Functions))**.** *Let (X, A, μ) be a measure space and $\{f_n\}_{n=1}^{\infty}$ be a sequence of Borel or Lebesgue measurable functions such that $\lim_{n\longrightarrow\infty} f_n(x)$ exists for μ almost every $x \in X$. Suppose also there is an integrable function $g : X \longrightarrow [0, \infty]$ such that for every n and μ almost every $x \in X$:*

$$|f_n(x)| \leq g(x).$$

Then

$$\lim_{n\longrightarrow\infty} \int_X f_n(x) d\mu = \int_X (\lim_{n\longrightarrow\infty} f_n)(x)\, d\mu,$$

and both sides are always finite.

Lemma 10 (Fatou)**.** *Let (X, A, μ) be a measure space and $\{f_n\}_{n=1}^{\infty}$ a sequence of non-negative functions. Then*

$$\int_X (\liminf f_n)(x)\, d\mu \leq \liminf \int_X f_n(x)\, d\mu.$$

Now we discuss CDF's.

Theorem 11. *Let $F : \mathbb{R} \longrightarrow \mathbb{R}$ be a right continuous, non-decreasing, and non-constant. There there exists a unique Borel measure denoted dF on $\mathbb{R}$ such that $dF\,(a, b]) = F(b) - F(a)$.*

Definition 11. Let $X : \Omega \longrightarrow \mathbb{R}$ be (Borel or Lebesgue) measurable. X is called a random variable.

X induces a probability distribution, which is a measure $\mu : \mathcal{B}(\mathbb{R}) \longrightarrow [0, \infty]$ given by $\mu(A) = P(X^{-1}(A))$. Using this notation, X also induces a CDF denoted $F_X : \mathbb{R} \longrightarrow [0, \infty]$, and defined by

$$F_X(t) = P(X^{-1}((-\infty, t])) = P(X \leq t).$$

Definition 12 (Independence For Sets)**.** Let (Ω, A, P) be a probability space. We say that $\{A_n\}_{n\in I}$ (I countable) are independent if for all $J \subset I$ finite one has

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j).$$

Definition 13 (Independence For σ-Algebras)**.** Let (Ω, A, P) be a probability space. We say that $\{F_n\}_{n\in I}$ (I countable) are independent σ-algebras if all collections of the form $\{A_n : A_n \in F_n\}_{n=1}^{\infty}$ are independent as sets.

Definition 14 (Independence of Random Variables)**.** Let (Ω, A, P) be a probability space. We say that $\{X_n\}_{n\in I}$ (I countable) are independent random variables if $\{\sigma(X_n)\}_{n\in I}$ are independent as σ-algebras.

Definition 15. Let $(X, A), (Y, B)$ be two measurable spaces. Then we can define the (tensor) product σ-algebra by

$$A \otimes B = \sigma(\{C \times D : C \in A \text{ and } D \in B\}).$$

Lemma 12. *$\{C \times D : C \in A$ and $D \in B\}$ forms a semi-algebra on $X \times Y$.*

Then the function $\pi(C \times D) = \mu(C)\nu(D)$ where μ, ν are measures on X and Y respectively induce a measure denoted $\mu \otimes \nu$ on $X \times Y$ by Caratheodory's extension theorem.

Theorem 13 (Tonelli's Theorem)**.** *Let $(X, A, \mu), (Y, B, \nu)$ be $\sigma-$finite measure spaces and assume that $f : X \times Y \longrightarrow [0, \infty]$ is an $A \otimes B$ measurable function. Then for ν almost every $y \in Y$ the function $f^y : X \longrightarrow [0, \infty]$ defined by $f^y(x) = f(x, y)$ is measurable and then $F : Y \longrightarrow [0, \infty]$ defined by*

$$F(y) = \int_X f^y(x)\, d\mu,$$

is measurable, and

$$\int_{X\times Y} f(x, y)\, d(\mu \otimes \nu) = \int_Y \int_X f^y(x)\, d\mu\, d\nu.$$

The same statement holds respectively for the functions f^x and then we obtain

$$\int_{X\times Y} f(x, y)\, d(\mu \otimes \nu) = \int_Y \int_X f^y(x)\, d\mu\, d\nu = \int_X \int_Y f^y(x)\, d\nu\, d\mu.$$

We define for a random variable X its expectation, given by

$$E(X) = \int_\Omega X(\omega)\, dP.$$

Lemma 14. *Let $X : \Omega \longrightarrow \mathbb{R}$ be a non-negative random variable. Then*

$$E(X) = \int_0^\infty P(X > t)\, dt.$$

Lemma 15. *Let $X : \Omega \longrightarrow \mathbb{R}$ be a non-negative random variable. Then*

$$E(|X|^r) = \int_0^\infty r t^{r-1} P(|X| > t)\, dt,$$

for $r \in (1, \infty)$.

Lemma 16. *Let $X : \Omega \longrightarrow \mathbb{R}$ be a non-negative random variable, and suppose that $E(X) < \infty$. Then*

$$\lim_{t \longrightarrow \infty} tP(X > t) = 0.$$

Theorem 17 (Markov's Inequality)**.** *Let $X : \Omega \longrightarrow \mathbb{R}$ be a non-negative random variable and suppose that $E(X) < \infty$. Then for every $t \geq 0$ we have*

$$P(X > t) \leq P(X \geq t) \leq \frac{1}{t} E(X).$$

This inequality is useful when $t > E(X)$. Now recall the L_p norms, and the L_∞ norm $\|f\|_\infty = \operatorname{esssup}|f| = \inf\{M \geq 0 : |f| \leq M \text{ a.e.}\}$.

Theorem 18. $\|fg\|_1 \leq \|f\|_p \|g\|_q$ *where*

$$1/p + 1/q = 1$$

for all measurable functions $f, g : X \longrightarrow \mathbb{R}$.

Corollary 1. *Let $X : \Omega \longrightarrow \mathbb{R}$ be a random variable. Then for all $p, q \in (0, \infty)$ with $p < q$, one has that*

$$E(|X|^p)^{1/p} \leq E(|X|^q)^{1/q}.$$

Using the above, one concludes that $E(X)^2 \leq E(X^2)$, and if $E(X^2) < \infty$ then $\operatorname{Var}(X) = E((X - E(X))^2)$ is defined and finite. One also notes that $\operatorname{Var}(X) = E(X^2) - E(X)^2$.

Theorem 19 (Chebychev's Inequality)**.** *Suppose that $X : \Omega \longrightarrow \mathbb{R}$ is a random variable and assume that $E(X^2) < \infty$. Then*

$$P\left(|X - E(X)| > r\sqrt{\operatorname{Var}(X)}\right) \leq 1/r^2$$

Theorem 20 (Jensen's Inequality)**.** *Let $\varphi : I \longrightarrow \mathbb{R}$ be a convex function and X be a random variable which takes on only values in I and is integrable (i.e. $E(|X|) < \infty$). Then $\varphi(X)$ is a random variable, $E(X) \in I$, and*

$$\varphi(E(X)) \leq E(\varphi(X)).$$

Note that we can apply linearity of expectation to apply an analogous statement for concave functions, but the inequality will be reversed.

Theorem 21 (Paley-Zygmund Inequality)**.** *Let $X \in \mathbb{R}$ be a non-negative random variable, and suppose that $E(X^2) < \infty$. Then for all $t \in [0, 1]$ we have that*

$$P(X \geq tE(X)) \geq P(X > tE(X)) \geq (1 - t)^2 \frac{E(X)^2}{E(X^2)}.$$

Theorem 22. *Let X and Y be independent random variables defined on the same probability space Ω. Suppose that $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ are Borel-Borel measurable functions. Then $f(X)$ and $g(Y)$ are random variables. If $g(X)$ and $h(Y)$ are integrable, that is $E(g(X))$ and $E(h(Y))$ are defined and finite, then*

$$E(g(X)h(Y)) = E(g(X))E(h(Y)).$$

Recall that two random variables X and Y are independent iff their $\sigma-$algebras are independent iff for every $A \in \sigma(X)$ and $B \in \sigma(Y)$ one has

$$P(A \cap B) = P(A)P(B).$$

To see why $f(X)$ is a random variable, note that for any Borel set A on $\mathbb{R}$ one has $(f(X))^{-1}(A) = X^{-1}(f^{-1}(A))$, but since f is Borel-Borel measurable one has $f^{-1}(A)$ is a Borel set, and thus $X^{-1}(f^{-1}(A))$ is an element of $\sigma(X)$ which is a subset of our σ-algebra.

Definition 16. Let X and Y be two random variables that are integrable. That is, $E(X), E(Y)$ exist and are finite. We define

$$\operatorname{cov}(X, Y) = E((X - E(X))(Y - E(Y))),$$

assuming this expectation exists and is finite.

Corollary 2 (Corollary of Theorem 21)**.** *Let X and Y be independent random variables and suppose that X and Y are both integrable. Then if* $\operatorname{cov}(X, Y)$ *is defined we must have* $\operatorname{cov}(X, Y) = 0$.

Recall that the $k-$th moment for a real valued random variable x is

$$E(X^k) = \int_\Omega X^k(\omega)\, dP.$$

Definition 17. The moment generating function of a random variable X is defined as

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\, dF_X(x).$$

Note for all random variables X one has $M_X(0) = 1$. We say that the moment generating function of X exists if there is a neighbourhood about 0 such that M_X is finite on this neighbourhood. Here are some properties of MGF:

Lemma 23. *For two random variables X and Y one has $M_X(t) = M_Y(t)$ iff X and Y have the same distribution.*

Lemma 24. *If $X_1, \ldots, X_n$ are independent random variabl;es and $c_0, \ldots, c_n \in \mathbb{R}$ then for $Z = c_0 + \sum_{i=1}^n c_i X_i$ one has $M_Z(t) = e^{c_0 t} \prod_{i=1}^n M_{x_i}(c_i t)$. In particular for $X_1, \ldots X_n$ i.i.d and $\overline{X} = (1/n)\sum_{i=1}^n X_i$ one has $M_{\overline{X}}(t) = M_X(t/n)^n$.*

Lemma 25. *For a random variable X with MGF $M_X(t)$ and $\lambda > 0$ one has*

$$P(X > t) \le e^{-\lambda t} M_X(\lambda).$$

Lemma 26. *For a non-negative random variable X with MGF $M_X(t)$ AND $t > 0$ one has*

$$E[X^k] \leq (\frac{k}{te})^k M_X(t)$$

Definition 18. The characteristic function of a real-valued random variable X with CDF F_X is $\varphi_X(t) = E[e^{itX}]$.

Lemma 27. *If $X_1, \dots, X_n$ are independent random variables and $c_0, \dots, c_n \in \mathbb{R}$ then for $Z = c_0 + \sum_{i=1}^n c_i X_i$ one has $\varphi_Z(t) = e^{ic_0 t} \prod_{i=1}^n \varphi_{X_i}(c_i t)$. In particular if $X_1, \dots, X_n$ i.i.d and $\overline{X} = (1/n) \sum_{i=1}^n X_i$ one has $\varphi_{\overline{X}}(t) = \varphi_X(t/n)^n$.*

Lemma 28. *If $E[X^k] < \infty$ then $\varphi_X^{(k)}(t) = i^k E[X^k] e^{itX}$.*

Lemma 29. *For random variables X and Y one has $F_X = F_Y$ iff $\varphi_X = \varphi_Y$.*

Lemma 30 (Convergence of Distributions)**.** *For a sequence of random variables (X_n) if $\varphi_{X_n} \longrightarrow \varphi_X$ point-wise for some random variable X then $F_{X_n} \longrightarrow F_X$ point-wise.*

Theorem 31 (Inversion)**.** *If φ_x is integrable w.r.t the Lebesgue measure then the pdf f_X exists and*

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \exp(-ixt)\, d\lambda(t),$$

and

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \exp(-iut)\, d\lambda(t)\, du.$$

If that is not the case, then nevertheless at all contunuity points x for $F_X(x)$ one has

$$F_X(x) = \lim_{\sigma^2 \longrightarrow 0} \int_{-\infty}^{x} \frac{1}{2\pi} \int_{-\infty}^{\infty} \varphi(t) \exp(-iut) \exp(-\sigma^2 t^2/2)\, d\lambda(t)\, du.$$

We now move onto convergences.

Definition 19 (Convergence in Distribution/Weakly/In Law)**.** For a sequence of real valued random variables, we say that $X_n \longrightarrow_d X$ if the CDFs pointwise converge at all continuity points $t \in \mathbb{R}$ of F_X. That is,

$$\lim_{n \longrightarrow \infty} P(X_n \leq t) = P(X \leq t),$$

for all $t \in \mathbb{R}$.

Definition 20. Let $C_B(\mathbb{R})$ be the space of all continuous bounded real-valued functions on $\mathbb{R}$. For the measure space $(R, \mathcal{B})$ we say that μ_n is to converge weakly to μ if

$$\lim_{n \longrightarrow \infty} \int_{-\infty}^{\infty} f(t)\, d\mu_n = \int_{-\infty}^{\infty} f(t) d\mu,$$

for all $f \in C_B(\mathbb{R})$.

Definition 21. Convergence in Distribution Again Let (X_n) be a sequence of real-valued random variables with common domain being a measure space (Ω, A, P). We say $X_n \longrightarrow_d X$ if the induced measures

$$\mu_n \longrightarrow \mu$$

in measure. Here the induced measure of a random variable Y is $P \circ Y^{-1}$.

Theorem 32 (Portmanteau). *Let (Ω, A, P) be a probability space with real-valued random variables (X_n). The following are equivalent:*

- $P(X_n \leq x) \longrightarrow P(X \leq x)$ *for all continuity points of* $P(X \leq x)$.
- $E[f(X_n)] \longrightarrow E[f(x)]$ *for all* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *bounded and continuous.*
- $E[f(X_n)] \longrightarrow E[f(x)]$ *for all* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *bounded and unif. continuous.*
- $\limsup_{n\longrightarrow\infty} P(X_n \in C) \leq P(X \in C)$ *for all* C *closed.*
- $\liminf_{n\longrightarrow\infty} P(X_n \in O) \geq P(X \in O)$ *for all* O *open.*
- $\lim_{n\longrightarrow\infty} P(X_n \in A) = P(X \in A)$ *for all* A *contunuity sets, (i.e. sets* B *which satisfy* $P(Y \in \partial B) = 0$. *for a given r.v.* Y*)*

Definition 22. Let (X_n) be a sequence of meausrable functions. We say that $X_n \longrightarrow_P X$ if for any $\varepsilon > 0$ one has $\lim_{n\longrightarrow\infty} P(|X_n - X| > \varepsilon) = 0$. If $P(\lim_{n\longrightarrow\infty} X_n \longrightarrow X) = 1$ then we say that $X_n \longrightarrow_{a.s.} X$, called almost sure convergence.

Lemma 33. *Let X_n be i.i.d. with mean* 0 *and variance* 1. *Furthermore, let* $S_n = \sum_{i=1}^{n} X_i$. *Then* $\mathrm{Var}(S_n) = n$ *and* $\mathrm{Var}(S_n/n) = 1/n$ *and* $P(|S_n/n| \geq \varepsilon) \leq 1/(n\varepsilon^2) \longrightarrow_{n\longrightarrow\infty} 0$, *so we have convergence to zero in probability.*

Definition 23. For a probability space (Ω, A, P) let (A_i) be any sequence of sets. Define

$$\limsup_{i\longrightarrow\infty} = \bigcap_{i=1}^{\infty} \bigcup_{j>1} A_j$$

and

$$\liminf_{i\longrightarrow\infty} = \bigcup_{i=1}^{\infty} \bigcap_{j>1} A_j.$$

In probability we say that each A_i are events of $\omega \in \Omega$. Then the $\limsup$ is the events of ω which happen infinitely often and $\liminf$ is the set of events which happen eventually.

Lemma 34 (Borel Cantelli Lemma 1). *Let (A_i) be a sequence of measurable sets with* $\sum P(A_i) < \infty$. *Then* $P(\limsup_i A_i) = 0$.

Proof. By assumption $\lim_{i\longrightarrow\infty} \sum_{j>i} P(A_j) = 0$. Then for any $k \in \mathbb{N}$ we have $P(\limsup_i A_i) = P(\bigcap_{i=1}^{\infty} \bigcup_{j>i} A_j) \leq P(\bigcup_{j>K} A_j) \leq \sum_{j>k} P(A_j) \longrightarrow 0$. □

Lemma 35 (Borel Cantelli Lemma 2)**.** *Let (A_i) be a sequence of measurable sets be independent such that $\sum_{i=1}^{\infty} P(A_i) = \infty$. Then $P(\limsup_i A_i) = 1$.*

Proof. Recall that $1-t \leq e^{-t}$ for all $t \in \mathbb{R}$ and that A_i are a sequence of independent sets implies that A_i^c is. Then for any $i \in \mathbb{N}$ and $k \geq i$ one has $P(\bigcap_{j=1}^{k} A_j^c) = \prod_{j=i}^{k} [1 - P(A_j)] \leq \exp(-\sum_{j=1}^{k} P(A_j)) \longrightarrow 0$. Thus, $P(\bigcap_{j>i} A_j^c) = 0$ for any choice of i, so $P(\limsup_i A_i) = P(\bigcup_{i=1}^{\infty} \bigcap_{j>i} A_j) = 1 - P(\bigcap_{i=1}^{\infty} \bigcup_{j>i} A_j) = 1$. $\square$

Corollary 3. *Let (X_n) be a sequence of random variables such that for all $\varepsilon > 0$ one has $\sum_{n=1}^{\infty} P(|X_n - X| \geq \varepsilon) < \infty$. Then $|X_n - X| \longrightarrow_{a.s.} 0$. That is, $X_n(\omega) \longrightarrow X(\omega)$ for almost all $\omega \in \Omega$.*

Proof. For each $k \in \mathbb{N}$, $\sum_{n=1}^{\infty} P(|X_n - X| \geq 2^{-k}) < \infty$. Then BC1 says that for n sufficiently large one has $|X_n - X| \leq 2^{-k}$ a.s., that is for all $\omega \notin N_k$ for some measure zero set N_k. The union of a countable amount of measure zero sets is again measure zero, so we have $|X_n - X|(\omega) \longrightarrow 0$ for all $\omega \in \bigcup_{k=1}^{\infty} N_k$, which is $|X_n - X| \longrightarrow_{a.s.} 0$ $\square$

Corollary 4. *Let (X_n) be a sequence of random variables since that $X_n \longrightarrow_P X$. Then there is a subsequence $n_k \longrightarrow \infty$ $|X_{n_k} - X| \longrightarrow_{a.s.} 0$.*

Proof. As $X_n \longrightarrow_P X$ we can choose n_k such that $P(|X_{n_k} - X| \geq 2^{-k}) < 2^{-k}$ and so $\sum P(|X_{n_k} - X| \geq 2^{-k}) < \infty$, so by the previous corollary we conclude that $|X_{n_k} - X| \longrightarrow_{a.s.} 0$. $\square$

Corollary 5. *Let (X_n) be a sequence of random variables such that for some $p > 0$ $\sum E[|X_n - X|]^p < \infty$. Then $|X_n - X| \longrightarrow_{a.s.} 0$.*

Definition 24. We say that a sequence of random variables $X_n \longrightarrow_{L^p} X$ (convergence in L^p) if $\|X_n - X\|_p \longrightarrow 0$. This is equivalent to $E|X_n - X|^p \longrightarrow 0$.

Convergence in L^p implies convergence in probability. Just consider Markov's inequality applied to the random variable $|X_n - X|^p$.

We now prove versions of the strong law of large numbers.

Theorem 36. *If (X_n) are i.i.d. random variables from Ω to $\mathbb{R}$. Denote $S_n = X_1 + \ldots + X_n$. If $E[X_1] < \infty$ then $S_n/n \longrightarrow_{a.s.} E[X_1]$.*

Proof. Lec 22. $\square$

Theorem 37 (Parseval's Formula)**.**

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} |\varphi_X(t)|^2 \, dt < \infty$$

iff

$$\int_{-\infty}^{\infty} f_X(t)^2 \, dt < \infty$$

and these integrals coincide.

Proof. Let X' be an i.i.d copy of X. Then $\varphi_{X-X'}(t) = \varphi_X(t)\varphi(-X')(t) = \varphi_X(t)\overline{\varphi}_X(t) = |\varphi_X(t)|^2$. Similarly,

$$f_{X-X'}(t) = \int_{-\infty}^{\infty} f_X(x) f_{-X}(t-x)\, dx \leq \int_{-\infty}^{\infty} |F_X(t)|^2\, dt < \infty.$$

But then,

$$f_{X-X'}(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_{X-X'}(t)\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\varphi_X(t)|^2\, dt = \int_{-\infty}^{\infty} f_X(x) f_{-X}(-x)\, dx = \int_{\infty}^{\infty} f_X(x)^2\, dx.$$

□

Now we discuss the Central Limit Theorem:

Theorem 38. *Let (X_n) be a sequence of i.i.d. random variables from Ω to $\mathbb{R}$ and denote $S_n = X_1 + \ldots + X_n$. If $E[X_1] = 0$ and $E[X_1^2] < \infty$ then $S_n/\sqrt{n} \longrightarrow_d Z \sim N(0, \sigma^2)$ where $\sigma^2 = \mathrm{Var}(X_1)$.*

Unlike SLLN, we now require a finite second momnet, otherwise our variance may be undefined. Before proving this we need some work:

Theorem 39 (Prokhorov's Theorem)**.** *For a sequence of probability measures (μ_i) if the sequence is uniformy tight then it is sequentially compact.*

Uniformly tight means that for all $\varepsilon > 0$ there is a compact set K such that $\mu_i(K) > 1 - \varepsilon$ for all i. Sequentially compact means that for every subsequence of (μ_i) there is a weakly convergent subsubsequence.

Lemma 40. *If (μ_i) and μ are probability measures such that (μ_i) is sequentially compact to μ, then $\mu_i \longrightarrow \mu$ weakly.*

Proof. Suppose we do not have weak convergence for μ_i. Then there is a bounded continuous function f such that $\int f d\mu_i \not\longrightarrow \int f d\mu$. Then there is a subsequence i_k and $\varepsilon > 0$ such that

$$\left| \int f\, d\mu_{i_k} - \int f d\mu \right| > \varepsilon,$$

for all i_k. But this must have a convergent sub-subsequence, which is a contradiction. □

Lemma 41. *Let X be a random variable and $\delta > 0$. Then*

$$P(|X| > 2/\delta) \leq \frac{1}{\delta}\int_{-\delta}^{\delta} [1 - \varphi_X(t)]\, dt.$$

Theorem 42 (Levy Continuity Theorem)**.** *Let (X_n) be a sequence of random variables with $\varphi_{X_n}(t) \longrightarrow \varphi(t)$ for all $t \in \mathbb{R}$, with φ continuous at $t = 0$. Then there exists a random variable X with $\varphi_X = \varphi$ and $X_n \longrightarrow_d X$.*

Proof. As we have continuity of φ at 0, we see that for all $\varepsilon > 0$ there is $\delta > 0$ such that $|1 - \varphi(t)| < \varepsilon$ for all $|t| < \delta$. Then

$$\lim_{n \longrightarrow \infty} \frac{1}{\delta} \int_{-\delta}^{\delta} [1 - \varphi_{X_n}(t)]\, dt = \frac{1}{\delta} \int_{-\delta}^{\delta} [1 - \varphi_X(t)]\, dt,$$

so for large enough n we have

$$\frac{1}{\delta} \int_{-\delta}^{\delta} [1 - \varphi_{X_n}(t)]\, dt < 3\varepsilon$$

and thus by the previous Lemma we see $P(|X_n| > 2/\delta) < \varepsilon$ so (X_n) is tight. Then by Prokhorov, there is a convergent subsequence $X_{n_k} \longrightarrow_d X$, which means that $\varphi_{X_{n_k}} \longrightarrow \varphi_X = \varphi$. For any such subsequence, we have a convergent subsubsequence, all of these must have char functions which converge to φ, so we conclude that they all converge in distribution to X, so by the previous lemma we see that $X_n \longrightarrow X$. □

Proof of CLT. By independence, $E(n^{-1/2}S_n)^2 = EX_1^2$. For any $\varepsilon > 0$ there is $M_\varepsilon > 0$ such that $EX_1^2/M_\varepsilon < \varepsilon$. Hence, $P(\left|n^{-1/2}S_n\right| > M_\varepsilon) < \varepsilon$. This means that $(n^{-1/2}S_n)$ is tight. Let φ be the characteristic function of X_1. Then $\varphi(0) = \varphi'(0) = 0$. Also, $\varphi''(0) = \sigma^2$, so by Taylor's theorem we have $\varphi(t) = 1 - \sigma^2 t^2/2 + o(t^2)$. Therefore

$$\varphi_{n^{-1/2}S_n}(t) = \varphi(n^{-1/2}t)^n = (1 - \sigma^2 t^2/2n + o(t^2/n))^n \longrightarrow e^{-\sigma^2 t^2/2}.$$

Thus we conclude that $n^{-1/2}S_n \longrightarrow_d Z \sim N(0, \sigma^2)$. □

Theorem 43 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent such that there exists a_i, b_i such that $a_i \leq X_i \leq b_i$. Then*

$$P(S_n - ES_n > t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. Let $X \in [a, b]$. Then for any $\lambda > 0$ we have

$$P(X - EX > t) \leq e^{-\lambda t} E[\exp(\lambda(X - EX))].$$

WLOG, $EX = 0$. Then since $e^{\lambda X}$ is convex in λ we have

$$e^{\lambda X} \leq \frac{b - X}{b - a} e^{\lambda a} + \frac{X - a}{b - a} e^{\lambda b}$$

which implies

$$Ee^{\lambda X} \leq \frac{be^{\lambda a} - ae^{\lambda b}}{b - a},$$

hence $Ee^{\lambda X} \leq e^{-\lambda^2 (b-a)^2/8}$. Then we obtain

$$P(S_n - ES_n > t) \leq \inf_{\lambda > 0} \exp\left(-\lambda t + (1/8)\lambda^2 \sum_{i=1}^n (b_i - a_i)^2\right)$$

which is optimized for $\lambda = 4t/(\sum_{i=1}^n (b_i - a_i)^2)$, see Lec 29. □

Theorem 44 (Bernstein's)**.** *Let $X_1, \dots, X_n$ be independent such that there is $c > 0$ such that $|X_i| \le c$ for all i and $\sum_{i=1}^n EX_i^2 \le v$. Then*

$$P(S_n - ES_n > t) \le \exp\left(\frac{-t^2}{2(v+ct)}\right)$$

Theorem 45 (Continuous Mapping Theorem)**.** *Let $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a function and let D_g be the set of discontinuities of g. Let $X, (X_k)$ be a sequence of random variables, vectors, matrices and assume that $P(X \in D_g) = 0$. Then*

$$X_k \longrightarrow_d X \implies g(X_k) \longrightarrow_d g(X),$$
$$X_k \longrightarrow_P X \implies g(X_k) \longrightarrow_P g(X),$$
$$X_k \longrightarrow_{a.s.} X \implies g(X_k) \longrightarrow_{a.s.} g(X),$$

In the above theorem, we make the natural generalization of convergences by using a distance function and viewing $\mathbb{R}^n$ as a metric space. Convergence in distribution is taken as a definition of the equivalencies by Portmanteau.

Theorem 46 (Slutsky's)**.** *Let $(X_k), (Y_k)$ be sequences of real, vector, matrix valued random variables and suppose that $X_k \longrightarrow_d X$ and $Y_k \longrightarrow_P c_0$ where X is a real, vector, matrix valued random variable and c_0 is either a real, vector, or matrix. Then*

$$X_k + Y_k \longrightarrow_d X + c,$$
$$X_k \cdot Y_k \longrightarrow_d Xc,$$
$$X_k / Y_k \longrightarrow_d X/c,$$

where the last one makes sense iff c is invertible.

Now we move onto comparing modes of convergence. We have the following implications:

$$X_k \longrightarrow_{a.s \text{ or } L^p} X \implies X_k \longrightarrow_P X \implies X_k \longrightarrow_d X$$
$$X_k \longrightarrow_d c \in \mathbb{R} \implies X_k \longrightarrow_P c \in \mathbb{R} \,(\text{here } c \text{ is a constant r.v.})$$
$$X_k \longrightarrow_d X \;\not\Longrightarrow\; X_k \longrightarrow_P X \;\not\Longrightarrow\; X_k \longrightarrow_{a.s. \text{ or } L^p} X$$

First Statement. For almost sure convergence implies probability, if $X_k \longrightarrow_{a.s.} X$ then $P(\lim_{k\longrightarrow\infty} X_k = X) = 1$, so if we set $\varepsilon > 0$ then $|X_n - X| \le \sup_{k\ge n} |X_k - X|$, hence $(\sup_{k\ge n} |X_k - X| < \varepsilon) \subset (|X_n - X| < \varepsilon)$, so $P(\sup_{k\ge n} |X_k - X| < \varepsilon) \le P(|X_n - X| < \varepsilon) \le 1$. Taking limits implies

$$P(|X_n - X| < \varepsilon) = 1,$$

so $P(|X_n - X| \ge \varepsilon) = 0$.

For convergence in probability implies convergence in distribution, if $X_k \longrightarrow_P X$ then for all $\varepsilon > 0$ one has $P(|X_k - X| \ge \varepsilon) \longrightarrow 0$. Let f be any bounded and continuous function. Then $E[|f(X_k) - f(X)|] \longrightarrow 0$ by dominated convergence, so by Portmanteau we conclude $X_k \longrightarrow_d X$.

□

Third Statement. For convergence in distribution does not imply convergence in probability note that if $\Omega = \{0,1\}$ and $P(\omega = 1) = P(\omega = 2) = 1/2$ take $X_n(0) = 0 = X(1)$ and $X_n(1) = 1 = X(0)$ then $|X_n(\omega) - X(\omega)| = 1$ so we don't have convergence in probability, but $F_X(x) = F_{X_n}(x)$ which is $1/2$ for $x < 1$ and 1 for $x \geq 1$.

Now for convergence in probability does not imply convergence in a.e., take $[0,1]$ and cut it uniformly into two evenly spaced intervals $[0,1/2]$ and $[1/2,1]$. Take f_1, f_2 to be the characteristic functions of these. Do the same thing for three intervals and so on, then $\{f_n\}$ is easily seen to converge to 0 in probability and in L^p, but it does not converge almost surely, because we can find subsequences of $\{f_n\}$ which will send our given input to 0 and to 1 because our point will lie inside and outside infinitely many intervals.

For convergence in probability does not imply convergence in L^p take (X_n) such that $P(X_n = 1) = 1/n$ and $P(X_n = 0) = 1 - 1/n$. Then $X_n \longrightarrow 0$ in probability, but we don't have convergence to zero in norm, and also since $\sum P(X_n = 1) = \infty$ and the events are independent we conclude that BCL says that $P(\limsup_n \{X_n = 1\}) = 1$. $\square$

Lemma 47. *Convergence in probability implies the existence of a subsequnece which converges a.e.*

Proof. Take a sequence (n_k) which works for almost all ω. To see this, fix a subsquence (n_k) such that for each k one has $P(|X_{n_k} - X| > 2^{-k}) \leq 2^{-k}$. We construct this subsequence by inductioon and applying convergence of probability with $\varepsilon = 1/2^{-k}$. Then $\square$

Lemma 48. *Suppose $(X_k), (Y_k)$ are sequences of random variables/vectors/matrices such that*

$$X_k \longrightarrow_d X.$$

If $\operatorname{dist}(X_k, Y_k) \longrightarrow_P 0$ *then* $Y_k \longrightarrow_d X$ *as well.*