Probability Theory

December 12, 2006

Contents

1 Probability Measures, Random Variables, and Expectation

A phenomena is called *random* if the exact outcome is uncertain. The mathematical study of randomness is called the *theory of probability.*

A probability model has two essential pieces of its description.

1. Ω, the sample space, the set of possible outcomes.
   An *event* is a collection of *outcomes.* and a subset of the sample space
   $$A \subset \Omega.$$
2. P, the probability assigns a number to each event.

1.1 Measures and Probabilities

Let Ω be a sample space $\{\omega_1, \ldots, \omega_n\}$ and for $A \subset \Omega$, let $|A|$ denote the number of elements in A. Then the probability associated with *equally likely events*

$$P(A) = \frac{|A|}{|\Omega|} \tag{1.1}$$

reports the fraction of outcomes in Ω that are also in A.

Some facts are immediate:

1. $P(A) \geq 0$.
2. If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.
3. $P(\Omega) = 1$.

From these facts, we can derive several others:

Exercise 1.1. *1. If $A_1, \ldots, A_k$ are pairwise disjoint or* mutually exclusive, *($A_i \cap A_j = \emptyset$ if $i \neq j$.) then*

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k).$$

2. For any two events A and B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

3. If $A \subset B$ then $P(A) \leq P(B)$.

4. For any A, $0 \leq P(A) \leq 1$.

5. Letting A^c denote the complement *of A, then $P(A^c) = 1 - P(A)$.*

The abstracting of the idea of probability beyond finite sample spaces and equally likely events begins with demanding that the domain of the probability have properties that allow for the operations in the exercise above. This leads to the following definition.

Definition 1.2. *A nonempty collection $\mathcal{A}$ of subsets of a set S is called an* algebra *if*

1. $S \in \mathcal{A}$.
2. $A \in \mathcal{A}$ *implies* $A^c \in \mathcal{A}$.
3. $A_1, A_2 \in \mathcal{A}$ *implies* $A_1 \cup A_2 \in \mathcal{A}$.
4. *If, in addition,* $\{A_n : n = 1, 2, \cdots\} \subset \mathcal{A}$ *implies* $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$, *then* $\mathcal{A}$ *is called a* σ*-algebra.*

Exercise 1.3.
1. *Let* $S = \mathbb{R}$, *then show that the collection* $\cup_{i=1}^{k}(a_i, b_i]$, $-\infty \le a_i < b_i \le \infty$, $k = 1, 2, \ldots$ *is an algebra.*
2. *Let* $\{\mathcal{F}_i; i \ge 1\}$ *be an increasing collection of* σ*-algebras, then* $\cup_{i=1}^{\infty}\mathcal{F}_i$ *is an algebra. Give an example to show that it is not a* σ*-algebra.*

We can use these ideas we can begin with $\{A_n : n \ge 1\} \subset \mathcal{A}$ and create other elements in $\mathcal{A}$. For example,

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m = \{A_n \text{ infinitely often}\} = \{A_n \text{ i.o.}\}, \tag{1.2}$$

and

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty} A_m = \{A_n \text{ almost always}\} = \{A_n \text{ a.a.}\}. \tag{1.3}$$

Exercise 1.4. *Explain why the terms* infinitely often *and* almost always *are appropriate. Show that* $\{A_n^c \text{ i.o.}\} = \{A_n \text{ a.a.}\}^c$

Definition 1.5. *If* $\mathcal{S}$ *is* σ*-algebra, then pair* $(S, \mathcal{S})$ *is called a* measurable space.

Exercise 1.6. *An arbitrary intersection of* σ*-algebras is a* σ*-algebra. The power set of* S *is a* σ*-algebra.*

Definition 1.7. *Let* $\mathcal{C}$ *be any collection of subsets. Then,* $\sigma(\mathcal{C})$ *will denote the smallest* σ*-algebra containing* $\mathcal{C}$.

By the exercise above, this is the (non-empty) intersection of all σ-algebras containing $\mathcal{C}$.

Example 1.8.
1. *For a single set* A, $\sigma(A) = \{\emptyset, A, A^c, S\}$.
2. *If* $\mathcal{C}$ *is a* σ*-algebra, then* $\sigma(\mathcal{C}) = \mathcal{C}$.
3. *If* $S \subset \mathbb{R}^d$, *or, more generally,* S *is a topological space, and* $\mathcal{C}$ *is the set of the open sets in* S, *then* $\sigma(\mathcal{C})$ *is called the* Borel σ-algebra *and denoted* $\mathcal{B}(S)$.
4. *Let* $\{(S_i, \mathcal{S}_i)1 \le i \le n\}$ *be a set of measurable spaces, then the* product σ-algebra *on the space* $S_1 \times \cdots \times S_n$ *is* $\sigma(\mathcal{S}_1 \times \cdots \times \mathcal{S}_n)$.

These σ-algebras form the domains of measures.

Definition 1.9. *Let* $(S, \mathcal{S})$ *be a measurable space. A function* $\mu : \mathcal{S} \to [0, \infty]$ *is called a* measure *if*

1. $\mu(\emptyset) = 0$.

2. *(Additivity) If* $A \cap B = \emptyset$ *then* $\mu(A \cup B) = \mu(A) + \mu(B)$.

3. *(Continuity) If* $A_1 \subset A_2 \subset \cdots$, *and* $A = \cup_{n=1}^{\infty} A_n$, *then* $\mu(A) = \lim_{n\to\infty} \mu(A_n)$.
If in addition,

4. *(Normalization)* $\mu(S) = 1$, μ *is called a* probability.

Only 1 and 2 are needed if $\mathcal{S}$ is an algebra. We need to introduce the notion of limit as in 3 to bring in the tools of calculus and analysis.

Exercise 1.10. *Property 3 is continuity from below. Show that measures have continuity from above. If* $A_1 \supset A_2 \supset \cdots$, *and* $A = \cap_{n=1}^{\infty} A_n$, *then* $\mu(A_1) < \infty$ *implies*

$$\mu(A) = \lim_{n\to\infty} \mu(A_n).$$

Give an example to show that the hypothesis $\mu(A_1) < \infty$ *is necessary.*

Definition 1.11. *The triple* $(S, \mathcal{S}, \mu)$ *is called a* measure space *or a* probability space *in the case that* μ *is a probability.*

We will generally use the triple $(\Omega, \mathcal{F}, P)$ for a probability space. An element in Ω is called an *outcome*, a *sample point* or *realization* and a member of $\mathcal{F}$ is called an *event*.

Exercise 1.12. *Show that property 3 can be replaced with:*
3'. (Countable additivity) If $\{A_n; n \geq 1\}$ *are pairwise disjoint (*$i \neq j$ *implies* $A_i \cap A_j = \emptyset$*), then*

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

Exercise 1.13. *Define*

$$\mathcal{A} = \{A \subset \mathbb{N}; \delta(A) = \lim_{n\to\infty} \frac{|A \cap \{1, 2, \ldots, n\}|}{n} \text{ exists.}\}.$$

Definition 1.14. *A measure* μ *is called* σ-finite *if can we can find* $\{A_n; n \geq 1\} \in \mathcal{S}$, *so that* $S = \cup_{n=1}^{\infty} A_n$ *and* $\mu(A_n) < \infty$ *for each* n.

Exercise 1.15. *(first two Bonferoni inequalities) Let* $\{A_n : n \geq 1\} \subset \mathcal{S}$. *Then*

$$P(\bigcup_{j=1}^{n} A_j) \leq \sum_{j=1}^{n} P(A_j) \tag{1.4}$$

and

$$P(\bigcup_{j=1}^{n} A_j) \geq \sum_{j=1}^{n} P(A_j) - \sum_{1\leq i<j\leq n} P(A_i \cap A_j). \tag{1.5}$$

Example 1.16. *1. (Counting measure,* ν*) For* $A \in \mathcal{S}$, $\nu(A)$ *is the number of elements in* A. *Thus,* $\nu(A) = \infty$ *if* A *has infinitely many elements.* ν *is not* σ*-finite if* Ω *is uncountable.*

2. *(Lebesgue measure m on $(\mathbb{R}^1, \mathcal{B}(\mathbb{R}^1))$ For the open interval (a,b), set $m(a,b) = b-a$. Lebesgue measure generalizes the notion of length. There is a maximum σ-algebra which is smaller than the power set in which this measure can be defined. Lebesgue measure restricted to the set $[0,1]$ is a probability measure.*

3. *(Product measure) Let $\{(S_i, \mathcal{S}_i, \nu_i); 1 \le i \le k\}$ be k σ-finite measure spaces. Then the product measure $\nu_1 \times \cdots \times \nu_k$ is the unique measure on $\sigma(\mathcal{S}_1 \times \cdots \times \mathcal{S}_n)$ such that*

$$\nu_1 \times \cdots \times \nu_k(A_1 \times \cdots \times A_k) = \nu_1(A_1) \cdots \nu_k(A_k) \quad \text{for all} \quad A_i \in \mathcal{S}_i, i = 1, \ldots k.$$

Lebesgue measure on $\mathbb{R}^k$ is the product measure of k copies of Lebesgue measure on R^1.

The events $A_1 \times \cdots \times A_k$ are called *measurable rectangles*. We shall learn soon why a measure is determined by its value on measurable rectangles.

Definition 1.17. *We say A* occurs almost everywhere *(A* a.e.*) if $\mu(A^c) = 0$. If μ is a probability, we say A* occurs almost surely *(A* a.s.*). If two functions f and g satsify $f = g$ a.e., then we say that g is a* version *of f.*

1.2 Random Variables and Distributions

Definition 1.18. *Let $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ be a function between measure spaces, then f is called measurable if*

$$f^{-1}(B) \in \mathcal{S} \text{ for every } B \in \mathcal{T}. \tag{1.6}$$

If $(S, \mathcal{S})$ has a probability measure, then f is called a random variable.

For random variables we often write $\{X \in B\} = \{\omega : X(\omega) \in B\} = X^{-1}(B)$. Generally speaking, we shall use capital letters near the end of the alphabet, e.g. X, Y, Z for random variables. The range of X is called the *state space*. X is often called a *random vector* if the state space is a Cartesian product.

Exercise 1.19. 1. *The composition of measurable functions is measurable.*

2. *If $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is a random variable, then the collection*

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{S}\} \tag{1.7}$$

is a σ-algebra in Ω. Thus, X is a random variable if and only if $\sigma(X) \subset \mathcal{F}$.

3. *The collection*

$$\{B \subset S : X^{-1}(B) \in \mathcal{F}\}$$

is a σ-algebra in S

4. *If S and T are topological spaces and $\mathcal{S}$ and $\mathcal{T}$ are their respective Borel σ-algebras, then any continuous function $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ is measurable.*

We would like to limit the number of events $\{X \in B\}$ we need to verify are in $\mathcal{F}$ to establish that X is a random variable. Here is an example in the case of real-valued X.

Proposition 1.20. *Let $X : \Omega \to [-\infty, \infty]$. Then X is a random variable if and only if*

$$X^{-1}([-\infty, x]) = \{X \leq x\} \in \mathcal{F} \tag{1.8}$$

for every $x \in \mathbb{R}$.

Proof. If X is a random variable, then obviously the condition holds.

By the exercise above,

$$\mathcal{C} = \{B \subset [-\infty, \infty] : X^{-1}(B) \in \mathcal{F}\}$$

is a σ-algebra. Thus, we need only show that it contains the open sets. Because an open subset of $[-\infty, \infty]$ is the countable collection of open intervals, it suffices to show that this collection $\mathcal{C}$ contains sets of the form

$$[-\infty, x_1), (x_2, x_1), \text{ and } (x_2, \infty].$$

However, the middle is the intersection of the first and third and the third is the complement of $[-\infty, x_2]$ whose iverse image is in $\mathcal{F}$ by assumption. Thus, we need only show that $\mathcal{C}$ contains sets of the form $[-\infty, x_1)$.

However, if we choose $s_n < x_1$ with $\lim_{n\to\infty} s_n = x_1$, then

$$[-\infty, x_1) = \bigcup_{n=1}^{\infty} [-\infty, s_n] = \bigcup_{n=1}^{\infty} (s_n, -\infty]^c.$$

□

Exercise 1.21. *If $\{X_n; n \geq 1\}$ is a sequence of random variables, then*

$$X = \limsup_{n\to\infty} X_n$$

is a random variable.

Example 1.22. *1. Let A be an event. The* indicator function *for A, $I_A(s)$ equals 1 if $s \in A$, and 0 is $s \notin A$.*

2. A simple function e take on a finite number of distinct values, $e(s) = \sum_{i=1}^{n} a_i I_{A_i}(s)$, $A_1, \cdots, A_n \in \mathcal{S}$, and $a_1, \cdots, a_n \in S$. Thus, $A_i = \{s : e(s) = a_i\}$. Call this class of functions $\mathcal{E}$.

Exercise 1.23. *For a countable collection of sets, $\{A_n : n \geq 1\}$*

$$s \in \liminf_{n\to\infty} A_n \text{ if and only if } \liminf_{n\to\infty} I_{A_n}(s) = 1.$$

$$s \in \limsup_{n\to\infty} A_n \text{ if and only if } \limsup_{n\to\infty} I_{A_n}(s) = 1.$$

Definition 1.24. *Given a sequence of sets, $\{A_n : n \geq 1\}$, if there exists a set A so that*

$$I_A = \lim_{n\to\infty} I_{A_n},$$

we write

$$A = \lim_{n\to\infty} A_n.$$

Exercise 1.25. *For two sets, A and B, define the* symmetric difference $A\Delta B = (A\backslash B) \cup (B\backslash A)$. *Let $\{A_n : n \geq 1\}$ and $\{B_n : n \geq 1\}$ be sequences of sets with a limit. Show that*

1. $\lim_{n\to\infty}(A_n \cup B_n) = (\lim_{n\to\infty} A_n) \cup (\lim_{n\to\infty} B_n)$.
2. $\lim_{n\to\infty}(A_n \cap B_n) = (\lim_{n\to\infty} A_n) \cap (\lim_{n\to\infty} B_n)$.
3. $\lim_{n\to\infty}(A_n\backslash B_n) = (\lim_{n\to\infty} A_n)\backslash(\lim_{n\to\infty} B_n)$.
4. $\lim_{n\to\infty}(A_n\Delta B_n) = (\lim_{n\to\infty} A_n)\Delta(\lim_{n\to\infty} B_n)$.

Definition 1.26. *For any random variable $X : \Omega \to S$, the* distribution *of X is the probability measure*

$$\mu(B) = P(X^{-1}(B)) = P\{X \in B\}. \tag{1.9}$$

Exercise 1.27. *μ is a probability measure on S.*

Definition 1.28. *If $X : \Omega \to \mathbb{R}$, then the* distribution function *is given by*

$$F_X(x) = P\{X \leq x\} = \mu(-\infty, x].$$

Theorem 1.29. *Any distribution function has the following properties.*

1. *F_X is nondecreasing.*
2. *$\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$.*
3. *F_X is right continuous.*
4. *Set $F_X(x-) = \lim_{p\to x-} F_X(p)$. Then $F_X(x-) = P\{X < x\}$.*
5. *$P\{X = x\} = F(x) - F(x-)$.*

Proof. Because we are determining limits in a metric space, checking the limits for sequences is sufficient.

1. Use the fact that $x_1 \leq x_2$ implies that $\{X \leq x_1\} \subset \{X \leq x_2\}$.
2. Let $\{s_n; n \geq 1\}$ be an increasing sequence with limit ∞, Then $\{X \leq s_1\} \subset \{X \leq s_2\} \subset \cdots$, and $\cup_{n=1}^{\infty}\{X \leq s_n\} = \Omega$. If $\{r_n; n \geq 1\}$ is a decreasing sequence with limit $-\infty$, Then $\{X \leq r_1\} \supset \{X \leq r_2\} \supset \cdots$, and $\cap_{n=1}^{\infty}\{X \leq r_n\} = \emptyset$. Now, use the continuity properties of a probability.
3. Now, let $\{r_n; n \geq 1\}$ be a decreasing sequence with limit x. Then $\{X \leq r_1\} \supset \{X \leq r_2\} \supset \cdots$, and $\cap_{n=1}^{\infty}\{X \leq r_n\} = \{X \leq x\}$. Again, use the continuity properties of a probability.
4. Also, if $\{s_n; n \geq 1\}$ is a strictly increasing sequence with limit x, Then $\{X \leq s_1\} \subset \{X \leq s_2\} \subset \cdots$, and $\cup_{n=1}^{\infty}\{X \leq s_n\} = \{X < x\}$. Once more, use the continuity properties of a probability.
5. Note that $P\{X = x\} + P\{X < x\} = P\{X \leq x\}$ and use 3 and 4.

□

Conversely, we have the following.

Theorem 1.30. *If F satisfies 1, 2 and 3 above, then it is the distribution of some random variable.*

Proof. Let $(\Omega, \mathcal{F}, P) = ((0,1), \mathcal{B}((0,1)), m)$ where m is Lebesgue measure and define for each ω,

$$X(\omega) = \sup\{\tilde{x} : F(\tilde{x}) < \omega\}.$$

Note that because F is nondecreasing $\{\tilde{x} : F(\tilde{x}) < \omega\}$ is an interval that is *not* bounded below.

Claim. $\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$.

Because P is Lebesgue measure on $(0,1)$, the claim shows that $P\{\omega : X(\omega) \leq x\} = P\{\omega : \omega \leq F(x)\} = F(x)$.

If

$$\tilde{\omega} \in \{\omega : \omega \leq F(x)\},$$

then

$$x \notin \{\tilde{x} : F(\tilde{x}) < \tilde{\omega}\}$$

and thus

$$X(\tilde{\omega}) \leq x.$$

Consequently,

$$\tilde{\omega} \in \{\omega : X(\omega) \leq x\}.$$

On the other hand, if

$$\tilde{\omega} \notin \{\omega : \omega \leq F(x)\},$$

then

$$\tilde{\omega} > F(x)$$

and by the right continuity of F,

$$\tilde{\omega} > F(x + \epsilon)$$

for some $\epsilon > 0$. Thus,

$$x + \epsilon \in \{\tilde{x} : F(\tilde{x}) < \tilde{\omega}\}.$$

and

$$X(\tilde{\omega}) \geq x + \epsilon > x$$

and

$$\tilde{\omega} \notin \{\omega : X(\omega) \leq x\}.$$

□

The definition of distribution function extends to random vectors $X : \Omega \to \mathbb{R}^n$. Write the components of $X = (X_1, X_2, \ldots, X_n)$ and define the distribution

$$F_n(x_1, \ldots, x_n) = P\{X_1 \leq x_1, \ldots, X_n \leq x_n\}.$$

For any function $G : \mathbb{R}^n \to \mathbb{R}$ define the difference operators

$$\Delta_{k,(a_k,b_k]} G(x_1, \ldots, x_n) = G(x_1, \ldots, x_{k-1}, b_k, x_{k+1}, \ldots, x_n) - G(x_1, \ldots, x_{k-1}, a_k, x_{k+1}, \ldots, x_n).$$

Then, for example,

$$\Delta_{k,(a_k,b_k]} F(x_1, \ldots, x_n) = P\{X_1 \leq x_1, \ldots, X_{k-1} \leq x_{k-1}, X_k \in (a_k, b_k], X_{k+1} \leq x_{k+1}, \ldots, X_n \leq x_n\}.$$

Exercise 1.31. *The distribution function F_n satisfies the following conditions.*

1. *For finite intervals $I_k = (a_k, b_k]$,*

$$\Delta_{1,I_1} \cdots \Delta_{n,I_n} F_n(x_1, \ldots, x_n) \geq 0.$$

2. *If each component of $s_m = (s_{1,m}, \ldots, s_{n,m})$ decreases to $x = (x_1, \ldots, x_n)$, then*

$$\lim_{m \to \infty} F_n(s_m) = F_n(x).$$

3. *If each of the components of s_m converge to ∞, then*

$$\lim_{m \to \infty} F_n(s_m) = 1.$$

4. *If one of the components of s_m converge to $-\infty$, then*

$$\lim_{m \to \infty} F_n(s_m) = 0.$$

5. *The distribution function satisfies the consistency property,*

$$\lim_{x_n \to \infty} F_n(x_1, \ldots, x_n) = F_{n-1}(x_1, \ldots, x_{n-1}).$$

Call any function F that satisfies these properties a *distribution function*. We shall postpone until the next section our discussion on the relationship between distribution functions and distributions for multivariate random variables.

Definition 1.32. *Let $X : \Omega \to \mathbb{R}$ be a random variable. Call X*

1. discrete *if there exists a countable set D so that $P\{X \in D\} = 1$,*
2. continuous *if the distribution function F is absolutely continuous.*

Discrete random variable have densities f with respect to counting measure on D in this case,

$$F(x) = \sum_{s \in D, s \leq x} f(s).$$

Thus, the requirements for a density are that $f(x) \geq 0$ for all $x \in D$ and

$$1 = \sum_{s \in D} f(s).$$

Continuous random variable have densities f with respect to Lebesgue measure on $\mathbb{R}$ in this case,

$$F(x) = \int_{-\infty}^{x} f(s)\, ds.$$

Thus, the requirements for a density are that $f(x) \geq 0$ for all $x \in \mathbb{R}$ and

$$1 = \int_{-\infty}^{\infty} f(s)\, ds.$$

Generally speaking, we shall use the density function to describe the distribution of a random variable. We shall leave until later the arguments that show that a distribution function characterizes the the distribution.

Example 1.33 (discrete random variables)**.**

1. *(Bernoulli)* $Ber(p)$, $D = \{0, 1\}$

$$f(x) = p^x (1-p)^{1-x}.$$

2. *(binomial)* $Bin(n, p)$, $D = \{0, 1, \ldots, n\}$

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

So $Ber(p)$ *is* $Bin(1, p)$.

3. *(geometric)* $Geo(p)$, $D = \mathbb{N}$

$$f(x) = p(1-p)^x.$$

4. *(hypergeometric)* $Hyp(N, n, k)$, $D = \{\max\{0, n-N+k\}, \ldots, \min\{n, k\}\}$

$$f(x) = \frac{\binom{n}{x}\binom{N-n}{k-x}}{\binom{N}{n}}.$$

For a hypergeometric random variable, consider an urn with N *balls,* k *green. Choose* n *and let* X *be the number of green under equally likely outcomes for choosing each subset of size* n.

5. *(negative binomial)* $Negbin(a, p)$, $D = \mathbb{N}$

$$f(x) = \frac{\Gamma(a+x)}{\Gamma(a)x!} p^a (1-p)^x.$$

Note that $Geo(p)$ *is* $Negbin(1, p)$.

6. *(Poisson)* $Pois(\lambda)$, $D = \mathbb{N}$,

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

7. *(uniform)* $U(a, b)$, $D = \{a, a+1, \ldots, b\}$,

$$f(x) = \frac{1}{b-a+1}.$$

Exercise 1.34. *Check that* $\sum_{x \in D} f(x) = 1$ *in the examples above.*

Example 1.35 (continuous random variables). *1. (beta) Beta*(α, β) *on* $[0, 1]$,

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

2. (Cauchy) Cau(μ, σ^2) *on* $(-\infty, \infty)$,

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

3. (chi-squared) χ^2_a *on* $[0, \infty)$

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

4. (exponential) Exp(θ) *on* $[0, \infty)$,

$$f(x) = \theta e^{-\theta x}.$$

5. (Fisher's F) $F_{q,a}$ *on* $[0, \infty)$,

$$f(x) = \frac{\Gamma((q+a)/2) q^{q/2} a^{a/2}}{\Gamma(q/2)\Gamma(a/2)} x^{q/2-1}(a+qx)^{-(q+a)/2}.$$

6. (gamma) $\Gamma(\alpha, \beta)$ *on* $[0, \infty)$,

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

Observe that Exp(θ) *is* $\Gamma(1, \theta)$.

7. (inverse gamma) $\Gamma^{-1}(\alpha, \beta)$ *on* $[0, \infty)$,

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}.$$

8. (Laplace) Lap(μ, σ) *on* $\mathbb{R}$,

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

9. (normal) $N(\mu, \sigma^2)$ *on* $\mathbb{R}$,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

10. (Pareto) Par(α, c) *on* $[c, \infty)$,

$$f(x) = \frac{c^\alpha \alpha}{x^{\alpha+1}}.$$

11. (Student's t) $t_a(\mu, \sigma^2)$ *on* $\mathbb{R}$,

$$f(x) = \frac{\Gamma((a+1)/2)}{\sqrt{\alpha\pi}\Gamma(\alpha/2)\sigma} \left(1 + \frac{(x-\mu)^2}{a\sigma^2}\right)^{-(a+1)/2}.$$

12. *(uniform)* $U(a,b)$ *on* $[a,b]$,

$$f(x) = \frac{1}{b-a}.$$

Exercise 1.36. *Check that some of the densities have integral 1.*

Exercise 1.37 (probability transform)**.** *Let the distribution function F for X be continuous and strictly increasing, then $F(X)$ is a $U(0,1)$ random variable.*

Exercise 1.38.

1. *Let X be a continuous real-valued random variable having density f_X and let $g : \mathbb{R} \to \mathbb{R}$ be continuously differential and monotone. Show that $Y = g(X)$ has density*

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|.$$

2. *If X is a normal random variable, then $Y = \exp X$ is called a* log-normal *random variable. Give its density.*

3. *A $N(0,1)$ random variable is call a* standard normal*. Show that its square is a χ^2_1 random variable.*

1.3 Integration and Expectation

Let μ be a measure. Our next goal is to define the integral of μ with respect to a sufficiently broad class of measurable function. This definition will give us a positive linear functional so that

$$I_A \text{ maps to } \mu(A). \tag{1.10}$$

For a simple function $e(s) = \sum_{i=1}^n a_i I_{A_i}(s)$ define the *integral of e with respect to the measure μ* as

$$\int e \, d\mu = \sum_{i=1}^n a_i \mu(A_i). \tag{1.11}$$

You can check that the value of $\int e \, d\mu$ does not depend on the choice for the representation of e. By convention $0 \times \infty = 0$.

Definition 1.39. *For f a non-negative measurable function, define the* integral of f with respect to the measure μ *as*

$$\int_S f(s)\,\mu(ds) = \int f \, d\mu = \sup\{\int e \, d\mu : e \in \mathcal{E}, e \le f\}. \tag{1.12}$$

Again, you can check that the integral of a simple function is the same under either definition. If the domain of f were an interval in $\mathbb{R}$ and A_i were subintervals, then this would be giving the supremum of lower Riemann sums. The added flexibility in the choice of the A_i allows us to avoid the corresponding upper sums in the definition of the Lebesgue integral.

For general functions, denote the positive part of f, $f^+(s) = \max\{f(s), 0\}$ and the negative part of f by $f^-(s) = -\min\{f(s), 0\}$. Thus, $f = f^+ - f^-$ and $|f| = f^+ + f^-$.

If f is a real valued measurable function, then define the *integral of f with respect to the measure μ* as

$$\int f(s)\,\mu(ds) = \int f^+(s)\,\mu(ds) - \int f^-(s)\,\mu(ds).$$

provided at least one of the integrals on the right is finite. If $\int |f|\, d\mu < \infty$, then we say that f is *integrable*.

We typically write $\int_A f(s)\, \mu(ds) = \int I_A(s) f(s)\, \mu(ds)$.

If the underlying measure is a probability, then we call the integral, the *expectation* or the *expected value* and write,

$$E_P X = \int_\Omega X(\omega)\, P(d\omega) = \int X\, dP$$

and

$$E_P[X; A] = E_P[X I_A].$$

The subscript P is often dropped when their is no ambiguity in the choice of probability.

Exercise 1.40. *1. Let $e \geq 0$ be a simple function and define $\nu(A) = \int_A e\, d\mu$. Show that ν is a measure.*

2. If $f = g$ a.e., then $\int f\, d\mu = \int g\, d\mu$.

3. If $f \geq 0$ and $\int f\, d\mu = 0$, then $f = 0$ a.e.

Example 1.41. *1. If μ is counting measure on S, then $\int f\ d\mu = \sum_{s \in S} f(s)$.*

2. If μ is Lebesgue measure and f is Riemann integrable, then $\int f d\mu = \int f\ dx$, the Riemann integral.

The integral is a *positive linear functional*, i.e.

1. $\int f\, d\mu \geq 0$ whenever f is non-negative and measurable.
2. $\int (af + bg)\, d\mu = a \int f\, d\mu + \int g\ d\mu$ for real numbers a, b and integrable functions f, g.

Together, these two properies guarantees that $f \geq g$ implies $\int f\, d\mu \geq \int g\, d\mu$ provided the integrals exist.

Exercise 1.42. *Suppose f is integrable, then*

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

Exercise 1.43. *Any non-negative real valued measurable function is the increasing limit of simple functions, e.g.,*

$$f_n(s) = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} I_{\{\frac{i-1}{2^n} < f \leq \frac{i}{2^n}\}}(s) + n I_{\{f > n\}}(s).$$

Exercise 1.44. *If $\{f_n : n \geq 1\}$ is a sequence of real valued measurable functions, then $f(s) = \liminf_{n\to\infty} f_n(s)$ is measurable.*

Theorem 1.45 (Monotone Convergence)**.** *Let $\{f_n : n \geq 1\}$ be an increasing sequence of non-negative measurable functions. Then*

$$\int \lim_{n\to\infty} f_n(s)\, \mu(ds) = \lim_{n\to\infty} \int f_n(s)\, \mu(ds). \tag{1.13}$$

Proof. By the definition, $\{\int f_n \, d\mu : n \geq 1\}$ is increasing sequence of real numbers. Call its limit $L \in [0, \infty]$. By the exercise, f is a measurable function. Because integration is a positive linear functional, $\int f_n \, d\mu \leq \int f \, d\mu$, and

$$L \leq \int f \, d\mu.$$

Let e, $0 \leq e \leq f$, be a simple function and choose $c \in (0, 1)$. Define the measure $\nu(A) = \int_A e \, d\mu$ and measurable sets $A_n = \{x : f_n(x) \geq ce(x)\}$. The sets A_n are increasing and have union S. Thus,

$$\int_S f_n \, d\mu \geq \int_{A_n} f_n \, d\mu \geq c \int_{A_n} e \, d\mu = c\nu(A_n).$$

L is an upper bound for the set

$$\{c \int_{A_n} e \, d\mu : n \geq 1, c \in (0, 1)\}.$$

Thus, L is greater than its supremum, $\int_S e \, d\mu$. Finally, L is an upper bound for the set

$$\{\int e \, d\mu : e \in \mathcal{E}, e \leq f\}$$

and thus L is greater than its supremum. In other words,

$$L \geq \int f \, d\mu.$$

□

Exercise 1.46. *1. Let $\{f_k : k \geq 1\}$ be a sequence of non-negative measurable functions. Then*

$$\int \sum_{k=1}^{\infty} f_k(s) \ \mu(dx) = \sum_{k=1}^{\infty} \int f_k(s) \ \mu(dx).$$

2. Let f be a non-negative measurable function, then

$$\nu(A) = \int_A f(x) \ \mu(dx)$$

is a measure.

Theorem 1.47 (Fatou's Lemma)**.** *Let $\{f_n : n \geq 1\}$ be a sequence of non-negative measurable functions. Then*

$$\int \liminf_{n \to \infty} f_n(s) \, \mu(ds) \leq \liminf_{n \to \infty} \int f_n(s) \, \mu(ds).$$

Proof. For $k = 1, 2, \ldots$ and $x \in S$, define $g_k(x) = \inf_{i \geq k} f_i(x)$, an increasing sequence of measurable functions. Note that $g_k(x) \leq f_k(x)$, and consequently,

$$\int g_k \ d\mu \leq \int f_k \ d\mu, \quad \liminf_{k \to \infty} \int g_k \, d\mu \leq \liminf_{k \to \infty} \int f_k \, d\mu.$$

By the definition $\lim_{k\to\infty} g_k(x) = \liminf_{k\to\infty} f_k(x)$.

By the monotone convergence theorem,

$$\lim_{k\to\infty} \int g_k \, d\mu = \int \liminf_{k\to\infty} f_k \, d\mu,$$

and the result follows. □

Corollary 1.48. *Let $\{A_n : n \geq 1\} \subset \mathcal{S}$, then*

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

Exercise 1.49. *Give examples for sets $\{A_n : n \geq 1\}$ for which the inequalilties above are strict.*

Theorem 1.50 (Dominated Convergence)**.** *Suppose that f_n and g_n are measurable functions*

$$|f_n| \leq g_n, \quad f_n \to^{a.e.} f, \quad g_n \to^{a.e.} g, \quad \lim_{n\to\infty} \int g_n \, d\mu = \int g \, d\mu < \infty.$$

Then

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu$$

.

Proof. Note that, for each n, $g_n + f_n \geq 0$, and $g_n - f_n \geq 0$. Thus, Fatou's lemma applies to give

$$\liminf_{n\to\infty} \left(\int g_n \, d\mu + \int f_n \, d\mu \right) \geq \int g \, d\mu + \int f \, d\mu$$

and therefore,

$$\liminf_{n\to\infty} \int f_n \, d\mu \geq \int f \, d\mu.$$

Similarly,

$$\liminf_{n\to\infty} \left(\int g_n \, d\mu - \int f_n \, d\mu \right) \geq \int g \, d\mu - \int f \, d\mu.$$

and therefore,

$$\limsup_{n\to\infty} \int f_n \, d\mu \leq \int f \, d\mu.$$

and the theorem follows by lining up the appropriate inequalities. □

Corollary 1.51 (Bounded Convergence)**.** *Suppose that $f_n : S \to \mathbb{R}$ are measurable functions satisfying*

$$|f_n| \leq M \quad f_n \to^{a.e.} f.$$

Then, if $\mu(S) < \infty$ inplies

$$\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu$$

.

Example 1.52. *Let X be a non-negative random variable with distribution function $F_X(x) = Pr\{X \leq x\}$. Set $X_n(\omega) = \sum_{i=1}^{n2^n} \frac{i-1}{2^n} I_{\{\frac{i-1}{2^n} < X(\omega) \leq \frac{i}{2^n}\}}$. Then by the monotone convergence theorem and the definition of the Riemann-Stieltjes integral*

$$\begin{aligned} EX &= \lim_{n\to\infty} EX_n \\ &= \lim_{n\to\infty} \sum_{i=1}^{n2^n} \frac{i-1}{2^n} P\{\frac{i-1}{2^n} < X \leq \frac{i}{2^n}\} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n2^n} \frac{i-1}{2^n} (F_X(\frac{i}{2^n}) - F_X(\frac{i-1}{2^n})) \\ &= \int_0^\infty x \, dF_X(x) \end{aligned}$$

Theorem 1.53 (Change of variables)**.** *Let $h : (S, \mathcal{S}) \to (T, \mathcal{T})$. For a measure μ on S, define the induced measure $\nu(A) = \mu(h^{-1}(A))$. If $g : T \to R$, is integrable with respect to the measure ν, then*

$$\int g(t) \; \nu(dt) = \int g(h(s)) \; \mu(ds).$$

To prove this, use the *"standard machine"*.

1. Show that the identity holds for indicator functions.
2. Show, by the linearity of the integral, that the identity holds for simple functions.
3. Show, by the monotone convergence theorem, that the identity holds for non-negative functions.
4. Show, by decomposing a function into its positive and negative parts, that it holds for integrable functions.

Typically, the desired identity can be seen to satisfy properties 2-4 and so we are left to verify 1.

To relate this to a familiar formula in calculus, let $h : [a, b] \to \mathbb{R}$ be differentiable and strictly increasing, and define μ so that $\mu(c, d) = h(d) - h(c) = \int_c^d h'(t) \; dt$, then $\nu(c, d) = d - c$, i.e., ν is Lebesgue measure. In this case the change of variable reads

$$\int_a^b g(t) \, dt = \int_{h(a)}^{h(b)} g(h(s))h'(s) \, ds,$$

the Riemann change of variables formula.

Example 1.54 (Law of the Unconscience Statistician)**.** *Let $X : \Omega \to S$ be a random variable with distribution ν and let $g : S \to \mathbb{R}$ be measurable so that $E[|g(X)|] < \infty$, then*

$$E[g(X)] = \int g(x) \, \nu(dx).$$

If X is $\mathbb{R}^d$-valued, with $E|g(X)| < \infty$, then g is Riemann integrable with respect to F, and g is Lebesgue integrable with respect to ν and

$$\int g(x)\,\nu(dx) = \int g(x)\,dF(x).$$

Exercise 1.55. *1. If X is a positive real valued random variable and $E[|g(X)|] < \infty$, then*

$$E[g(X)] = \int g'(x)P\{X > x\}\,dx.$$

2. Let $h : S \to \mathbb{R}$ be integrable with respect to μ and define the measure

$$\nu(A) = \int_A h(s)\,\mu(ds).$$

If $g : S \to \mathbb{R}$ is integrable with respect to ν, then

$$\int g(s)\,\nu(ds) = \int g(s)h(s)\,\mu(ds).$$

Example 1.56. *Several choice for g have special names.*

1. *If $g(x) = x$, then $\mu = EX$ is call variously the* expectation*, the* mean*, and the* first moment.
2. *If $g(x) = x^k$, then EX^k is called the k-th* moment.
3. *If $g(x) = (x)_k$, where $(x)_k = x(x-1)\cdots(x-k+1)$, then $E(X)_k$ is called the k-th* factorial moment.
4. *If $g(x) = (x-\mu)^k$, then $E(X-\mu)^k$ is called the k-th* central moment.
5. *The second central moment $\sigma_X^2 = E(X-\mu)^2$ is called the* variance*. Note that*

$$Var(X) = E(X-\mu)^2 = EX^2 - 2\mu EX + \mu^2 = EX^2 - 2\mu^2 + \mu^2 = EX^2 - \mu^2.$$

6. *If X is $\mathbb{R}^d$-valued and $g(x) = e^{i\langle\theta,x\rangle}$, where $\langle\cdot,\cdot\rangle$ is the standard inner product, then $\phi(\theta) = Ee^{i\langle\theta,X\rangle}$ is called the* Fourier transform *or the* characteristic function*.*
7. *Similarly, if X is $\mathbb{R}^d$-valued and $g(x) = e^{\langle\theta,x\rangle}$, then $m(\theta) = Ee^{\langle\theta,X\rangle}$ is called the* Laplace transform *or the* moment generating function*.*
8. *If X is $\mathbb{Z}^+$-valued and $g(x) = z^x$, then $\rho(z) = Ez^X = \sum_{x=0}^{\infty} P\{X = x\}z^x$ is called the* (probability) generating function*.*

Exercise 1.57. *1. Show that the characteristic function is uniformly continuous.*

2. Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ be a multi-index *and define D_α be the differential operator that takes α_i derivatives of the i-th coordinate. Assume that the moment generating function m for $(X_1, \ldots, X_n)$ exists for θ in a neighborhood of the origin, then*

$$D_\alpha m(0) = E[X_1^{\alpha_1} \cdots X_n^{\alpha_n}].$$

3. *Let X be $\mathbb{Z}^+$-valued random variable with generating function ρ. If ρ has radius of convergence greater than 1, show that*

$$\rho^{(n)}(0) = E(X)_n.$$

Exercise 1.58. 1. *For a geometric random variable X, find $E(X)_2$.*

2. *For a Poisson random variable X, find $E(X)_k$.*

Exercise 1.59. *Compute parts of the following table for discrete random variables.*

random variable	parameters	mean	variance	generating function
Bernoulli	p	p	$p(1-p)$	$(1-p)+pz$
binomial	n, p	np	$np(1-p)$	$((1-p)+pz)^n$
hypergeometric	N, n, k	$\frac{nk}{N}$	$\frac{nk}{N}\left(\frac{N-k}{N}\right)\left(\frac{N-n}{N-1}\right)$	
geometric	p	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$\frac{p}{1-(1-p)z}$
negative binomial	a, p	$a\frac{1-p}{p}$	$a\frac{1-p}{p^2}$	$\left(\frac{p}{1-(1-p)z}\right)^a$
Poisson	λ	λ	λ	$\exp(-\lambda(1-z))$
uniform	a, b	$\frac{b-a+1}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\frac{z^a}{b-a+1}\frac{1-z^{b-a+1}}{1-z}$

Exercise 1.60. *Let X be an exponential random variable with parameter θ, show that the k-th moment is $k!/\theta^k$.*

Exercise 1.61. *Compute parts of the following table for continuous random variables.*

random variable	parameters	mean	variance	characteristic function
beta	α, β	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$F_{1,1}(a, b; \frac{i\theta}{2\pi})$
Cauchy	μ, σ^2	none	none	$\exp(i\mu\theta - \sigma^2)$
chi-squared	a	a	$2a$	$\frac{1}{(1-2i\theta)^{a/2}}$
exponential	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{i\lambda}{\theta+i\lambda}$
F	q, a	$\frac{a}{a-2}, a>2$	$2a^2\frac{q+a-2}{q(a-4)(a-2)^2}$	
gamma	α, β	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\left(\frac{i\beta}{\theta+i\beta}\right)^\alpha$
Laplace	μ, σ	μ	$2\sigma^2$	$\frac{\exp(i\mu\theta)}{1+\sigma^2\theta^2}$
normal	μ, σ^2	μ	σ^2	$\exp(i\mu\theta - \frac{1}{2}\sigma^2\theta^2)$
Pareto	α, c	$\frac{c\alpha}{\alpha-1}, \alpha>1$	$\frac{c^2\alpha}{(\alpha-2)(\alpha-1)^2}$	
t	a, μ, σ^2	$\mu, a>1$	$\sigma^2\frac{a}{a-2}, a>1$	
uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$-i\frac{\exp(i\theta b)-\exp(i\theta a)}{\theta(b-a)}$

2 Measure Theory

We now introduce the notion of a Sierpinski class and show how measures are uniquely determined by their values for events in this class.

2.1 Sierpinski Class Theorem

Definition 2.1. *A collection of subsets $\mathcal{S}$ of S is called a* Sierpinski class *if*

1. $A, B \in \mathcal{S}, A \subset B$ *implies* $B \backslash A \in \mathcal{S}$
2. $\{A_n; n \geq 1\} \subset \mathcal{S}$, $A_1 \subset A_2 \subset \cdots$ *implies that* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{S}$.

Exercise 2.2. *An arbitrary intersection of Sierpinski classes is a Sierpinski class. The power set of S is a Sierpinski class.*

By the exercise above, given a collection of subsets $\mathcal{C}$ of S, there exists a smallest Sierpinski class that contains the set.

Exercise 2.3. *If, in addition to the properties above,*

3. $A, B \in \mathcal{S}, A \cap B \in \mathcal{S}$
4. $S \in \mathcal{S}$

Then $\mathcal{S}$ is a σ-algebra.

Theorem 2.4 (Sierpinski class)**.** *Let $\mathcal{C}$ be a collection of subsets of a set S and suppose that $\mathcal{C}$ is closed under pairwise intersections and contains S. Then the smallest Sierpinski class of subsets of S that contains $\mathcal{C}$ is $\sigma(\mathcal{C})$.*

Proof. Let $\mathcal{D}$ be the smallest Sierpinski class containing $\mathcal{C}$. Clearly, $\mathcal{C} \subset \mathcal{D} \subset \sigma(\mathcal{C})$.

To show this, select $D \subset S$ and define

$$\mathcal{N}_D = \{A; A \cap D \in \mathcal{D}\}.$$

Claim. If $D \in \mathcal{D}$, then $\mathcal{N}_D$ is a Sierpinski class.

- If $A, B \in \mathcal{N}_D, A \subset B$ then $A \cap D, B \cap D \in \mathcal{D}$, a Sierpinski class. Therefore, $(A \cap D) \backslash (B \cap D) = (A \backslash B) \cap D \in \mathcal{D}$. Thus, $A \backslash B \in \mathcal{N}_D$.
- If $\{A_n; n \geq 1\} \subset \mathcal{N}_D$, $A_1 \subset A_2 \subset \cdots$, then $\{(A_n \cap D); n \geq 1\} \subset \mathcal{D}$. Therefore, that $\cup_{n=1}^{\infty}(A_n \cap D) = (\cup_{n=1}^{\infty} A_n) \cap D \in \mathcal{D}$. Thus, $\cup_{n=1}^{\infty} A_n \in \mathcal{N}_D$.

Claim. If $C \in \mathcal{C}$, then $\mathcal{C} \subset \mathcal{N}_C$.

Because $\mathcal{C}$ is closed under pairwise intersections, for any $A \in \mathcal{C}$, $A \cap C \in \mathcal{C} \subset \mathcal{D}$ and so $A \in \mathcal{N}_C$.

This claim has at least two consequences:

- $\mathcal{N}_C$ is a Sierpinski class that contains $\mathcal{C}$ and, consequently, $\mathcal{D} \subset \mathcal{N}_C$.
- The intersection of any element of $\mathcal{C}$ with any element of $\mathcal{D}$ is an element of $\mathcal{D}$.

Claim. If $D \in \mathcal{D}$, then $\mathcal{C} \subset \mathcal{N}_D$.

Let $C \in \mathcal{C}$. Then, by the second claim, $C \cap D \in \mathcal{D}$ and therefore, $C \in \mathcal{N}_D$.

Consequently, $\mathcal{N}_D$ is a Sierpinski class that contains $\mathcal{D}$. However, the statement that $\mathcal{D} \subset \cap_{D \in \mathcal{D}} \mathcal{N}_D$ implies that that $\mathcal{D}$ is closed under pairwise intersections. Thus, by the exercise, $\mathcal{D}$ is a σ-algebra. □

Theorem 2.5. *Let $\mathcal{C}$ be a set closed under pairwise intersection and let P and Q be probability measures on $(\Omega, \sigma(\mathcal{C}))$. If P and Q agree on $\mathcal{C}$, then they agree on $\sigma(\mathcal{C})$.*

Proof. The set $\{A; P(A) = Q(A)\}$ is easily seen to be a Sierpinski class that contains Ω. □

Example 2.6. *Consider the collection $\mathcal{C} = \{(-\infty, c]; -\infty \leq c \leq +\infty\}$. Then $\mathcal{C}$ is a set closed under pairwise intersection and $\sigma(\mathcal{C})$ is the Borel σ-algebra. Consequently, a measure is uniquely determined by its values on the sets in $\mathcal{C}$.*

More generally, in $\mathbb{R}^d$, let $\mathcal{C}$ be sets of the form

$$(-\infty, c_1] \times \cdots \times (-\infty, c_d], \qquad -\infty < c_1, \ldots, c_d \leq +\infty$$

is a set closed under pairwise intersection and $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R}^d)$.

For an infinite sequence of random variables, we will need to make additional considerations in order to state probabilities uniquely.

This give us a uniqueness of measures criterion. We now move on to finding conditions in which a finitely additive set function defined on an algebra of sets can be extended to a countably additive set function.

2.2 Finitely additive set functions and their extensions to measures

The next two lemmas look very much like the completion a metric space via equivalence classes of Cauchy sequences.

Lemma 2.7. *Let $\mathcal{Q}$ be an algebra of sets on Ω and let R be a countably additive set function on $\mathcal{Q}$ so that $R(\Omega) = 1$. Let $\{A_n; n \geq 1\} \subset \mathcal{Q}$ satisfying $\lim_{n\to\infty} A_n = \emptyset$. Then*

$$\lim_{n\to\infty} R(A_n) = 0.$$

Proof. Case I. $\{A_n; n \geq 1\}$ decreasing.

The proof is the same as in the case of a σ-algebra.

Case II. The general case.

The idea is that $\limsup_{n\to\infty} A_n = \emptyset$. For each m, p define

$$v_m(p) = R(\bigcup_{n=m}^{p} A_n).$$

Then,

$$v_m = \lim_{p\to\infty} v_m(p)$$

exists. Let $\epsilon > 0$ and choose a strictly increasing sequence $p(m)$ (In particular, $p(m) \geq m$.) so that

$$v_m - R(\bigcup_{n=m}^{p(m)} A_n) < \frac{\epsilon}{2^m}.$$

Note that $\lim_{m\to\infty} \cup_{n=m}^{p(m)} A_n = \emptyset$. Set

$$C_k = \bigcap_{m=1}^{k} (\bigcup_{n=m}^{p(m)} A_n).$$

Then $\{C_k; k \geq 1\}$ decreases with $\lim_{k\to\infty} C_k = \emptyset$. Therefore,

$$\lim_{k\to\infty} R(C_k) = 0.$$

Clearly,

$$R(A_k) \leq R(C_k) + R(A_k \backslash C_k).$$

Because $R(A_k) \geq 0$ for each k, if suffices to show that

$$R(A_k \backslash C_k) \leq \epsilon.$$

To this end, write $B_m = \cup_{n=m}^{p(m)} A_n$. Then $C_k = \cap_{m=1}^{k} B_m$

$$\begin{aligned} R(A_k \backslash C_k) &\leq R(B_k \backslash \bigcap_{m=1}^{k} B_m) \leq R(\bigcup_{m=1}^{k} (B_k \backslash B_m)) \leq \sum_{m=1}^{k} R(B_k \backslash B_m) \\ &\leq \sum_{m=1}^{k} R((\bigcup_{n=m}^{p(k)} A_n) \backslash B_m) \leq \sum_{m=1}^{k} (v_m - R(B_m)) \leq \sum_{m=1}^{k} \frac{\epsilon}{2^m} < \epsilon. \end{aligned}$$

□

Exercise 2.8. *Let $\{a_n; n \geq 1\}$ and $\{b_n; n \geq 1\}$ be bounded sequences. Then $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ both exist and are equal if and only if for every increasing sequence $\{m(n); n \geq 1\}$,*

$$\lim_{n\to\infty} (a_n - b_{m(n)}) = 0.$$

Lemma 2.9. *Let $\mathcal{Q}$ be an algebra of subsets of Ω and let R be a nonnegative countably additive set function defined on $\mathcal{Q}$ satisfying $R(\Omega) = 1$. Let $\{A_n; n \geq 1\} \subset \mathcal{Q}$ and $\{B_n; n \geq 1\} \subset \mathcal{Q}$ be sequences of sets with a common limit. Then,*

$$\lim_{n\to\infty} R(A_n) \quad \textit{and} \quad \lim_{n\to\infty} R(B_n)$$

exist and have the same limit.

Note that the limit need not be a set in $\mathcal{Q}$.

Proof. Choose an increasing sequence $\{m(n); n \geq 1\}$, and note that

$$\lim_{n\to\infty}(A_n \Delta B_{m(n)}) = (\lim_{n\to\infty} A_n)\Delta(\lim_{n\to\infty} B_{m(n)}) = \emptyset.$$

By the previous lemma,

$$\limsup_{n\to\infty} |R(A_n) - R(B_{m(n)})| \leq \lim_{n\to\infty} R(A_n \Delta B_{m(n)}) = 0.$$

Now, the lemma follows from the exercise. □

In the case that $A \in \mathcal{Q}$, taking the constant sequence for $\{B_n; n \geq 1\}$, we obtain that $R(A) = \lim_{n\to\infty} R(A_n)$.

Exercise 2.10. *A finitely additive set function R on an algebra $\mathcal{Q}$, $R(\Omega) = 1$ is countably additive if and only if $\lim_{n\to\infty} R(A_n) = 0$ for every deceasing sequence $\{A_n; n \geq 1\} \subset \mathcal{Q}$ for which $\lim_{n\to\infty} A_n = \emptyset$.*

We now go on to establish a procedure for the extension of measures. We will begin with an algebra of set $\mathcal{Q}$. Our first extension, to $\mathcal{Q}_1$, plays the role of open and closed sets. The second extension, to $\mathcal{Q}_2$, plays the role of F_σ and G_δ sets. Recall that for a regular Borel measure μ, and any measurable set E, there exists an F_σ set A and a G_δ set B. $A \subset E \subset B$ so that $\mu(B\backslash A) = 0$.

Definition 2.11. *Let $\mathcal{E}$ be an algebra of subsets of Ω and let $\tilde{R}$ be a countably additive set funtion on $\mathcal{E}$ such that $\tilde{R}(\Omega) = 1$. The* completion *of $\tilde{R}$ with respect to $\mathcal{E}$ is the collection $\mathcal{D}$ of all sets E such that there exist $F, G \in \mathcal{E}$, $F \subset E \subset G$ so that $\tilde{R}(G\backslash F) = 0$.*

Thus, if this completion with respect to $\mathcal{Q}_2$ yields a collection $\mathcal{D}$ that contains $\sigma(\mathcal{Q})$, then we can stop the procedure at this second step. We will emulate that process here.

With this in mind, set $\mathcal{Q}_0 = \mathcal{Q}$ and let $\mathcal{Q}_i$ be the limit of sequences from $\mathcal{Q}_{i-1}$. By considering constant sequences, we see that $\mathcal{Q}_{i-1} \subset \mathcal{Q}_i$.

In addition, set $R_0 = R$ and for $A \in \mathcal{Q}_i$, write $A = \lim_{n\to\infty} A_n$, $A_n \in \mathcal{Q}_{i-1}$. If R_{i-1} is countably additive, then we can extend R_i to $\mathcal{Q}_i$ by

$$R_i(A) = \lim_{n\to\infty} R_{i-1}(A_n).$$

The lemmas guarantee us that the limit does not depend on the choice of $\{A_n; n \geq 1\}$.

To check that R_i is finitely additive, choose $A = \lim_{n\to\infty} A_n$ and $B = \lim_{n\to\infty} B_n$, $A \cap B = \emptyset$. Verify that

$$A = \lim_{n\to\infty} \tilde{A}_n \quad \text{and} \quad B = \lim_{n\to\infty} \tilde{B}_n,$$

where $\tilde{A}_n = A_n\backslash B_n$ and $\tilde{B}_n = B_n\backslash A_n$. Then

$$R_i(A \cup B) = \lim_{n\to\infty} R_{i-1}(\tilde{A}_n \cup \tilde{B}_n) = \lim_{n\to\infty} R_{i-1}(\tilde{A}_n) + \lim_{n\to\infty} R_{i-1}(\tilde{B}_n) = R_i(A) + R_i(B).$$

Lemma 2.12. *If R_{i-1} is countably additive, then so is R_i.*

Proof. We have that R_{i-1} is finitely additive. Thus, it suffices to choose a decreasing sequence $\{A_n; n \geq 1\} \subset \mathcal{Q}_i$ converging to $\emptyset$ and show that $\lim_{n\to\infty} R_{i-1}(A_n) = 0$. Write

$$A_n = \lim_{m\to\infty} B_{m,n}, \quad \{B_{m,n}; m, n \geq 1\} \subset \mathcal{Q}_{i-1}.$$

Because the A_n are decreasing, we arrange that $C_{m,1} \supset C_{m,2} \supset \cdots$ by setting

$$C_{m,n} = \bigcap_{j=1}^{n} B_{m,j}$$

so that

$$\lim_{m\to\infty} C_{m,n} = \lim_{m\to\infty} (\bigcap_{j=1}^{n} B_{m,j}) = \bigcap_{j=1}^{n} A_j = A_n.$$

By definition,

$$R_i(A_n) = \lim_{m\to\infty} R_{i-1}(C_{m,n}).$$

By choosing a subsequence $m(1) < m(2) < \cdots$ appropriately, we can guarantee the convergence

$$\lim_{n\to\infty} (R_i(A_n) - R_{i-1}(C_{m(n),n})) = 0.$$

By the lemma above, the following claim completes the proof.

Claim. $\lim_{n\to\infty} C_{m(n),n} = \emptyset$.

Recall that $C_{m,1} \supset C_{m,2} \supset \cdots$. Therefore,

$$\lim_{n\to\infty} C_{m(n),n} \subset \limsup_{n\to\infty} C_{m(n),n} \subset \lim_{n\to\infty} C_{m(n),k} = A_k.$$

Now, let $k \to \infty$ □

Thus, by induction, we have that $R_i : \mathcal{Q}_i \to [0, 1]$ is countably additive.

Lemma. Let $E \in \sigma(\mathcal{Q})$. Then there exists $A, B \in \mathcal{Q}_2$ such that $A \subset E \subset B$ and $R_2(B\backslash A) = 0$.

Proof. Let $\mathcal{Q}_\uparrow$ (respectively, $\mathcal{Q}_\downarrow$) be the collection the limits of all increasing (respectively, decreasing) sequences in $\mathcal{Q}$. Note that $\mathcal{Q}_\uparrow \cup \mathcal{Q}_\downarrow \subset \mathcal{Q}_1$, the domain of R_1. Define

$$\mathcal{S} = \{E; \text{for each choice of } \epsilon > 0, \text{ there exists } F \in \mathcal{Q}_\downarrow, G \in \mathcal{Q}_\uparrow, F \subset E \subset G, R_1(G\backslash F) < \epsilon\}.$$

Note that the lemma holds for all $E \in \mathcal{S}$ and that $\mathcal{Q} \subset \mathcal{S}$. Also, $\mathcal{S}$ is closed under pairwise intersection and by taking $F = G = \Omega$, we see that $\Omega \in \mathcal{S}$. Thus, the theorem follows from the following claim.

Claim. $\mathcal{S}$ is a Sierpinski class.

To see that $\mathcal{S}$ is closed under proper set differences, choose $E_1, E_2 \in \mathcal{S}$, $E_1 \subset E_2$ and $\epsilon > 0$, then, for $i = 1, 2$ there exists

$$F_i \in \mathcal{Q}_\downarrow,\ G_i \in \mathcal{Q}_\uparrow,\ F_i \subset E_i \subset G_i,\ R_1(G_i\backslash F_i) < \frac{\epsilon}{2}.$$

Then $F_2\backslash G_1 \in \mathcal{Q}_\uparrow$, $F_1\backslash G_2 \in \mathcal{Q}_\downarrow$,

$$F_2\backslash G_1 \subset E_2\backslash E_1 \subset G_2\backslash F_1.$$

Check that $(G_2\backslash F_1)\backslash(F_2\backslash G_1) = (G_2\backslash(F_1 \cup F_2)) \cup ((G_1 \cap G_2)\backslash F_1) \subset (G_2\backslash F_2) \cup (G_1\backslash F_1)$. Thus,

$$R_1((G_2\backslash F_1)\backslash(F_2\backslash G_1)) \leq R_1(G_2\backslash F_2) + R_1(G_1)\backslash F_1) < \epsilon.$$

Now let $\{E_n; n \geq 1\} \subset \mathcal{S}$, $E_1 \subset E_2 \subset \cdots$, $E = \cup_{n=1}^{\infty} E_n$ and let $\epsilon > 0$. Consequently, we can choose

$$F_m \subset E_m \subset G_m,\ F_m \in \mathcal{Q}_\downarrow, G_m \in \mathcal{Q}_\uparrow, \quad R_1(G_m\backslash F_m) < \frac{\epsilon}{2^{m+1}}.$$

Note that

$$G = \bigcup_{n=1}^{\infty} G_n \in \mathcal{Q}_\uparrow$$

and therefore

$$R_1(G) = \lim_{N\to\infty} R_1(\bigcup_{n=1}^{N} G_n).$$

So choose N_0 so that

$$R_1(G) - R_1(\bigcup_{n=1}^{N_0} G_n) < \frac{\epsilon}{2}.$$

Now set

$$F = \bigcup_{n=1}^{N_0} F_n \in \mathcal{Q}_\downarrow,$$

and note that $F \subset E_{N_0} \subset E \subset G$. Now, by the finite additivity of R_1, we have

$$\begin{aligned} R_1(G\backslash F) &= R_1(G\backslash \bigcup_{n=1}^{N_0} G_n) + R_1((\bigcup_{n=1}^{N_0} G_n)\backslash F) \\ &< \frac{\epsilon}{2} + \sum_{n=1}^{N_0} R_1(G_n\backslash F_n) < \epsilon \end{aligned}$$

and $E \in \mathcal{S}$. □

Summarizing the discussion above, we have

Theorem 2.13. *Let $\mathcal{E}$ be an algebra of subsets of a space Ω and let R be a countably additive set function on $\mathcal{E}$ such that $R(\Omega) = 1$. Then there exist a unique probability measure P defined on $\sigma(\mathcal{E})$ such that*

$$P(A) = R(A) \quad \textit{for every } A \in \mathcal{E}.$$

Exercise 2.14. *Check the parts in the previous theorem.*

Exercise 2.15. *On $\mathbb{R}^n$, consider a collection $\mathcal{E}$ of finite unions of sets of the form*

$$(a_1, b_1] \times \cdots \times (a_n, b_n].$$

Verify that $\mathcal{E}$ is an algebra. Let F_n be a distribution function on $\mathbb{R}^n$ and define

$$R((a_1, b_1] \times \cdots \times (a_n, b_n]) = \Delta_{1,I_1} \cdots \Delta_{n,I_n} F_n(x_1, \ldots, x_n)$$

where $I_k = (a_k, b_k]$. Show that R is countably additive on $\mathcal{E}$.

3 Multivariate Distributions

How do we modify the probability of an event in light of the fact that something is known?

In a standard deck of cards, if the top card is $A\spadesuit$, what is the probability that the second card is an ace? a $\spadesuit$? a king?

All of your answers have 51 in the denominator. You have mentally restricted the sample space from Ω with 52 outcomes to $B = \{\text{all cards but } A\spadesuit\}$ with 51 outcomes. We call the answer the *conditional probability.*

For equally likely outcomes, we have a formula.

$$\begin{aligned} P(A|B) &= \text{the proportion of outcomes in } A \text{ that are also in } B \\ &= \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} \end{aligned}$$

The last identity for $P(A|B)$ with equally likely outcomes can be interpreted as the ratio of probabilities:

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

Exercise 3.1. *Let $P(B) > 0$, then*

$$Q(A) = P(A|B)$$

is a probability meaasure.

We now say that A *independent* of B if

$$P(A|B) = P(A)$$

or, using the formula above,

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

and B is independent of A.

Exercise 3.2. *If A and B are independent, then so are A and B^c, A^c and B, and A^c and B^c. Thus, every event in $\sigma(A)$ is independent of every event in $\sigma(B)$.*

We now look to a definition that works more generally.

3.1 Independence

Definition 3.3. *1. A collection of σ-algebras $\{\mathcal{F}_\lambda; \lambda \in \Lambda\}$ are* independent *if for any finite choice $A_1 \in \mathcal{F}_{\lambda_1}, \ldots, A_n \in \mathcal{F}_{\lambda_n}$*

$$P(\bigcap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i).$$

2. *A collection of events* $\{A_\lambda : \lambda \in \Lambda\}$ *are* independent *if* $\{\sigma(A_\lambda); \lambda \in \Lambda\}$ *are independent.*

3. *A collection of random variables* $\{X_\lambda : \lambda \in \Lambda\}$ *are* independent *if* $\{\sigma(X_\lambda); \lambda \in \Lambda\}$ *are independent. In other words, for events* B_λ *in the state space for* X_λ*,*

$$P(\bigcap_{i=1}^{n}\{X_{\lambda_i} \in B_{\lambda_i}\}) = \prod_{i=1}^{n} P\{X_{\lambda_i} \in B_{\lambda_i}\}.$$

Exercise 3.4. 1. *A collection of events* $\{A_\lambda : \lambda \in \Lambda\}$ *are independent if and only if the collection of random variables* $\{I_{A_\lambda} : \lambda \in \Lambda\}$ *are independent.*

2. *If a sequence* $\{X_k; k \geq 1\}$ *of random variables are independent, then*

$$P\{X_1 \in A_1, X_2 \in A_2 \ldots\} = \prod_{i=1}^{\infty} P\{X_i \in A_i\}.$$

3. *If* $\{X_\lambda : \lambda \in \Lambda\}$ *are independent and* $\{f_\lambda : \lambda \in \Lambda\}$ *are measurable function on the range of* X_λ *then,* $\{f(X_\lambda) : \lambda \in \Lambda\}$ *are independent.*

Theorem 3.5. *Let* Λ *be finite and write* $\Lambda = \Lambda_1 \cup \Lambda_2$*, with* $\Lambda_1 \cap \Lambda_2 = \emptyset$*, then*

$$\mathcal{F}_1 = \sigma\{X_\lambda : \lambda \in \Lambda_1\} \text{ and } \mathcal{F}_2 = \sigma\{X_\lambda : \lambda \in \Lambda_2\}$$

are independent.

Proof. Let $\{\lambda_1, \ldots, \lambda_m\} \subset \Lambda_2$ and define

$$D = \{X_{\lambda_1} \in B_1, \ldots, X_{\lambda_m} \in B_m\}.$$

Assume $P(D) > 0$ and define

$$P_1(C) = P(C|D),\ C \in \mathcal{F}_1.$$

If $C = \{X_{\tilde{\lambda}_1} \in \tilde{B}_1, \ldots, X_{\tilde{\lambda}_m} \in \tilde{B}_k\}$ Then,

$$P(C \cap D) = P(C)P(D)$$

and

$$P_1(C) = P(C).$$

But such sets form a Sierpinski class $\mathcal{C}$ closed under pairwise intersection with $\sigma(\mathcal{C}) = \mathcal{F}_1$. Thus, $P_1 = P$ on $\mathcal{F}_1$.

Now fix an arbitrary $C \in \mathcal{F}_1$ with $P(C) > 0$ and define

$$P_2(D) = P(D|C),\ D \in \mathcal{F}_2.$$

Arguing as before we obtain

$$P_2(D) = P(D),\ D \in \mathcal{F}_2.$$

Therefore,

$$P(C \cap D) = P(C)P(D),\quad C \in \mathcal{F}_1, D \in \mathcal{F}_2$$

whenever $P(C) > 0$. But this identity is immediate if $P(C) = 0$. Thus, $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent. □

When we learn about infinite products and the product topology, we shall see that the theorem above holds for arbitrary Λ with the same proof.

Exercise 3.6. *Let $\{\Lambda_j; j \in J\}$ be a partition of a finite set Λ. Then the σ-algebras $\mathcal{F}_j = \sigma\{X_\lambda; \lambda \in \Lambda_j\}$ are independent.*

Thus, if X_i has distribution ν_i, then for $X_1, \ldots, X_n$ independent and for measurable sets B_i, subsets of the range of X_i, we have

$$P\{X_1 \in B_1, \ldots, X_n \in B_n\} = \nu_1(B_1) \cdots \nu_n(B_n) = (\nu_1 \times \cdots \nu_n)(B_1 \times \cdots \times B_n),$$

the product measure.

We now relate this to the distribution functions.

Theorem 3.7. *The random variables $\{X_n; n \geq 1\}$ are independent if and only if their distribution functions satisfy*

$$F_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n).$$

Proof. The necessity follows by considering sets $\{X_1 \leq x_1, \ldots, X_n \leq x_n\}$.

For sufficiency, note that the case $n = 1$ is trivial. Now assume that this holds for $n = k$, i.e., the product representation for the distribution function implies that for all Borel sets $B_1, \ldots, B_k$,

$$P\{X_1 \in B_1, \ldots, X_k \in B_k\} = P\{X_1 \in B_1\} \cdots P\{X_k \in B_k\}.$$

Define

$$Q_1(B) = P\{X_{k+1} \in B\} \text{ and } \tilde{Q}_1(B) = P\{X_{k+1} \in B | X_1 \leq x_1, \ldots, X_k \leq x_k\}.$$

Then $Q_1 = \tilde{Q}_1$ on sets of the form $(-\infty, x_{k+1}]$ and thus, by the Sierpinski class theorem, for all Borel sets. Thus,

$$P\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_k \leq x_k, X_{k+1} \in B\} = P\{X_1 \leq x_1, \ldots, X_k \leq x_k\} P\{X_{k+1} \in B\}$$

and $X_1, \ldots, X_{k+1}$ are independent. □

Exercise 3.8. *1. For independent random variables X_1, X_2 choose measurable functions f_1 and f_2 so that $E|f_1(X_1)f_2(X_2)| < \infty$, then*

$$E[f_1(X_1)f_2(X_2)] = E[f_1(X_1)]\, E[f_2(X_2)].$$

*(*Hint: *Use the standard machine.)*

2. If X_1, X_2 are independent random variables having finite variance, then

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2).$$

Corollary 3.9. *For independent random variables $X_1, \ldots X_n$ choose measurable functions $f_1, \ldots f_n$ so that*

$$E|\prod_{i-1}^{n} f_i(X_i)| < \infty,$$

then

$$E[\prod_{i-1}^{n} f_i(X_i)] = \prod_{i-1}^{n} E[f_i(X_i)].$$

Thus, we have three equivalent identities to establish independence, using either the distribution, the distribution function, and products of functions of random variables.

We begin the proofs of equivalence with the fact that measures agree on a Sierpinski class, $\mathcal{S}$. If we can find a collection of events $\mathcal{C} \subset \mathcal{S}$ that contains the whole space and in closed under intersection, then we can conclude by the Sierpinski class theorem that they agree on $\sigma(\mathcal{C})$.

The basis for this choice, in the case where the state space S^n is a product of topological spaces, is that a collection $U_1 \times \cdots \times U_n$ forms a subbasis for the topology whenever U_i are arbitrary choices from a subbasis for the topology of S.

Exercise 3.10. *1. Let $\mathbb{Z}^+$-valued random variables $X_1, \ldots, X_n$ have generating functions $\rho_{X_1}, \ldots, \rho_{X_n}$, then*

$$\rho_{X_1+\cdots+X_n} = \rho_{X_1} \times \cdots \times \rho_{X_n}.$$

Show when the sum of independent

(a) binomial random variables is a binomial random variable,

(b) negative binomial random variables is a negative binomial random variable,

(c) Poisson random variables is a Poisson random variable.

Definition 3.11. *Let X_1 and X_2 have finite variance. If their means are μ_1 and μ_2 respectively, then their* covariance *is defined to be*

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = EX_1X_2 - \mu_2 EX_1 - \mu_1 EX_2 + \mu_1\mu_2 = EX_1X_2 - \mu_1\mu_2.$$

If both of these random variables have positive variance, then their correlation coefficient

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)\, Var(X_2)}}.$$

For a vector valued random variable $X = (X_1, \ldots, X_n)$ define the *covariance matrix* $\mathrm{Var}(X)$ as a matrix whose i, j entry is $\mathrm{Cov}(X_i, X_j)$

Exercise 3.12. *1. If X_1 and X_2 are independent, then $\rho(X_1, X_2) = 0$. Give an example to show that the converse is not true.*

2. Let $\sigma^2_{X_i} = Var(X_i), i = 1, 2$, then

$$\sigma^2_{X_1+X_2} = \sigma^2_{X_1} + \sigma^2_{X_2} + 2\sigma_{X_1}\sigma_{X_2}\rho(X_1, X_2).$$

3. $-1 \le \rho(X_1, X_2) \le 1$. Under what circumstances is $\rho(X_1, X_2) = \pm 1$?

4. Assume that the random variables $\{X_1, \ldots, X_n\}$ have finite variance and that each pair is uncorrelated. Then

$$Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n).$$

5. Check that the covariance satisfies

$$Cov(a_1X_1 + b_1, a_2X_2 + b_2) = a_1a_2\, Cov(X_1, X_2).$$

In particular $Var(aX) = a^2\, Var(X)$.

6. *Let $a_1, a_2 > 0$, and $b_1, b_2 \in \mathbb{R}$, then $\rho(a_1X_1 + b_1, a_2X_2 + b_2) = \rho(X_1, X_2)$.*

7. *Let A be a $d \times n$ matrix and define $Y = AX$, then $Var(Y) = A\,Var(X)A^T$. The case $d = 1$ shows that the covariance matrix in non-negative definite.*

3.2 Fubini's theorem

Theorem 3.13. *Let $(S_i, \mathcal{A}_i, \mu_i), i = 1, 2$ be two σ-finite measures. If $f : S_1 \times S_2 \to \mathbb{R}$ is integrable with respect to $\mu_1 \times \mu_2$, then*

$$\int f(s_1, s_2)\ (\mu_1 \times \mu_2)(ds_1 \times ds_2) = \int [\int f(s_1, s_2)\ \mu_1(ds_1)]\mu_2(ds_2) = \int [\int f(s_1, s_2)\ \mu_2(ds_2)]\mu_1(ds_1).$$

Use the "standard machine" to prove this. Use the Sierpinski class theorem to argue that it suffices to begin with indicators of sets of the form $A_1 \times A_2$. The identity for non-negative functions is known as Tonelli's theorem.

Example 3.14. *If f_n is measurable, then consider the measure $\mu \times \nu$ where ν is counting measure on $\mathbb{Z}^+$ to see that*

$$\sum_{n=1}^{\infty} \int |f_n|\, d\mu < \infty,$$

implies

$$\sum_{n=1}^{\infty} \int f_n\, d\mu = \int \sum_{n=1}^{\infty} f_n\, d\mu.$$

Exercise 3.15. *Assume that $(X_1, \ldots, X_n)$ has distribution function $F_{(X_1,\ldots,X_n)}$ and density $f_{(X_1,\ldots,X_n)}$ with respect to Lebesgue measure.*

1. *The random variables $(X_1, \ldots, X_n)$ with density $f_{(X_1,\ldots,X_n)}$ are independent if and only if*

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

where f_{X_k} is the density of $X_k, k = 1, 2, \ldots, n$

2. *The* marginal density

$$f_{(X_1,\ldots,X_{n-1})}(x_1, \ldots, x_{n-1}) = \int f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n)\, dx_n.$$

Let X_1 and X_2 be independent $\mathbb{R}^d$-valued random variables having distributions ν_1 and ν_2 respectively. Then the distribution of their sum,

$$\nu(B) = P\{X_1 + X_2 \in B\} = \int \int I_B(x_1 + x_2)\, \nu_1(dx_1)\nu_2(dx_2) = \int \nu_1(B - x_2)\nu_2(dx_2) = (\nu_1 * \nu_2)(B),$$

the *convolution* of the measures ν_1 and ν_2.

If ν_1 and ν_2 have densities f_1 and f_2 with respect to Lebesgue measure, then

$$\nu(B) = \int \int I_B(x_1 + x_2) f_1(x_1) f_2(x_2)\, dx_1 dx_2 = \int \int I_B(s) f_1(s - y) f_2(y)\, dy ds = \int_B (f_1 * f_2)(s)\, ds,$$

the *convolution* of the functions f_1 and f_2. Thus, ν has the convolution $f_1 * f_2$ as its density with respect to Lebesgue measure.

Exercise 3.16. *Let X and Y be independent random variables and assume that the distribution of X has a density with respect to Lebesgue measure. Show that the distribution of $X+Y$ has a density with respect to Lebesgue measure.*

A similar formula holds if we have a $\mathbb{Z}^d$ valued random variable and look at random variable that are absolutely continuous with respect to counting measure.

$$(f_1 * f_2)(s) = \sum_{y\in\mathbb{Z}^d} f_1(s-y)f_2(y), \quad \text{and} \quad \nu(B) = \sum_{s\in B}(f_1 * f_2)(s).$$

Exercise 3.17.

1. *Let X_i be independent $N(\mu_i, \sigma_i^2)$ random variables, $i = 1, 2$. Then $X_1 + X_2$ is a $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ random variable.*
2. *Let X_i be independent $\chi^2_{a_i}$ random variables, $i = 1, 2$. Then $X_1 + X_2$ is a $\chi^2_{a_1+a_2}$ random variable.*
3. *Let X_i be independent $\Gamma(\alpha_i, \beta)$ random variables, $i = 1, 2$. Then $X_1 + X_2$ is a $\Gamma(\alpha_1 + \alpha_2, \beta)$ random variable.*
4. *Let X_i be independent $Cau(\mu_i, \sigma_i)$ random variables, $i = 1, 2$. Then $X_1 + X_2$ is a $Cau(\mu_1 + \mu_2, \sigma_1 + \sigma_2)$ random variable.*

Exercise 3.18. *If X_1 and X_2 have joint density $f_{(X_1,X_2)}$ with respect to Lebesgue measure, then their sum Y has density*

$$f_Y(y) = \int f(x, y-x)\, dx.$$

Example 3.19 (Order statistics)**.** *Let $X_1, \ldots, X_n$ be independent random variables with common distribution function F. Assume F has density f with respect to Lebesgue measure. Let $X_{(k)}$ be the k-th smallest of $X_1, \ldots, X_n$. (Note that the probability of a tie is zero.) To find the density of the order statistcs, note that $\{X_{(k)} \le x\}$ if and only if at least k of the random variables lie in $(-\infty, x]$. Its distribution function*

$$F_{(k)}(x) = \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1-F(x))^{n-j}$$

and its density

$$\begin{aligned} f_{(k)}(x) &= f(x) \sum_{j=k}^{n} \left(j\binom{n}{j} F(x)^{j-1}(1-F(x))^{n-j} - (j-1)\binom{n}{j+1} F(x)^j (1-F(x))^{n-j+1} \right) \\ &= f(x) k \binom{n}{k} F(x)^{k-1}(1-F(x))^{n-k}. \end{aligned}$$

Note that in the case that the random variable are $U(0,1)$, we have that the order statistics are beta random variables.

3.3 Transformations of Continuous Random Variables

For a one-to-one transformation g of a continuous random variable X, we saw how that the density of $Y = g(X)$ is

$$f_Y(y) = f_X(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|.$$

In multiple dimensions, we will need to use the *Jacobian*. Now, let $g : S \to \mathbb{R}^n$, $S \subset \mathbb{R}^n$ be one-to-one and differentiable and write $y = g(x)$. Then the Jacobian we need is based on the inverse function $x = g^{-1}(y)$.

$$Jg^{-1}(y) = \det \begin{pmatrix} \frac{\partial g_1^{-1}(y)}{\partial y_1} & \frac{\partial g_1^{-1}(y)}{\partial y_2} & \dots & \frac{\partial g_1^{-1}(y)}{\partial y_n} \\ \frac{\partial g_2^{-1}(y)}{\partial y_1} & \frac{\partial g_2^{-1}(y)}{\partial y_2} & \dots & \frac{\partial g_2^{-1}(y)}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}(y)}{\partial y_1} & \frac{\partial g_n^{-1}(y)}{\partial y_2} & \dots & \frac{\partial g_n^{-1}(y)}{\partial y_n} \end{pmatrix}$$

Then

$$f_Y(y) = f_X(g^{-1}(y))|Jg^{-1}(y)|.$$

Example 3.20. *1. Let A be an invertible $d \times d$ matrix and define*

$$Y = AX + b.$$

Then, for $g(x) = Ax + b$, $g^{-1}(y) = A^{-1}(y - b)$, and

$$Jg^{-1}(y) = A^{-1}, \text{ and } f_Y(y) = \frac{1}{|det(A)|} f_X(A^{-1}(y - b)).$$

2. Let X_1 and X_2 be independent $Exp(1)$ random variables. Set

$$Y_1 = X_1 + X_2, \text{ and } Y_2 = \frac{X_1}{X_1 + X_2}. \text{ Then, } X_1 = Y_1Y_2, \text{ and } X_2 = Y_1(1 - Y_2).$$

The Jacobian for $g^{-1}(y_1, y_2) = (y_1y_2, y_1(1 - y_2))$,

$$Jg^{-1}(y) = det \begin{pmatrix} y_2 & y_1 \\ (1 - y_2) & -y_1 \end{pmatrix} = -y_1.$$

Therefore,

$$f_{(Y_1,Y_2)}(y_1, y_2) = y_1e^{-y_1}$$

on $[0, \infty) \times [0, 1]$. Thus, Y_1 and Y_2 are independent. Y_1 is χ^2_2 and Y_2 is $U(0, 1)$.

3. Let X_1 and X_2 be independent standard normals and define

$$Y_1 = \frac{X_1}{X_2}, \text{ and } Y_2 = X_2. \text{ Then, } X_1 = Y_1Y_2, \text{ and } X_2 = Y_2.$$

The Jacobian for $g^{-1}(y_1, y_2) = (y_1 y_2, y_2)$,

$$Jg^{-1}(y) = det\begin{pmatrix} y_2 & y_1 \\ 0 & 1 \end{pmatrix} = y_2.$$

Therefore,

$$f_{(Y_1,Y_2)}(y_1, y_2) = \frac{1}{2\pi} \exp \frac{-y_2^2(y_1^2+1)}{2} |y_2|$$

and

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{(Y_1,Y_2)}(y_1, y_2)|y_2|\, dy_2 = \frac{1}{\pi}\int_0^{\infty} \exp \frac{-y_2^2(y_1^2+1)}{2} y_2\, dy_2 \\ &= \frac{1}{\pi}\frac{1}{y_1^2+1} \exp \frac{-y_2^2(y_1^2+1)}{2}\Big|_0^{\infty} = \frac{1}{\pi}\frac{1}{y_1^2+1}. \end{aligned}$$

and Y_1 *is a Cauchy random variable.*

Exercise 3.21. *Let* U_1 *and* U_2 *be independent* $U(0,1)$ *random variables. Define*

$$R = \sqrt{-2\ln U_1} \text{ and } \Theta = 2\pi U_2.$$

Show that

$$X_1 = R\sin\Theta \text{ and } X_2 = R\cos\Theta.$$

are independent $N(0,1)$ *random variables.*

Example 3.22. *Let* X_1 *be a standard normal random variable and let* X_2 *be a* χ_a^2 *random variable. Assume that* X_1 *and* X_2 *are independent. Then their joint density is*

$$f_{(X_1,X_2)}(x_1, x_2) = \frac{1}{\sqrt{2\pi}\Gamma(a/2)2^{a/2}} e^{-x_1^2/2} x_2^{a/2-1} e^{-x_2/2}.$$

A random variable T *having the* t *distribution with* a *degrees of freedom is obtained by*

$$T = \frac{X_1}{\sqrt{X_2/a}}.$$

To find the density of T *consider the transformation*

$$(y_1, y_2) = g(x_1, x_2) = \left(\frac{x_1}{\sqrt{x_2/a}}, x_2\right).$$

This map is a one-to-one transformation from $\mathbb{R}\times(0,\infty)$ *to* $\mathbb{R}\times(0,\infty)$ *with inverse*

$$(x_1, x_2) = g^{-1}(y_1, y_2) = (y_1\sqrt{y_2/a}), y_2).$$

The Jacobian

$$Jg^{-1}(y) = det\begin{pmatrix} \sqrt{y_2/a} & y_1/(2\sqrt{y_2 a}) \\ 0 & 1 \end{pmatrix} = \sqrt{y_2/a}.$$

Therefore,

$$f_{(Y_1,Y_2)}(y_1,y_2) = \frac{1}{\sqrt{2\pi}\Gamma(a/2)2^{a/2}} y_2^{a/2-1} \exp\frac{-y_2}{2}\left(1+\frac{y_1^2}{a}\right).$$

The marginal density for T is

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi}\Gamma(a/2)2^{a/2}} \int_0^\infty y_2^{a/2-1} \exp\frac{-y_2}{2}\left(1+\frac{t^2}{a}\right)\sqrt{\frac{y_2}{a}}\, dy_2,\ u = \frac{-y_2}{2}\left(1+\frac{t^2}{a}\right) \\ &= \frac{1}{\sqrt{2\pi a}\Gamma(a/2)2^{a/2}} \int_0^\infty \left(\frac{2u}{1+t^2/a}\right)^{a/2-1/2} e^{-u}\left(\frac{2}{1+t^2/a}\right) du \\ &= \frac{\Gamma((a+1)/2)}{\sqrt{2\pi a}\Gamma(a/2)} \frac{1}{(1+t^2/a)^{a/2+1/2}} \end{aligned}$$

Exercise 3.23. *Let $X_i, i = 1,2$, be independent $\chi^2_{a_i}$ random variables. Find the density with respect to Lebesgue measure for*

$$F = \frac{X_1/a_1}{X_2/a_2}.$$

Verify that this is the density of an F-distribution with parameters a_1 and a_2

3.4 Conditional Expectation

In this section, we shall define conditional expectation with respect to a random variable. Later, this definition with be genrealized to conditional expectation with respect to a σ-algebra.

Definition 3.24. *Let Z be an integrable random variable on $(\Omega, \mathcal{F}, P)$ and let X be any random variable. The* conditional expectation *of Z given X, denoted $E[Z|X]$ is the a.s. unique random variable satisfying the following two conditions.*

1. *$E[Z|X]$ is a measurable function of X.*
2. *$E[E[Z|X]]; \{X \in B\}] = E[Y; \{X \in B\}]$ for any measurable B.*

The uniqueness follows from the following:
Let $h_1(X)$ and $h_2(X)$ be two candidates for $E[Y|X]$. Then, by property 2,

$$E[h_1(X); \{h_1(X) > h_2(X)\}] = E[h_2(X); \{h_1(X) > h_2(X)\}] = E[Y; \{h_1(X) > h_2(X)\}].$$

Thus,

$$0 = E[h_1(X) - h_2(X); \{h_1(X) > h_2(X)\}].$$

Consequently, $P\{h_1(X) > h_2(X)\} = 0$. Similarly, $P\{h_2(X) > h_1(X)\} = 0$ and $h_1(X) = h_2(X)$ a.s.

Existence follows from the Radon-Nikodym theorem. Recall from Chapter 2, that given a measure μ and a nonnegative measurable function h, we can define a new measure ν by

$$\nu(A) = \int_A h(x)\ \mu(dx). \tag{3.1}$$

The Radon-Nikodym theorem answers the question: What conditions must we have on μ and ν so that we can find a function h so that (3.1) holds. In the case of a discrete state space, equation (3.1) has the form

$$\nu(A) = \sum_{x \in A} h(x)\mu\{x\}.$$

For the case A equals a singleton set $\{\tilde{x}\}$, this equation becomes

$$\nu\{\tilde{x}\} = h(\tilde{x})\mu\{\tilde{x}\}.$$

If $\nu\{\tilde{x}\} = 0$, the we can set $h(\tilde{x}) = 0$. Otherwise, we set

$$h(\tilde{x}) = \frac{\nu\{\tilde{x}\}}{\mu\{\tilde{x}\}}.$$

This choice answers the question as long as we do not divide by zero. In other words, we have the condition that $\nu\{\tilde{x}\} > 0$ implies $\mu\{\tilde{x}\} > 0$. Extending this to sets in general, we must have $\nu(A) > 0$ implies $\nu(A) > 0$. Stated in the contrapositive,

$$\mu(A) = 0 \quad \text{implies} \quad \nu(A) = 0. \tag{3.2}$$

.

If any two measures μ and ν have the relationship described by (3.2), we say that ν *is absolutely continuous with respect to* μ and write $\nu << \mu$.

The Radon-Nikodym theorem states that this is the appropriate condition. If $\nu << \mu$, then there exists an integrable function h so that (3.1) holds. In general, one can construct a proof by looking at ratios $\nu(A)/\mu(A)$ for small sets that contain a given point $\tilde{x}$ and try to define h by shrinking these sets down to a point. For this reason, we sometimes write

$$h(\tilde{x}) = \frac{\nu(d\tilde{x})}{\mu(d\tilde{x})}$$

and call h the *Radon-Nikodym derivative*.

Returning the the issue of the definition of conditional expectation, assume that Z is a non-negative random variable and consider the two measures

$$\mu(B) = P\{X \in B\} \text{ and } \nu(B) = E[Z; \{X \in B\}].$$

Then $\nu << \mu$. Thus, by the Radon-Nikodym theorem, there exist a measurable function h so that

$$E[Z; \{X \in B\}] = \nu(B) = \int_B h(x)\, \nu(dx) = E[h(X); \{X \in B\}]$$

and property 2 in the definition of conditional expectation is satisfied and $h(X) = E[Z|X]$.

For an arbitrary integrable Z, consider its positive and negative parts separately.

Often we will write $h(x) = E[Z|X = x]$. Then, for example

$$EY = E[E[Z|X]] = \int E[Z|X = x]\nu(dx).$$

If X is a discrete random varible, then we have $\mu\{x\} = P\{X = x\}, \nu\{x\} = E[Z; \{X = x\}]$ and

$$h(x) = \frac{E[Z; \{X = x\}]}{P\{X = x\}}. \tag{3.3}$$

Definition 3.25. *The* conditional probability $P(A|X) = E[I_A|X]$.

Exercise 3.26. *A random variable is $\sigma(X)$ measurable if and only if it can be written as $h(X)$ for some measurable function h.*

Exercise 3.27. *1. $E[g(X)Z|X] = g(X)E[Z|X]$.*

2. If X and Z are independent, then $E[Z|X] = EY$.

3. Assume that Z is square integrable, then

$$E[Zg(X)] = E[E[Z|X]g(X)]$$

for every square integrable $g(X)$.

We can give a Hilbert space perspective to conditional expectation by writing EX_1X_2 as an inner product $\langle X_1, X_2\rangle$. Then, the identity above becomes

$$\langle E[Z|X], g(X)\rangle = \langle Z, g(X)\rangle \text{ for every } g(X) \in L^2(\Omega, \sigma(X), P).$$

Now consider $L^2(\Omega, \sigma(X), P)$ as a closed subspace of $L^2(\Omega, \mathcal{F}, P)$. Then this identity implies that

$$E[Z|X] = \Pi_X(Z)$$

where Π_X is orthogonal projection onto $L^2(\Omega, \sigma(X), P)$. This can be viewed as a minimization problem

$$\min\{E(Z - h(X))^2; h(X) \in L^2(\Omega, \sigma(X), P)\}.$$

The unique solution occurs by taking $g(X) = E[Z|X]$. In statistics, this is called "least squares".

For the case that $Z = g(X, Y)$ and X takes values on a discrete state space S, then by conditional expectation property 2,

$$\begin{aligned} E[g(X,Y); \{X = x\}] &= E[E[g(X,Y)|X]; \{X = x\}] \\ &= E[h(X); \{X = x\}] \\ &= h(x)P\{X = x\}. \end{aligned}$$

Thus, if $P\{X = x\} > 0$,

$$h(x) = \frac{E[g(X,Y); \{X = x\}]}{P\{X = x\}}$$

as in (3.3).

If, in addition, Y is S-valued and the pair (X, Y) has joint density

$$f_{(X,Y)}(x, y) = P\{X = x, Y = y\}$$

with respect to counting measure on $S \times S$. Then,

$$E[g(X,Y); \{X = x\}] = \sum_{y \in S} g(x, y) f_{(X,Y)}(x, y).$$

Taking $h(x) = E[g(X,Y)|X = x]$, and $f_X(x) = P\{X = x\}$, we then have

$$h(x) = \sum_{y \in S} g(x,y) \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \sum_{y \in S} g(x,y) f_{Y|X}(y|x).$$

Let's see if this definition of h works more generally. Let ν_1 and ν_2 be σ-finite measures and consider the case in which (X,Y) has a density $f_{(X,Y)}$ with respect to $\nu_1 \times \nu_2$. i.e.,

$$P\{(X,Y) \in A\} = \int_A f_{(X,Y)}(x,y)\ (\nu_1 \times \nu_2)(dx \times dy).$$

Then the marginal density $f_X(x) = \int f_{(X,Y)}(x,y)\, \nu_2(dy)$ and the conditional density

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)}.$$

if $f_X(x) > 0$ and 0 if $f_X(x) = 0$. Set

$$h(x) = \int g(x,y) f_{Y|X}(y|x)\, \nu_2(dy).$$

Claim. If $E|g(X,Y)| < \infty$, then $E[g(X,Y)|X] = h(X)$

We only need to show that

$$E[h(X); \{X \in B\}] = E[g(X,Y); \{X \in B\}].$$

Thus,

$$\begin{aligned} E[h(X); \{X \in B\}] &= \int_B h(x) f_X(x)\, \nu_1(dx) \\ &= \int_B (\int g(x,y) f_{Y|X}(y|x)\, \nu_2(dy)) f_X(x)\ \nu_1(dx) \\ &= \int \int g(x,y) I_B(x) f_{(X,Y)}(x,y)\, \nu_2(dy) \nu_1(dx) \\ &= E[g(X,Y); \{X \in B\}]. \end{aligned}$$

Definition 3.28. *The* conditional variance

$$Var(Z|X) = E[(Z - E[Z|X])^2|X] = E[Z^2|X] - (E[Z|X])^2$$

and the conditional covariance

$$Cov(Z_1, Z_2|X) = E[(Z_1 - E[Z_1|X])(Z_2 - E[Z_2|X])|X] = E[Z_1 Z_2|X] - E[Z_1|X]E[Z_2|X].$$

Exercise 3.29. *1. If (X,Y) has joint density $f_{(X,Y)}$ with respect to Lebesgue measure, then*

$$P\{Y \le y|X = x\} = \lim_{h \to 0} P\{Y \le y | x \le X \le x + h\}.$$

2. Show that $E[E[Z|X]] = EZ$ *and* $Var(Z) = E[\,Var(Z|X)] + Var(E[Z|X])$.

Exercise 3.30. *1. (conditional bounded convergence theorem) Let* $\{Z_n; n \geq 1\}$ *be a bounded sequence of random variables that converges to* Z *almost surely, then*

$$\lim_{n\to\infty} E[Z_n|X] = E[Z|X].$$

2. (tower property) $E[E[Z|X_1, X_2]|X_2] = E[Z|X_2]$.

Example 3.31. *Let* X*, a* $Pois(\lambda)$ *random variable, and* Y*, a* $Pois(\mu)$ *random variable be independent, Then*

$$\begin{aligned} f_{X+Y}(z) &= (f_X * f_Y)(z) = \sum_x f_X(x) f_Y(z-x) = \sum_{x=0}^{z} \frac{\lambda^x}{x!} e^{-\lambda} \frac{\mu^{z-x}}{(z-x)!} e^{-\mu} \\ &= \frac{1}{z!} e^{-(\lambda+\mu)} \sum_{x=0}^{z} \frac{z!}{x!(z-x)!} \lambda^x \mu^{z-x} = \frac{(\lambda+\mu)^z}{z!} e^{-(\lambda+\mu)}, \end{aligned}$$

and $Z = X + Y$ *is a* $Pois(\lambda + \mu)$ *random variable. Also*

$$\begin{aligned} f_{X|Z}(x|z) &= \frac{f_{(X,Z)}(x,z)}{f_Z(z)} = \frac{f_{(X,Y)}(x,z-x)}{f_Z(z)} = \frac{f_X(x) f_Y(z-x)}{f_Z(z)} \\ &= \left(\frac{\lambda^x}{x!} e^{-\lambda} \frac{\mu^{z-x}}{(z-x)!} e^{-\mu} \right) \Big/ \left(\frac{(\lambda+\mu)^z}{z!} e^{-(\lambda+\mu)} \right) \\ &= \binom{z}{x} \left(\frac{\lambda}{\lambda+\mu} \right)^x \left(\frac{\mu}{\lambda+\mu} \right)^{(z-x)}. \end{aligned}$$

the distribution of a $Bin(z, \lambda/(\lambda+\mu)$ *random variable.*

Example 3.32. *Let* (X, Y) *have joint density* $f_{(X,Y)}(x, y) = e^{-y},\ 0 < x < y < \infty$ *with respect to Lebesgue measure in the plane. Then the marginal density*

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\ dy = \int_x^{\infty} e^{-y}\ dy = e^{-x}.$$

Thus, X *is an* $Exp(1)$ *random variable. The conditional density is*

$$f_{Y|X}(y|x) = \begin{cases} e^{-(y-x)}, & \text{if } x < y, \\ 0 \qquad\quad , & \text{if } y \geq x. \end{cases}$$

Thus, given that $X = x$*,* Y *is equal to* x *plus an* $Exp(1)$ *random variable. Thus,* $E[Y|X] = X + 1$ *and*$Var(Y|X) = 1$*. Consequently,* $EY = 2$ *and*

$$\begin{aligned} Cov(X,Y) &= E[XY] - (EX \cdot EY) = E[E[XY|X]] - (1 \cdot 2) \\ &= E[XE[Y|X]] - 2 = E[X(X+1)] - 2 = E[X^2] + EX - 2 = 2 + 1 - 2 = 1. \end{aligned}$$

Exercise 3.33. *1. Let* S_m *and* S_n *be independent* $Bin(m, p)$ *and* $Bin(n, p)$ *random variables. Find* $P\{S_m + S_n = y|S_m = x\}$ *and* $P\{S_m = x|s_m + S_n = y\}$.

2. *Let X_1 be uniformly distributed on $[0,1]$ and X_2 be uniformly distributed on $[0,X_2]$. Find the density of X_2. Find the mean and variance of X_2 directly and by using the conditional mean and variance formula.*

3. *Let X be $Pois(\lambda)$ random variable and let Y be a $Bin(X,p)$ random variable. Find the distribution of Y.*

4. *Consider the independent random variables with common continuous distribution F. Show that*

   (a) $P\{X_{(n)} \le x_n, X_{(1)} > x_1\} = (F(x_n) - F(x_1))$, *for* $x_1 < x_n$.

   (b) $P\{X_{(1)} > x_1 | X_{(n)} = x_n\} = ((F(x_n) - F(x_1))/F(x_n))$, *for* $x_1 < x_n$.

   (c)

   $$P\{X_1 \le x | X_{(n)} = x_n\} = \frac{n-1}{n}\frac{F(x)}{F(x_n)}^{n-1}, \text{ for } x \le x_n.$$

   and 1 for $x > x_n$.

   (d)

   $$E[X_1|X_{(n)}] = \frac{n-1}{n}\frac{1}{F(X_{(n)})}\int_{-\infty}^{X_{(n)}} x\, dF(x) + \frac{X_{(n)}}{n}.$$

5. *Consider the density* $f_{(X_1,X_2)}(x_1,x_2)$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\frac{-1}{(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right).$$

   Show that

   (a) $f_{(X_1,X_2)}$ *is a probability density function.*

   (b) X_i *is* $N(\mu_i, \sigma_i^2), i = 1,2$.

   (c) ρ *is the correlation of* X_1 *and* X_2.

   (d) Find $f_{X_2|X_1}$.

   (e) Show that $E[X_2|X_1] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1)$.

3.5 Normal Random Variables

Definition 3.34 (multivariate normal random variables)**.** *Let Q be a $d \times d$ symmetric matrix and let*

$$q(x) = xQx^T = \sum_{i=1}^{d}\sum_{j=1}^{d} x_i q_{ij} x_j$$

be the associated quadratic form. A normal random variable *X on $\mathbb{R}^d$ is defined to be one that has density*

$$f_X(x) \propto \exp\left(-q(x-\mu)/2\right).$$

For the case $d = 2$ we have seen that

$$Q = \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}.$$

Exercise 3.35. *For the quadratic form above, Q is the inverse of the variance matrix $Var(X)$.*

We now look at some of the properties of normal random variables.

- *The collection of normal random variables is closed under invertible affine transformations.*

  If $Y = X - a$, then Y is also normal. Call a normal random variable *centered* if $\mu = 0$.
  Let A be a non-singular matrix and let X be a centered normal. If $Y = XA$ then,

  $$f_Y(y) \propto \exp\left(-yA^{-1}Q(A^{-1})^T y^T/2\right).$$

  Note that $A^{-1}Q(A^{-1})^T$ is symmetric and consequently, Y is normal.

- *The diagonal elements of Q are non-zero.*

  For example, if $q_{dd} = 0$, then we have that the marginal density

  $$f_{X_d}(x_d) \propto \exp(-ax_d + b),$$

  for some $a, b \in \mathbb{R}$. Thus, $\int f_{X_n}(x_n)\, dx_n = \infty$ and f_{X_n} cannot be a density.

- *All marginal densities of a normal density are normal.*

  Consider the invertible transformation

  $$y_1 = x_1, \ldots, y_{d-1} = x_{d-1}, y_d = q_{1d}x_1 + \cdots + q_{dd}x_d.$$

  (We can solve for x_d because $q_{dd} \neq 0$.) Then

  $$A^{-1} = \begin{pmatrix} 1 & 0 & \cdots & -q_{1d}/q_{dd} \\ 0 & 1 & \cdots & -q_{2d}/q_{dd} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/q_{dd} \end{pmatrix}.$$

  Write $\tilde{Q} = A^{-1}Q(A^{-1})^T$. Then

  $$\tilde{q}_{dd} = \sum_{j,k=1}^{d} A^{-1}_{dj} q_{jk} A^{-1}_{dk} = \frac{1}{q_{dd}} q_{dd} \frac{1}{q_{dd}} = \frac{1}{q_{dd}}$$

  and in addition, note that for $i \neq d$,

  $$\tilde{q}_{di} = \sum_{j,k=1}^{d} A^{-1}_{dj} q_{jk} A^{-1}_{ik} = \frac{1}{q_{dd}} \sum_{k=1}^{d} q_{dk} A^{-1}_{ik} = \frac{1}{q_{dd}} \left(q_{di} + q_{dd} \left(-\frac{q_{di}}{q_{dd}} \right) \right) = 0.$$

Consequently,

$$\tilde{q}(y) = \frac{1}{q_{dd}} y_d^2 + \tilde{q}^{(d-1)}(y)$$

where $\tilde{q}^{(d-1)}$ is a quadratic form on $y_1, \ldots, y_{d-1}$. Note that

$$(X_1, \ldots, X_{d-1}) = (Y_1, \ldots, Y_{d-1})$$

to see that it is a normal random variable.

Noting that $q_{dd} > 0$, an easy induction argument yields:

- *There exists a matrix C with positive determinant such that $\tilde{Z} = XC$ in which the components $\tilde{Z}_i$ are independent normal random variables.*
- *Conditional expectations are linear functions.*

$$0 = E[Y_d | Y_1, \ldots, Y_{d-1}] = E[q_{1d} X_1 + \cdots + q_{dd} X_d | X_1, \ldots, X_{d-1}]$$

or

$$E[X_d | X_1, \ldots, X_{d-1}] = \frac{1}{q_{dd}} q_{1d} X_1 + \cdots + q_{d,d-1} X_{d-1}.$$

Thus, the Hilbert space minimization problem for $E[X_d | X_1, \ldots, X_{d-1}]$ reduces to the multidimensional calculus problem for the coefficients of linear function of $X_1, \ldots, X_{d-1}$. This is the basis of least squares linear regression for normal random variables.

- *The quadratic form Q is the inverse of the variance matrix* $\mathrm{Var}(X)$.

Set

$$D = \mathrm{Var}(\tilde{Z}) = C^T \mathrm{Var}(X) C,$$

a diagonal matrix with diagonal elements $\mathrm{Var}(\tilde{Z}_i) = \sigma_i^2$. Thus the quadratic form for the density of Z is

$$\begin{pmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1/\sigma_d^2 \end{pmatrix} = D^{-1}.$$

Write $xC = z$, then the density

$$f_X(x) = |\det(C)| f_{\tilde{Z}}(Cx) \propto \exp(-\frac{1}{2} x^T C^T D^{-1} C x).$$

and

$$\mathrm{Var}(X) = (C^{-1})^T D C^{-1} = Q^{-1}.$$

Now write

$$Z_i = \frac{\tilde{Z}_i - \mu_i}{\sigma_i}.$$

Thus,

- *Every normal random variable is an affine transformation of the vector-valued random variable whose components are independent standard normal random variables.*

  We can use this to extend the definition of normal to X is a d-dimensional normal random variable if and only if

  $$X = ZA + c$$

  for some constant $c \in \mathbb{R}^d$, $d \times r$ matrix A and Z, a collection of r independent standard normal random variables.

  By checking the 2×2 case, we find that:

- *Two normal random variables (X_1, X_2) are independent if and only if $\mathrm{Cov}(X_1, X_2) = 0$, that is, if and only if X_1 and X_2 are uncorrelated.*

We now relate this to the t-distribution.

For independent $N(\mu, \sigma^2)$ random variables $X_1, \cdots, X_n$ write

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

Then, $X_i - \bar{X}$ and $\bar{X}$ together form a bivariate normal random variable. To see that they are independent note that

$$\mathrm{Cov}(X_i - \bar{X}, \bar{X}) = \mathrm{Cov}(X_i, \bar{X}) - \mathrm{Cov}(\bar{X}, \bar{X}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.$$

Thus,

$$\bar{X}, \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are independent.

Exercise 3.36. *Call S^2 the sample variance.*

1. *Check that S^2 is* unbiased*: For X_i independent $N(\mu, \sigma^2)$ random variables, $ES^2 = \sigma^2$.*
2. *Define the T statistic to be*

   $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

   Show that the T statistic is invariant under an affine transformation of the X_i's.
3. *If the X_i's are $N(0,1)$ then $(n-1)S^2$ is χ^2_{n-1}.*

4 Notions of Convergence

In this chapter, we shall introduce a variety of modes of convergence for a sequence of random variables. The relationship among the modes of convergence is sometimes established using some of the inequalities established in the next section.

4.1 Inequalities

Theorem 4.1 (Chebyshev's inequality)**.** *Let $g : \mathbb{R} \to [0.\infty)$ be a measurable function, and set $m_A = \inf\{g(x) : x \in A\}$. Then*

$$m_A P\{X \in A\} \le E[g(X); \{X \in A\}] \le Eg(X).$$

Proof. Note that

$$m_A I_{\{X \in A\}} \le g(X) I_{\{X \in A\}} \le g(X).$$

Now take expectations. □

One typical choice is to take g increasing, and $A = (a, \infty)$, then

$$P\{g(X) > a\} \le \frac{Eg(X)}{g(a)}.$$

For example,

$$P\{|Y - \mu_Y| > a\} = P\{(Y - \mu_Y)^2 > a^2\} \le \frac{\mathrm{Var}(Y)}{a^2}.$$

Exercise 4.2. *1. Prove* Cantelli's inequality.

$$P\{X - \mu > a\} \le \frac{Var(X)}{Var(X) + a^2}.$$

2. Choose X so that its moment generating function is finite in some open interval I containing 0. Then

$$P\{X > a\} = P\{e^{\theta X} > e^{\theta a}\} \le \frac{m(\theta)}{e^{\theta a}}, \quad \theta > 0.$$

Thus,

$$\ln P\{X > a\} \le \inf\{\ln m(\theta) - \theta a; \theta \in (I \cap (0, \infty))\}.$$

Exercise 4.3. *Use the inequality above to find upper bounds for $P\{X > a\}$ where X is normal, Poisson, binomial.*

Definition 4.4. *For an open and convex set $D \in \mathbb{R}^d$, call a function $\phi : D \to \mathbb{R}$* convex *if for every pair of points $x, \tilde{x} \in S$ and every $\alpha \in [0, 1]$*

$$\phi(\alpha x + (1 - \alpha)\tilde{x}) \le \alpha\phi(x) + (1 - \alpha)\phi(\tilde{x}).$$

Exercise 4.5. *Let D be convex. Then ϕ is convex function if and only if the set $\{(x, y); y \ge \phi(x)\}$ is a convex set.*

The definition of ϕ being a convex function is equivalent to the *supporting hyperplane condition*. For every $\tilde{x} \in D$, there exist a linear operator $A(x) : \mathbb{R}^d \to \mathbb{R}$ so that

$$\phi(x) \geq \phi(\tilde{x}) + A(\tilde{x})(x - \tilde{x}).$$

If the choice of $A(\tilde{x})$ is unique, then it is called the tangent hyperplane.

Theorem 4.6 (Jensen's inequality)**.** *Let ϕ be the convex function described above and let X be an D-valued random variable chosen so that each component is integrable and that $E|\phi(X)| < \infty$. Then*

$$E\phi(X) \geq \phi(EX).$$

Proof. Let $\tilde{x} = EX$, then

$$\phi(X(\omega)) \geq \phi(EX) + A(EX)(X(\omega) - EX).$$

Now, take expectations and note that $E[A(EX)(X - EX)] = 0$. □

Exercise 4.7. *1. Show that for ϕ convex, for $\{x_1, \ldots, x_k\} \subset D$, a convex subset of $\mathbb{R}^n$ and for $\alpha_i \geq 0, i = 1, \cdots, k$ with $\sum_{i=1}^k \alpha_i = 1$,*

$$\phi(\sum_{i=1}^k \alpha_i x_i) \leq \sum_{i=1}^k \phi(\alpha_i x_i).$$

2. *Prove the conditional Jensen's inequaltiy: Let ϕ be the convex function described above and let Y be an D-valued random variable chosen so that each component is integrable and that $E|\phi(X)| < \infty$. Then $E[\phi(Y)|X] \geq \phi(E[Y|X])$.*

3. *Let $d = 2$, then show that a function ϕ that has continuous second derivatives is convex if*

$$\frac{\partial^2 \phi}{\partial x_1^2}(x_1, x_2) \geq 0, \quad \frac{\partial^2 \phi}{\partial x_2^2}(x_1, x_2) \geq 0, \quad \frac{\partial^2 \phi}{\partial x_1^2}(x_1, x_2)\frac{\partial^2 \phi}{\partial x_2^2}(x_1, x_2) \geq \frac{\partial^2 \phi}{\partial x_1 \partial x_2}(x_1, x_2)^2.$$

4. *Call L^p the space of measurable functions Z so that $|Z|^p$ is integrable. If $1 \leq q < p < \infty$, then L^p is contained in L^q. In particular show that the function*

$$n(p) = E[|Z|^p]^{1/p}$$

is increasing in p and has limit $ess \sup |Z|$ where $\operatorname{ess\,sup} \mathrm{X} = \inf\{\mathrm{x} : \mathrm{P}\{\mathrm{X} \leq \mathrm{x}\} = 1\}$.

5. *(Hölder's inequality). Let X and Y be non-negative random variables and show that $E[X^{1/p}Y^{1/q}] \leq (EX)^{1/p}(EY)^{1/q}$, $p^{-1} + q^{-1} = 1$.*

6. *(Minkowski's inequality). Let X and Y be non-negative random variables and let $p \geq 1$. Show that $E[(X^{1/p} + Y^{1/p})^p] \leq ((EX)^{1/p} + (EY)^{1/p})^p$. Use this to show that $||Z||_p = E[|Z|^p]^{1/p}$ is a norm.*

4.2 Modes of Convergence

Definition 4.8. *Let $X, X_1, X_2, \cdots$ be a sequence of random variables taking values in a metric space S with metric d.*

1. *We say that* X_n converges to X almost surely *($X_n \to^{a.s.} X$) if*

$$\lim_{n\to\infty} X_n = X \qquad a.s..$$

2. *We say that* X_n converges to X in L^p, $p > 0$, *($X_n \to^{L^p} X$) if,*

$$\lim_{n\to\infty} E[d(X_n, X)^p] = 0.$$

3. *We say that* X_n converges to X in probability *($X_n \to^{P} X$) if, for every $\epsilon > 0$,*

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

4. *We say that* X_n converges to X in distribution *($X_n \to^{\mathcal{D}} X$) if, for every bounded continuous $h : S \to \mathbb{R}$.*

$$\lim_{n\to\infty} Eh(X_n) = Eh(X).$$

Convergence in distribution differs from the other modes of convergence in that it is based not on a direct comparison of the random variables X_n with X but rather on a comparision of the distributions $\mu_n(A) = P\{X_n \in A\}$ and $\mu(A) = P\{X \in A\}$. Using the change of variables formula, convergence in distribution can be written

$$\lim_{n\to\infty} \int h \, d\mu_n = \int h \, d\mu.$$

Thus, it investigates the behavior of the distributions $\{\mu_n : n \geq 1\}$ using the continuous bounded functions as a class of test functions.

Exercise 4.9. *1. $X_n \to^{a.s.} X$ implies $X_n \to^{P} X$.*

*(*Hint: *Almost sure convergence is the same as $P\{d(X_n, X) > \epsilon \text{ i.o.}\} = 0$.)*

2. $X_n \to^{L^p} X$ implies $X_n \to^{P} X$.

3. Let $p > q$, then $X_n \to^{L^p} X$ then $X_n \to^{L^q} X$.

Exercise 4.10. *Let $g : S \to \mathbb{R}$ be continuous. Then*

1. *$X_n \to^{a.s.} X$ implies $g(X_n) \to^{a.s.} g(X)$*
2. *$X_n \to^{\mathcal{D}} X$ implies $g(X_n) \to^{\mathcal{D}} g(X)$*
3. *$X_n \to^{a.s.} X$ implies $X_n \to^{\mathcal{D}} X$.*

We would like to show that the same conclusion hold for convergence in probability.

Theorem 4.11 (first Borel-Cantelli lemma)**.** *Let $\{A_n : n \geq 1\} \subset \mathcal{F}$, if*

$$\sum_{n=1}^{\infty} P(A_n) < \infty \quad \textit{then} \quad P(\limsup_{n\to\infty} A_n) = 0.$$

Proof. For any $m \in N$

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

Let $\epsilon > 0$, then, by hypothesis, this sum can be made to be smaller than ϵ with an appropriate choice of m. □

Theorem 4.12. *If $X_n \to^P X$, then there exists a subsequence $\{n_k : k \geq 1\}$ so that $X_{n_k} \to^{a.\ s.} X$.*

Proof. Let $\epsilon > 0$. Choose $n_k > n_{k-1}$ so that

$$P\{d(X_{n_k}, X) > 2^{-k}\} < 2^{-k}.$$

Then, by the first Borel-Cantelli lemma,

$$P\{d(X_{n_k}, X) > 2^{-k} \text{ i.o.}\} = 0.$$

The theorem follows upon noting that $\{d(X_{n_k}, X) > \epsilon \text{ i.o.}\} \subset \{d(X_{n_k}, X) > 2^{-k} \text{ i.o.}\}$. □

Exercise 4.13. *Let $\{a_n; n \geq 1\}$ be a sequence of real numbers. Then*

$$\lim_{n\to\infty} a_n = L$$

if and only if for every subsequence of $\{a_n; n \geq 1\}$ there exist a further subsequence that converges to L.

Theorem 4.14. *Let $g : S \to \mathbb{R}$ be continuous. Then $X_n \to^P X$ implies $g(X_n) \to^P g(X)$.*

Proof. Any subsequence $\{X_{n_k}; k \geq 1\}$ converges to X in probability. Thus, by the theorem above, there exists a further subsequence $\{X_{n_k(m)}; m \geq 1\}$ so that $X_{n_k(m)} \to^{a.s.} X$. Then $g(X_{n_k(m)}) \to^{a.s.} g(X)$ and consequently $g(X_{n_k(m)}) \to^P g(X)$. □

If we identify versions of a random variable, then we have the L^p-norm for real valued random variables

$$||X||^p = E[|X|^p]^{1/p}.$$

The triangle inequality is given by Minkowski's inequality. This gives rise to a metric via $\rho_p(X, Y) = ||X - Y||_p$.

Convergence in probability is also a metric convergence.

Theorem 4.15. *Let X, Y be random variables with values in a metric space (S, d) and define*

$$\rho_0(X, Y) = \inf\{\epsilon > 0 : P\{d(X, Y) > \epsilon\} < \epsilon\}.$$

Then ρ_0 is a metric.

Proof. If $\rho_0(X,Y) > 0$, then $X \neq Y$.

$$P\{d(X,X) > \epsilon\} = 0 < \epsilon.$$

Thus, $\rho_0(X,X) = 0$.

Because d is symmetric, so is ρ_0. To establish the triangle inequality, note that

$$\{d(X,Y) \leq \epsilon_1\} \cap \{d(Y,Z) \leq \epsilon_2\} \subset \{d(X,Z) \leq \epsilon_1 + \epsilon_2\}$$

or, by writing the complements,

$$\{d(X,Z) > \epsilon_1 + \epsilon_2\} \subset \{d(X,Y) > \epsilon_1\} \cup \{d(Y,Z) > \epsilon_2\}.$$

Thus,

$$P\{d(X,Z) > \epsilon_1 + \epsilon_2\} \leq P\{d(X,Y) > \epsilon_1\} + P\{d(Y,Z) > \epsilon_2\}.$$

So, if $\epsilon_1 > \rho_0(X,Y)$ and $\epsilon_2 > \rho_0(Y,Z)$ then

$$P\{d(X,Y) > \epsilon_1\} < \epsilon_1 \text{ and } P\{d(Y,Z) > \epsilon_2\} < \epsilon_2$$

then

$$P\{d(X,Z) > \epsilon_1 + \epsilon_2\} < \epsilon_1 + \epsilon_2.$$

and, consequently, $\rho_0(X,Z) \leq \epsilon_1 + \epsilon_2$. Thus,

$$\rho_0(X,Z) \leq \inf\{\epsilon_1 + \epsilon_2; \epsilon_1 > \rho_0(X,Y), \epsilon_2 > \rho_0(Y,Z)\} = \rho_0(X,Y) + \rho_0(Y,Z).$$

□

Exercise 4.16. *1. $X_n \to^P X$ if and only if $\lim_{n\to\infty} \rho_0(X_n, X) = 0$.*

2. Let $c > 0$. Then $X_n \to^P X$ if and only if

$$\lim_{n\to\infty} E[\max\{d(X_n, X), c\}] = 0.$$

We shall explore more relationships in the different modes of convergence using the tools developed in the next section.

4.3 Uniform Integrability

Let $\{X_k, k \geq 1\}$ be a sequence of random variables converging to X almost surely. Then by the bounded convergence theorem, we have for each fixed n that

$$E[|X|; \{X < n\}] = \lim_{k\to\infty} E[|X_k; \{X_k < n\}].$$

By the dominated convergence theorem,

$$E|X| = \lim_{n\to\infty} E[|X|; \{X < n\} = \lim_{n\to\infty} \lim_{k\to\infty} E[|X_k|; \{X_k < n\}.$$

If we had a sufficient condition to reverse the order of the double limit, then we would have, again, by the dominated convergence theorem that

$$E|X| = \lim_{k\to\infty} \lim_{n\to\infty} E[|X_k|; \{X_k < n\}] = \lim_{k\to\infty} E[|X_k|].$$

In other words, we would have convergence of the expectations. The uniformity we require to reverse this order is the subject of this section.

Definition 4.17. *A collection of real-valued random variables* $\{X_\lambda; \lambda \in \Lambda\}$ *is* uniformly integrable *if*

1. $\sup_{\lambda\in\Lambda} E|X_\lambda| < \infty$, *and*
2. *for every* $\epsilon > 0$, *there exists a* $\delta > 0$ *such that for every* λ,

$$P(A_\lambda) < \delta \qquad \textit{implies} \qquad |E[X_\lambda; A_\lambda]| < \epsilon.$$

Exercise 4.18. *The criterion above is equivalent to the seemingly stronger condition:*

$$P(A_\lambda) < \delta \qquad \textit{implies} \qquad E[|X_\lambda|; A_\lambda] < \epsilon.$$

Consequently, $\{X_\lambda : \lambda \in \Lambda\}$ is uniformly integrable if and only if $\{|X_\lambda| : \lambda \in \Lambda\}$ is uniformly integrable.

Theorem 4.19. *The following are equivalent:*

1. $\{X_\lambda : \lambda \in \Lambda\}$ *is uniformly integrable.*
2. $\lim_{n\to\infty} \sup_{\lambda\in\Lambda} E[|X_\lambda|; \{|X_\lambda| > n\}] = 0.$
3. $\lim_{n\to\infty} \sup_{\lambda\in\Lambda} E[|X_\lambda| - \min\{n, |X_\lambda|\}] = 0.$
4. *There exists an increasing convex function* $\phi : [0,\infty) \to R$ *such that* $\lim_{x\to\infty} \phi(x)/x = \infty$, *and*

$$\sup_{\lambda\in\Lambda} E[\phi(|X_\lambda|)] < \infty.$$

Proof. $(1 \to 2)$ Let $\epsilon > 0$ and choose δ as defined in the exercise. Set $M = \sup_\lambda E|X_\lambda|$, choose $n > M/\delta$ and define $A_\lambda = \{|X_\lambda| > n\}$. Then by Chebyshev's inequality,

$$P(A_\lambda) \le \frac{1}{n} E|X_\lambda| \le \frac{M}{n} < \delta.$$

$(2 \to 3)$ Note that,

$$nP\{|X_\lambda| > n\} \le E[|X_\lambda|; \{|X_\lambda| > n\}]$$

Therefore,

$$\begin{aligned} |E[|X_\lambda| - \min\{n, |X_\lambda|\}]| &= |E[|X_\lambda| - n; |X_\lambda| > n]| \\ &= |E[|X_\lambda|; |X_\lambda| > n]| - nP\{|X_\lambda| > n\}| \\ &\le 2E[|X_\lambda|; \{|X_\lambda| > n\}]. \end{aligned}$$

$(3 \to 1)$ If n is sufficiently large,

$$M = \sup_{\lambda \in \Lambda} E[|X_\lambda| - \min\{n, |X_\lambda|\}] < \infty$$

and consequently

$$\sup_{\lambda \in \Lambda} E|X_\lambda| \leq M + n < \infty.$$

If $P(A_\lambda) < 1/n^2$, then

$$E[|X_\lambda|; A_\lambda] \leq E[|X_\lambda| - \min\{n, |X_\lambda|\}] + n; A_\lambda] \leq E[|X_\lambda| - \min\{n, |X_\lambda|\}] + nP(A_\lambda) \leq E[|X_\lambda| - \min\{n, |X_\lambda|\}] + \frac{1}{n}.$$

For $\epsilon > 0$, choose n so that the last term is less than ϵ, then choose $\delta < 1/n^2$.

$(4 \to 2)$ By subracting a constant, we can assume that $\phi(0) = 0$. Then, by the convexity of ϕ, $\phi(x)/x$ is increasing. Let $\epsilon > 0$ and let $M = \sup_{\lambda \in \Lambda} E[\phi(|X_\lambda|)]$. Choose N so that

$$\frac{\phi(n)}{n} > \frac{M}{\epsilon} \quad \text{whenever } n \geq N.$$

If $x > n$,

$$\frac{\phi(x)}{x} \geq \frac{\phi(n)}{n}, \quad x \leq \frac{n\phi(x)}{\phi(n)}.$$

Therefore,

$$E[|X_\lambda|; \{|X_\lambda| > n\}] \leq \frac{nE[\phi(|X_\lambda|); |X_\lambda| > n]}{\phi(n)} \leq \frac{nE[\phi(|X_\lambda|)]}{\phi(n)} \leq \frac{nM}{\phi(n)} < \epsilon.$$

$(2 \to 4)$ Choose a decreasing sequence $\{a_k : k \geq 1\}$ of positive numbers so that $\sum_{k=1}^{\infty} k a_k < \infty$. By 2, we can find a strictly increasing sequence $\{n_k : k \geq 1\}$ satisfying $n_0 = 0$.

$$\sup_{\lambda \in \Lambda} E[|X_\lambda|; \{|X_\lambda| > n_k\}] \leq a_k.$$

Define ϕ by $\phi(0) = 0$, $\phi'(0) = 0$ on $[n_0, n_1)$ and

$$\phi'(x) = k - \frac{n_{k+1} - x}{n_{k+1} - n_k}, \ x \in [n_k, n_{k+1}).$$

On this interval, ϕ' increases from $k - 1$ to k.

Because ϕ is convex, the slope of the tangent at x is greater than the slope of the secant line between $(x, \phi(x))$ and $(0, 0)$, i.e,

$$\frac{\phi(x)}{x} \leq \phi'(x) \leq k \quad \text{for } x \in [n_k, n_{k+1}).$$

Thus,

$$\phi(x) \leq kx \quad \text{for } x \in [n_k, n_{k+1}).$$

Consequently,

$$\sup_{\lambda \in \Lambda} E[\phi(|X_\lambda|)] = \sup_{\lambda \in \Lambda} \sum_{k=1}^{\infty} E[\phi(|X_\lambda|); n_{k+1} \geq |X_\lambda| > n_k] \leq \sup_{\lambda \in \Lambda} \sum_{k=1}^{\infty} kE[|X_\lambda|; \{|X_\lambda| \geq n_k\}] < \infty.$$

□

Exercise 4.20. *1. If a collection of random variables is bounded in $L^p, p > 1$, then it is uniformly integrable.*

2. *A finite collection of integrable random variables is uniformly integrable.*

3. *If $|X_\lambda| \leq Y_\lambda$ and $\{Y_\lambda; \lambda \in \Lambda\}$ is uniformly integrable, then so is $\{X_\lambda; \lambda \in \Lambda\}$.*

4. *If $\{X_\lambda : \lambda \in \Lambda\}$ and $\{Y_\lambda : \lambda \in \Lambda\}$ are uniformly integrable, then so is $\{X_\lambda + Y_\lambda : \lambda \in \Lambda\}$.*

5. *Assume that Y is integrable and that $\{X_\lambda; \lambda \in \Lambda\}$ form a collection of real valued random variables, then $\{E[Y|X_\lambda] : \lambda \in \Lambda\}$ is uniformly integrable.*

6. *Assume that $\{X_n : n \geq 1\}$ is a uniformly integrable sequence and define $\bar{X}_n = (X_1 + \cdots + X_n)/n$, then $\{\bar{X}_n : n \geq 1\}$ is a uniformly integrable sequence*

Theorem 4.21. *If $X_k \to^{a.s.} X$ and $\{X_k; k \geq 1\}$ is uniformly integrable, then $\lim_{k\to\infty} EX_k = EX$.*

Proof. Let $\epsilon > 0$ and write

$$\begin{aligned}(E|X_k| - E|X|) &= (E[|X_k| - \max\{|X_k|, n\}] \\ &\quad -E[|X| - \max\{|X|, n\}]) \\ &\quad +(E[\max\{|X_k|, n\}] - E[\max\{|X|, n\}]).\end{aligned}$$

If $\{X_k; k \geq 1\}$ is uniformly integrable, then by the appropriate choice on N, the first term on the right can be made to have absolutely value less than $\epsilon/3$ uniformly in k for all $n \geq N$. The same holds for the second term by the integrability of X. Note that the function $f(x) = \max\{|x|, n\}$ is continuous and bounded and therefore, because almost sure convergence implies convergence in distribution, the last pair of terms can be made to have absolutely value less than $\epsilon/3$ for k sufficiently large. This proves that $\lim_{n\to\infty} E|X_n| = E|X|$.

Now, the theorem follows from the dominated convergence theorem. □

Corollary 4.22. *If $X_k \to^{a.s.} X$ and $\{X_k; k \geq 1\}$ is uniformly integrable, then $\lim_{k\to\infty} E|X_k - X| = 0$.*

Proof. Use the facts that $|X_k - X| \to^{a.s.} 0$, and $\{|X_k - X|; k \geq 1\}$ is uniformly integrable in the theorem above. □

Theorem 4.23. *If the X_k are integrable, $X_k \to^{\mathcal{D}} X$ and $\lim_{k\to\infty} E|X_k| = E|X|$, then $\{X_k; k \geq 1\}$ is uniformly integrable.*

Proof. Note that

$$\lim_{k\to\infty} E[|X_k| - \min\{|X_k|, n\}] = E[|X| - \min\{|X|, n\}].$$

Choose N_0 so that the right side is less than $\epsilon/2$ for all $n \geq N_0$. Now choose K so that $|E[|X_k| - \min\{|X_k|, n\}]| < \epsilon$ for all $k > K$ and $n \geq N_0$. Because the finite sequence of random variables $\{X_1, \ldots, X_K\}$ is uniformly integrable, we can choose N_1 so that

$$E[|X_k| - \min\{|X_k|, n\}] < \epsilon$$

for $n \geq N_1$ and $k < K$. Finally take $N = \max\{N_0, N_1\}$. □

Taken together, for a sequence $\{X_n : n \geq 1\}$ of integrable real valued random variables satisfying $X_n \to^{a.s.} X$, the following conditions are equivalent:

1. $\{X_n : n \geq 1\}$ is uniformly integrable.
2. $E|X| < \infty$ and $X_n \to^{L^1} X$.
3. $\lim_{n\to\infty} E|X_n| = E|X|$.

5 Laws of Large Numbers

Definition 5.1. *A* stochastic process X *(or a* random process*, or simply a process) with index set* Λ *and a measurable state space* $(S, \mathcal{B})$ *defined on a probability space* $(\Omega, \mathcal{F}, P)$ *is a function*

$$X : \Lambda \times \Omega \to S$$

such that for each $\lambda \in \Lambda$*,*

$$X(\lambda, \cdot) : \Omega \to S$$

is an S*-valued random variable.*

Note that Λ is not given the structure of a measure space. In particular, it is not necessarily the case that X is measurable. However, if Λ is countable and has the power set as its σ-algebra, then X is automatically measurable.

$X(\lambda, \cdot)$ is variously written $X(\lambda)$ or X_λ. Throughout, we shall assume that S is a metric space with metric d.

Definition 5.2. *A* realization *of* X *or a* sample path *for* X *is the function*

$$X(\cdot, \omega_0) : \Lambda \to S \quad \textit{for some } \omega_0 \in \Omega.$$

Typically, for the processes we study Λ will be the natural numbers, and $[0, \infty)$. Occasionally, Λ will be the integers or the real numbers. In the case that Λ is a subset of a multi-dimensional vector space, we often call X a *random field*.

The laws of large numbers state that somehow a statistical average

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

is near their common mean value. If near is measured in the almost sure sense, then this is called a *strong law*. Otherwise, this law is called a *weak law*.

In order for us to know that the stong laws have content, we must know when there is a probability measure that supports, in an appropriate way, the distribution of a sequence of random variable, $X_1, X_2, \ldots$. That is the topic of the next section.

5.1 Product Topology

A function

$$x : \Lambda \to S$$

can also be considered as a point in a product space,

$$x = \{x_\lambda : \lambda \in \Lambda\} \in \prod_{\lambda \in \Lambda} S_\lambda.$$

with $S_\lambda = S$ for each $\lambda \in \Lambda$.

One of simplest questions to ask of this set is to give its value for the λ_0 coordinate. That is, to evaluate the function

$$\pi_{\lambda_0}(x) = x_{\lambda_0}.$$

In addition, we will ask that this evaluation function π_{λ_0} be continuous. Thus, we would like to place a topology on $\prod_{\lambda\in\Lambda} S_\lambda$ to accomodate this. To be precise, let $\mathcal{O}_\lambda$ be the open subsets of S_λ. We want

$$\pi_\lambda^{-1}(U) \text{ to be an open set for any } U \in \mathcal{O}_\lambda$$

Let $F \subset \Lambda$ be a finite subset, $U_\lambda \in \mathcal{O}_\lambda$ and $\pi_F : \prod_{\lambda\in\Lambda} S_\lambda \to \prod_{\lambda\in F} S_\lambda$ evaluation on the coordinates in F. Then, the topology on $\prod_{\lambda\in\Lambda} S_\lambda$ must contain

$$\pi_F^{-1}(\prod_{\lambda\in F} U_\lambda) = \bigcap_{\lambda\in F} \pi_\lambda^{-1}(U_\lambda) = \{x : x_\lambda \in U_\lambda \text{ for } \lambda \in F\} = \prod_{\lambda\in\Lambda} U_\lambda$$

where $U_\lambda \in \mathcal{O}_\lambda$ for all $\lambda \in \Lambda$ and $U_\lambda = S_\lambda$ for all $\lambda \notin F$.

This collection

$$\mathcal{Q} = \{\prod_{\lambda\in\Lambda} U_\lambda : U_\lambda \in \mathcal{O}_\lambda \text{ for all } \lambda \in \Lambda,\ U_\lambda = S_\lambda \text{ for all } \lambda \notin F\}.$$

forms a *basis* for the *product topology* on $\prod_{\lambda\in\Lambda} S$. Thus, every open set in the product topology is the arbitrary union of open sets in $\mathcal{Q}$. From this we can define the Borel σ-algebra as $\sigma(\mathcal{Q})$.

Note that $\mathcal{Q}$ is closed under the finite union of sets. Thus, the collection $\tilde{\mathcal{Q}}$ obtained by replacing the open sets above in S_λ with measurable sets in S_λ is an algebra. Such a set

$$\{x : x_{\lambda_1} \in B_1, \ldots, x_{\lambda_n} \in B_n\}, \quad B_i \in \mathcal{B}(S_{\lambda_i}), \quad F = \{\lambda_1 \ldots, \lambda_n\},$$

is called an *F-cylinder set* or a *finite dimensional set* having dimension $|F| = n$. Note that if $F \subset \tilde{F}$, then any F-cylinder set is also an $\tilde{F}$-cylinder set.

5.2 Daniell-Kolmogorov Extension Theorem

The Daniell-Kolmogorov extension theorem is the precise articulation of the statement: "The finite dimensional distributions determine the distribution of the process."

Theorem 5.3 (Daniell-Kolmogorov Extension)**.** *Let $\mathcal{E}$ be an algebra of cylinder sets on $\prod_{\lambda\in\Lambda} S_\lambda$. For each finite subset $F \subset \Lambda$, let R_F be a countably additive set function on $\pi_F(\mathcal{E})$, a collection of subsets of $\prod_{\lambda\in F} S_\lambda$ and assume that the collection of R_F satisfies the compatibility condition:*

For any F-cylinder set E, and any $\tilde{F} \supset F$,

$$R_F(\pi_F(E)) = R_{\tilde{F}}(\pi_{\tilde{F}}(E))$$

Then there exists a unique measure P on $(\prod_{\lambda\in\Lambda} S_\lambda, \sigma(\mathcal{E}))$ so that for any F cylinder set E,

$$P(E) = R_F(\pi_F(E)).$$

Proof. The compatibility condition guarantees us that P is defined in $\mathcal{E}$. To prove that P is countably additive, it suffices to show for every decreasing sequence $\{C_n : n \geq 1\} \subset \mathcal{E}$ that $\lim_{n\to\infty} C_n = \emptyset$ implies

$$\lim_{n\to\infty} P(C_n) = 0.$$

We show the contrapositive by showing that

$$\lim_{n\to\infty} P(C_n) = \epsilon > 0$$

implies $\lim_{n\to\infty} C_n \neq \emptyset$

Each R_F can be extended to a unique probability measure P_F on $\sigma(\pi(E))$. Note that because the C_n are decreasing, they can be viewed as cylinder sets of nondecreasing dimension. Thus, by perhaps repeating some events or by viewing an event C_n as a higher dimensional cylinder set, we can assume that C_n is an F_n-cylinder set with $F_n = \{\lambda_1, \ldots, \lambda_n\}$, i.e.,

$$C_n = \{x : x_{\lambda_1} \in \tilde{C}_{1,n}, \ldots, x_{\lambda_n} \in \tilde{C}_{n,n}\}.$$

Define

$$Y_{n,n}(x_{\lambda_1}, \ldots, x_{\lambda_n}) = I_{C_n}(x) = \prod_{k=1}^{n} I_{\tilde{C}_{k,n}}(x_{\lambda_k})$$

and for $m < n$, use the probability P_{F_n} to take the conditional expectation over the first m coordinates to define

$$Y_{m,n}(x_{\lambda_1}, \ldots, x_{\lambda_m}) = E_{F_n}[Y_{n,n}(x_{\lambda_1}, \ldots, x_{\lambda_n})|x_{\lambda_1}, \ldots, x_{\lambda_m}].$$

Use the tower property to obtain the identity

$$\begin{aligned} Y_{m-1,n}(x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}) &= E_{F_n}[Y_{n,n}(x_{\lambda_1}, \ldots, x_{\lambda_n})|x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}] \\ &= E_{F_n}[E_{F_n}[Y_{n,n}(x_{\lambda_1}, \ldots, x_{\lambda_n})|x_{\lambda_1}, \ldots, x_{\lambda_m}]|x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}] \\ &= E_{F_n}[Y_{m,n}(x_{\lambda_1}, \ldots, x_{\lambda_m})|x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}] \end{aligned}$$

Conditional expectation over none of the coordinates yields $Y_{0,n} = P(C_n)$.

Now, note that $\tilde{C}_{k,n+1} \subset \tilde{C}_{k,n}$. Consequently,

$$\begin{aligned} Y_{m,n+1}(x_{\lambda_1}, \ldots, x_{\lambda_m}) &= E_{F_{n+1}}[\prod_{k=1}^{n+1} I_{\tilde{C}_{k,n+1}}(x_{\lambda_k})|x_{\lambda_1}, \ldots, x_{\lambda_m}] \\ &\leq E_{F_{n+1}}[\prod_{k=1}^{n} I_{\tilde{C}_{k,n}}(x_{\lambda_k})|x_{\lambda_1}, \ldots, x_{\lambda_m}] \\ &= E_{F_n}[\prod_{k=1}^{n} I_{\tilde{C}_{k,n}}(x_{\lambda_k})|x_{\lambda_1}, \ldots, x_{\lambda_m}] \\ &= Y_{m,n}(x_{\lambda_1}, \ldots, x_{\lambda_m}) \end{aligned} \tag{5.1}$$

The compatible condition allow us to change the probability from $P_{F_{n+1}}$ to P_{F_n} in the second to last inequality.

Therefore, this sequence, decreasing in n for each value of $(x_{\lambda_1}, \ldots, x_{\lambda_m})$ has a limit,

$$Y_m(x_{\lambda_1}, \ldots, x_{\lambda_m}) = \lim_{n\to\infty} Y_{m,n}(x_{\lambda_1}, \ldots, x_{\lambda_m}).$$

Now apply the conditional bounded convergence theorem to (5.1) with $n = m$ to obtain

$$Y_{m-1}(x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}) = E_{F_m}[Y_m(x_{\lambda_1}, \ldots, x_{\lambda_m})|x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}]. \tag{5.2}$$

The random variable $Y_m(x_{\lambda_1}, \ldots, x_{\lambda_m})$ cannot be for all values strictly below $Y_{m-1}(x_{\lambda_1}, \ldots, x_{\lambda_{m-1}})$, its conditional mean. Therefore, identity (5.2) cannot hold unless, for every choice of $(x_{\lambda_1}, \ldots, x_{\lambda_{m-1}})$, there exists x_{λ_m} so that

$$Y_m(x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}, x_{\lambda_m}) \geq Y_{m-1}(x_{\lambda_1}, \ldots, x_{\lambda_{m-1}}).$$

Now, choose a sequence $\{x^*_{\lambda_m} : m \geq 1\}$ for which this inequality holds and choose $x^* \in \prod_{\lambda\in\Lambda} S_\lambda$ with λ_m-th coordinate equal to $x^*_{\lambda_m}$. Then,

$$I_{C_n}(x^*) = Y_{n,n}(x^*_{\lambda_1}, \ldots, x^*_{\lambda_n}) \geq Y_n(x^*_{\lambda_1}, \ldots, x^*_{\lambda_n}) \geq Y_0 = \lim_{n\to\infty} P(C_n) > 0.$$

Therefore, $I_{C_n}(x^*) = 1$ and $x^* \in C_n$ for every n. Consequently, $\lim_{n\to\infty} C_n \neq \emptyset$. □

Exercise 5.4. *Consider the S_λ-valued random variables X_λ with distribution ν_λ. Then the case of independent random variable on the product space is obtained by taking*

$$R_F = \prod_{\lambda\in F} \nu_\lambda.$$

Check that the conditions of the Daniell-Kolmogorov extension theorem are satisfied.

In addition, we know have:

Theorem 5.5. *Let $\{X_\lambda; \lambda \in \Lambda\}$ to be independent random variable. Write $\Lambda = \Lambda_1 \cup \Lambda_2$, with $\Lambda_1 \cap \Lambda_2 = \emptyset$, then*

$$\mathcal{F}_1 = \sigma\{X_\lambda : \lambda \in \Lambda_1\} \text{ and } \mathcal{F}_2 = \sigma\{X_\lambda : \lambda \in \Lambda_2\}$$

are independent.

This removes the restriction that Λ be finite. With the product topology on $\prod_{\lambda\in\Lambda} S_\lambda$, we see that this improved theorem holds with the same proof.

Definition 5.6 (canonical space)**.** *The distribution ν of any S-valued random variable can be realized by having the probability space be $(S, \mathcal{B}, \nu)$ and the random variable be the x variable on S. This is called the* canonical space.

Similarly, the Daniell-Kolmogorov extension theorem finds a measure on the canonical space S^Λ so that the random process is just the variable x.

For a countable Λ, this is generally satisfactory. For example, in the strong law of large numbers, we have that

$$\frac{1}{n}\sum_{k=1}^{n} X_k$$

is measurable. However, for $\Lambda = [0, \infty)$, the corresponding limit of averages

$$\frac{1}{N} \int_0^N X_\lambda \, d\lambda$$

is not necessarily measurable. Consequently, we will look to place the probability for the stochastic process on a space of continuous fuunctions or right continuous functions to show that the sample paths have some regularity.

5.3 Weak Laws of Large Numbers

We begin with an L^2-weak law.

Theorem 5.7. *Assume that $X_1, X_2, \ldots$ for a sequence of real-valued uncorrelated random variable with common mean μ. Futher assume that their variances are bounded by some constant C. Write*

$$S_n = X_1 + \cdots + X_n.$$

Then

$$\frac{1}{n} S_n \to^{L^2} \mu.$$

Proof. Note that $E[S_n/n] = \mu$. Then

$$E[(\frac{1}{n} S_n - \mu)^2] = \mathrm{Var}(\frac{1}{n} S_n) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) \leq \frac{1}{n^2} Cn.$$

Now, let $n \to \infty$ □

Because L^2 convergence implies convergence in probability, we have, in addition,

$$\frac{1}{n} S_n \to^{P} \mu.$$

Note that this result does not require the Daniell-Kolmogorov extension theorem. For each n, we can evaluate the the variance of S_n on a probability space that contains the random variables $(X_1, \ldots, X_n)$.

Many of the classical limit theorems begin with *triangular arrays*, a doubly indexed collection

$$\{X_{n,k}; 1 \leq n, 1 \leq k \leq k_n\}.$$

For the classical laws of large numbers, $X_{nk} = X_k/n$ and $k_n = n$.

Exercise 5.8. *For the triangular array $\{X_{n,k}; 1 \leq n, 1 \leq k \leq k_n\}$. Let $S_n = X_{n,1} + \cdots + X_{n,k_n}$ be the n-th row rum. Assume that $ES_n = \mu_n$ and that $\sigma_n^2 = Var(S_n)$. If*

$$\frac{\sigma_n^2}{b_n^2} \to 0 \text{ then } \frac{S_n - \mu_n}{b_n} \to^{L^2} 0.$$

Example 5.9. *1. (*Coupon Collectors Problem) *Let $Y_1, Y_2, \ldots,$ be independent random variables uniformly distributed on $\{1, 2, \ldots, n\}$ (sampling with replacement). Define the random sequence $T_{n,k}$ to be minimum time m such that the cardinality of the range of $(Y_1, \ldots, Y_m)$ is k. Thus, $T_{n,0} = 0$. Define the triangular array*

$$X_{n,k} = T_{n,k} - T_{n,k-1},\ k = 1, \ldots, n.$$

For each n, $X_{k,n} - 1$ are independent $Geo(1 - (k-1)/n)$ random variables. Therefore

$$EX_{n,k} = (1 - \frac{k-1}{n})^{-1} = \frac{n}{n-k-1}, \quad Var(X_{n,k}) = \frac{(k-1)/n}{((n-k-1)/n)^2}.$$

Consequently, for $T_{n,n}$ the first time that all numbers are sampled,

$$ET_{n,n} = \sum_{k=1}^{n} \frac{n}{n-k-1} = \sum_{k=1}^{n} \frac{n}{k} \approx n \log n, \quad Var(T_{n,n}) = \sum_{k=1}^{n} \frac{(k-1)/n}{((n-k-1)/n)^2} \leq \sum_{k=1}^{n} \frac{n^2}{k^2}.$$

By taking $b_n = n \log n$, we have that

$$\frac{T_{n,n} - \sum_{k=1}^{n} \frac{n}{k}}{n \log n} \rightarrow^{L^2} 0$$

and

$$\frac{T_{n,n}}{n \log n} \rightarrow^{L^2} 1.$$

2. We can sometimes have an L^2 law of large numbers for correlated random variables if the correlation is sufficiently weak. Consider r balls to be placed at random into n urns. Thus each configuration has probability n^{-r}. Let N_n be the number of empty urns. Set the triangular array

$$X_{n,k} = I_{A_{n,k}}$$

where $A_{n,k}$ is the event that the k-th of the n urns is empty. Then,

$$N_n = \sum_{k=1}^{n} X_{n,k}.$$

Note that

$$EX_{n,k} = P(A_{n,k}) = (1 - \frac{1}{n})^r.$$

Consider the case that both n and r tend to ∞ so that $r/n \rightarrow c$. Then,

$$EX_{n,k} \rightarrow e^{-c}.$$

For the variance $Var(N_n) = EN_n^2 - (EN_n)^2$ and

$$EN_n^2 = E\left(\sum_{k=1}^{n} X_{k,n}\right)^2 = \sum_{j=1}^{n}\sum_{k=1}^{n} P(A_{n,j} \cap A_{n,k}).$$

The case $j = k$ is computed above. For $j \neq k$,

$$P(A_{n,j} \cap A_{n,k}) = \frac{(n-2)^r}{n^r} = (1 - \frac{2}{n})^r \to e^{-2c}.$$

and Var(N_n)

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

Take $b_n = n$. Then Var(N_n)$/n^2 \to 0$ and

$$\frac{N_n}{n} \to^{L^2} e^{-c}$$

Theorem 5.10 (Weak law for triangular arrays)**.** *Assume that each row in the triangular array $\{X_{n,k}; 1 \leq k \leq k_n\}$ is a finite sequence of independent random variables. Choose an increasing unbounded sequence of positive numbers b_n. Suppose*

1. $\lim_{n\to\infty} \sum_{k=1}^{k_n} P\{|X_{n,k}| > b_n\} = 0$, *and*
2. $\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{k=1}^{k_n} E[X_{n,k}^2 : \{|X_{n,k}| \leq b_n\}] = 0$.

Let $S_n = X_{n,1} + \cdots + X_{n,k_n}$ be the row sum and set $a_n = \sum_{k=1}^{k_n} E[X_{n,k} : \{|X_{n,k}| \leq b_n\}]$. Then

$$\frac{S_n - a_n}{b_n} \to^P 0.$$

Proof. Truncate $X_{n,k}$ at b_n by defining

$$Y_{n,k} = X_{n,k} I_{\{|X_{n,k}| \leq b_n\}}.$$

Let T_n be the row sum of the $Y_{n,k}$ and note that $a_n = ET_n$. Consequently,

$$P\{|\frac{S_n - a_n}{b_n}| > \epsilon\} \leq P\{S_n \neq T_n\} + P\{|\frac{T_n - a_n}{b_n}| > \epsilon\}.$$

To estimate the first term,

$$P\{S_n \neq T_n\} \leq P(\bigcup_{k=1}^{k_n} \{Y_{n,k} \neq X_{n,k}\}) \leq \sum_{k=1}^{k_n} P\{|X_{n,k}| > b_n\}$$

and use hypothesis 1. For the second term, we have by Chebyshev's inequality that

$$\begin{aligned} P\{|\frac{T_n - a_n}{b_n}| > \epsilon\} &\leq \frac{1}{\epsilon^2} E\left(\frac{T_n - a_n}{b_n}\right)^2 = \frac{1}{\epsilon^2 b_n^2} \mathrm{Var}(T_n) \\ &= \frac{1}{\epsilon^2 b_n^2} \sum_{k=1}^{k_n} \mathrm{Var}(Y_{n,k}) \leq \frac{1}{\epsilon^2 b_n^2} \sum_{k=1}^{k_n} EY_{n,k}^2 \end{aligned}$$

and use hypothesis 2. □

The next theorem requires the following exercise.

Exercise 5.11. *If a measurable function $h : [0, \infty) \to \mathbb{R}$ satisfies*

$$\lim_{t\to\infty} h(t) = L, \text{ then } \lim_{T\to\infty} \frac{1}{T}\int_0^T h(t)\,dt = L.$$

Theorem 5.12 (Weak law of large numbers)**.** *Let $X_1, X_2, \dots$ be a sequence of independent random variable having a common distribution. Assume that*

$$\lim_{x\to\infty} xP\{|X_1| > x\} = 0. \tag{5.3}$$

Let $S_n = X_1 + \cdots + X_n$, $\mu_n = E[X_1; \{|X_1| \le n\}]$. Then

$$\frac{S_n}{n} - \mu_n \to^P 0.$$

Proof. We shall use the previous theorem with $X_{n,k} = X_k$, $k_n = n$, $b_n = n$ and $a_n = n\mu_n$. To see that 1 holds, note that

$$\sum_{k=1}^{n} P\{|X_{k,n}| > n\} = nP\{|X_1| > n\}.$$

To check 2, write $Y_{n,k} = X_k I_{\{|X_k|\le n\}}$. Then,

$$EY_{n,1}^2 = \int_0^\infty 2yP\{|Y_{n,1}| > y\}\,dy = \int_0^n 2yP\{|Y_{n,1}| > y\}\,dy \le \int_0^n 2yP\{|X_1| > y\}\,dy.$$

By the hypothesis of the theorem and the exercise with $L = 0$,

$$\lim_{n\to\infty} \frac{1}{n} EY_{n,1}^2 = 0.$$

Therefore,

$$\lim_{n\to\infty} \frac{1}{n^2}\sum_{k=1}^{n} E[X_{n,k}^2 : \{|X_{n,k}| \le n\}] = \lim_{n\to\infty} \frac{n}{n^2} EY_{n,1}^2 = 0.$$

□

Corollary 5.13. *Let $X_1, X_2, \dots$ be a sequence of independent random variable having a common distribution with finite mean μ. Then*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \to^P \mu.$$

Proof.

$$xP\{|X_1| > x\} \le E[|X_1|; \{|X_1| > x\}]. \tag{5.4}$$

Now use the integrability of X_1 to see that the limit is 0 as $x \to \infty$.

By the dominated convergence theorem

$$\lim_{n\to\infty} \mu_n = \lim_{n\to\infty} E[X_1; \{|X_1| \le n\}] = EX_1 = \mu.$$

□

Remark 5.14. *Any random variable X satisfying (5.3) is said to belong to* weak L^1. *The inequality in (5.4) constitutes a proof that weak L^1 contains L^1.*

Example 5.15 (Cauchy distribution)**.** *Let X be $Cau(0,1)$. Then*

$$xP\{|X| > x\} = x\frac{2}{\pi}\int_x^\infty \frac{1}{1+t^2}\,dt = x(1 - \frac{2}{\pi}\tan^{-1} x).$$

which has limit 1 as $x \to \infty$ and the conditions for the weak law fail to hold. We shall see that the average of $Cau(0,1)$ is $Cau(0,1)$.

Example 5.16 (The St. Petersburg paradox)**.** *Let $X_1, X_2, \ldots$ be independent payouts from the game "receive 2^j if the first head is on the j-th toss."*

$$P\{X_1 = 2^j\} = 2^{-j},\ j \geq 1.$$

Check that $EX_1 = \infty$ and that

$$P\{X_1 \geq 2^m\} = 2^{-m+1} = 22^{-m}.$$

If we set $k_n = n, X_{k,n} = X_k$ and write $b_n = 2^{m(n)}$, then, because the payouts have the same distribution, the two criteria in the weak law become

1. $$\lim_{n\to\infty} nP\{X_1 \geq 2^{m(n)}\} = \lim_{n\to\infty} 2n2^{-m(n)}.$$

2. $$\begin{aligned}\lim_{n\to\infty} \frac{n}{2^{2m(n)}} E[X_1^2; \{|X_1| \leq 2^{m(n)}\}] &= \lim_{n\to\infty} \frac{n}{2^{2m(n)}} \sum_{j=1}^{m(n)} 2^{2j} P\{X_1 = 2^j\} \\ &= \lim_{n\to\infty} \frac{n}{2^{2m(n)}}(2^{m(n)+1} - 2) \leq \lim_{n\to\infty} 2n2^{-m(n)}.\end{aligned}$$

   Thus, if the limit in 1 is zero, then so is the limit in 2 and the sequence $m(n)$ must be $o(\log_2 n)$. Next, we compute

$$a_n = nE[X_1; \{|X_1| \leq 2^{m(n)}\}] = n\sum_{j=1}^{m(n)} 2^{2j} P\{X_1 = 2^j\} = nm(n).$$

If $m(n) \to \infty$ as $n \to \infty$, the weak law gives us that

$$\frac{S_n - nm(n)}{2^{m(n)}} \to^P 0.$$

The best result occurs by taking $m(n) \to \infty$ as slowly as possible so that 1 and 2 continue to hold. For example, if we take $m(n)$ to be the nearest integer to $\log_2 n + \log_2 \log_2 n$

$$\frac{S_n - n(\log_2 n + \log_2 \log_2 n)}{n \log_2 n} \to^P 0 \quad \textit{or} \quad \frac{S_n}{n \log_2 n} \to^P 1.$$

Thus, to be fair, the charge for playing n times is approximately $\log_2 n$ per play.

5.4 Strong Law of Large Numbers

Theorem 5.17 (Second Borel-Cantelli lemma)**.** *Assume that events $\{A_n; n \geq 1\}$ are independent and satisfy $\sum_{n=1}^{\infty} P(A_n) = \infty$, then*

$$P\{A_n \text{ i.o.}\} = 1.$$

Proof. Recall that for any $x \in \mathbb{R}$, $1 - x \leq e^{-x}$. For any integers $0 < M < N$,

$$P(\bigcap_{n=M}^{N} A_n^c) = \prod_{n=M}^{N} (1 - P(A_n)) \leq \prod_{n=M}^{N} \exp(-P(A_n)) = \exp\left(-\sum_{n=M}^{N} P(A_n)\right).$$

This has limit 0 as $N \to \infty$. Thus, for all M,

$$P(\bigcup_{n=M}^{\infty} A_n) = 1.$$

Now use the definition of infinitely often and the continuity from of above of a probability to obtain the theorem. □

Taken together, the two Borel-Cantelli lemmas give us our first example of a *zero-one law*. For independent events $\{A_n; n \geq 1\}$,

$$P\{A_n \text{ i.o.}\} = \begin{cases} 0 & \text{if} \quad \sum_{n=1}^{\infty} P(A_n) < \infty, \\ 1 & \text{if} \quad \sum_{n=1}^{\infty} P(A_n) = \infty. \end{cases}$$

Exercise 5.18. *1. Let $\{X_n; n \geq 1\}$ be the outcome of independent coin tosses with probability of heads p. Let $\{\epsilon_1, \ldots, \epsilon_k\}$ be any sequence of heads and tails, and set*

$$A_n = \{X_n = \epsilon_1, \ldots, X_{n+k-1} = \epsilon_k\}.$$

Then, $P\{A_n \text{ i.o.}\} = 1$.

2. Let $\{X_n; n \geq 1\}$ be the outcome of independent coin tosses with probability of heads p_n. Then

(a) $X_n \to^P 0$ if and only if $p_n \to 0$, and

(b) $X_n \to^{a.s.} 0$ if and only if $\sum_{n=1}^{\infty} p_n < \infty$.

3. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables. Then, they have common finite mean if and only if $P\{X_n > n \text{ i.o.}\} = 0$.

4. Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables. Find necessary and sufficient conditions so that

(a) $X_n/n \to^{a.s.} 0$,

(b) $(\max_{m \leq n} X_m)/n \to^{a.s.} 0$,

(c) $(\max_{m \leq n} X_m)/n \to^{P} 0$,

(d) $X_n/n \to^{P} 0$,

5. For the St. Petersburg's paradox, show that

$$\limsup_{n\to\infty} \frac{X_n}{n\log_2 n} = \infty$$

almost surely and hence

$$\limsup_{n\to\infty} \frac{S_n}{n\log_2 n} = \infty$$

.

Theorem 5.19 (Strong Law of Lange Numbers)**.** *Let $X_1, X_2, \ldots$ be independent identically distributed random variables and set $S_n = X_1 + \cdots + X_n$, then*

$$\lim_{n\to\infty} \frac{1}{n} S_n$$

exists almost surely if and only if $E|X_1| < \infty$. In this case the limit is $EX_1 = \mu$ with probability 1.

The following proof, due to Etemadi in 1981, will be accomplished in stages.

Lemma 5.20. *Let $Y_k = X_k I_{\{|X_k|\le k\}}$ and $T_n = Y_1 + \cdots + Y_n$, then it is sufficient to prove that*

$$\lim_{n\to\infty} \frac{1}{n} T_n = \mu$$

Proof.

$$\sum_{k=1}^{\infty} P\{X_k \neq Y_k\} = \sum_{k=1}^{\infty} P\{|X_k| \ge k\} = \sum_{k=1}^{\infty} \int_{k-1}^{k} P\{|X_k| \ge k\}\, dx \le \int_0^{\infty} P\{|X_1| > x\}\, dx = E|X_1| < \infty.$$

Thus, by the first Borel-Cantelli, $P\{X_k \neq Y_k \text{ i.o.}\} = 0$. Fix $\omega \notin \{X_k \neq Y_k \text{ i.o.}\}$ and choose $N(\omega)$ so that $X_k(\omega) = Y_k(\omega)$ for all $k \ge N(\omega)$. Then

$$\lim_{n\to\infty} \frac{1}{n}(S_n(\omega) - T_n(\omega)) = \lim_{n\to\infty} \frac{1}{n}(S_{N(\omega)}(\omega) - T_{N(\omega)}(\omega)) = 0.$$

□

Lemma 5.21.

$$\sum_{k=1}^{\infty} \frac{1}{k^2} Var(Y_k) < \infty.$$

Proof. Set $A_{j,k} = \{j-1 \le X_k < j\}$ and note that $P(A_{j,k}) = P(A_{j,1})$. Then, noting that reversing the order of summation holds if the summands are non-negative, we have that

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{1}{k^2} \mathrm{Var}(Y_k) &\le \sum_{k=1}^{\infty} \frac{1}{k^2} E[Y_k^2] = \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{j=1}^{k} E[Y_k^2; A_{j,k}] \\
&\le \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{j=1}^{k} j^2 P(A_{j,k}) = \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} \frac{j^2}{k^2} P(A_{j,1})
\end{aligned}$$

Note that for

$$j > 1, \quad \sum_{k=j}^{\infty} \frac{1}{k^2} \leq \int_{j-1}^{\infty} \frac{1}{x^2}\, dx = \frac{1}{j-1} \leq \frac{2}{j}$$

$$j = 1, \quad \sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \sum_{k=2}^{\infty} \frac{1}{k^2} \leq 2 = \frac{2}{j}.$$

Consequently,

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \mathrm{Var}(Y_k) \leq \sum_{j=1}^{\infty} j^2 \frac{2}{j} P(A_{j,1}) = 2 \sum_{j=1}^{\infty} j P(A_{j,1}) = 2EZ$$

where $Z = \sum_{j=1}^{\infty} j I_{A_{j,k}}$.

Because $Z \leq X_1 + 1$,

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \mathrm{Var}(Y_k) \leq 2E[X_1 + 1] < \infty.$$

□

Theorem 5.22. *The strong law holds for non-negative random variables.*

Proof. Choose $\alpha > 1$ and set $\beta_k = [\alpha^k]$. Then

$$\beta_k \geq \frac{\alpha^k}{2}, \quad \frac{1}{\beta_k^2} \leq \frac{4}{\alpha^{2k}}.$$

Thus, for all $m \geq 1$,

$$\sum_{k=m}^{\infty} \frac{1}{\beta_k^2} \leq 4 \sum_{k=m}^{\infty} \alpha^{-2k} = \frac{4\alpha^{-2m}}{1-\alpha^{-2}} = \frac{4}{1-\alpha^{-2}} \frac{1}{\alpha^{2m}} \leq A \frac{1}{\beta_m^2}.$$

As shown above, we may prove the stong law for T_n. Let $\epsilon > 0$, then by Chebyshev's inequality,

$$\sum_{n=1}^{\infty} P\{\frac{1}{\beta_n} |T_{\beta_n} - ET_{\beta_n}| > \epsilon\} \leq \sum_{n=1}^{\infty} \frac{1}{(\epsilon\beta_n)^2} \mathrm{Var}(T_n) \leq \frac{1}{\epsilon^2} \sum_{n=1}^{\infty} \frac{1}{\beta_n^2} \sum_{k=1}^{\beta_n} \mathrm{Var}(Y_k)$$

by the independence of the X_n.

To interchange the order of summation, let

$$\gamma_k = j \text{ if and only if } \beta_j = k.$$

Then the double sum above is

$$\frac{1}{\epsilon^2} \sum_{k \in im(\beta)} \sum_{n=\gamma_k}^{\infty} \frac{1}{\beta_n^2} \mathrm{Var}(Y_k) \leq \frac{A}{\epsilon} \sum_{k=1}^{\infty} \frac{1}{\beta_{\gamma_k}^2} \mathrm{Var}(Y_k) = \frac{A}{\epsilon} \sum_{k=1}^{\infty} \frac{1}{k^2} \mathrm{Var}(Y_k) < \infty.$$

By the first Borel-Cantelli lemma,

$$P\{\frac{1}{\beta_n} |T_{\beta_n} - ET_{\beta_n}| > \epsilon \text{ i.o.}\} = 0.$$

Consequently,

$$\lim_{n\to\infty} \frac{1}{\beta_n}(T_{\beta_n} - ET_{\beta_n}) = 0 \text{ almost surely.}$$

Now we have the convergence along any geometric subsequence because

$$EY_k = E[X_k; \{X_k < k\}] = E[X_1; \{X_1 < k\}] \to EX_1 = \mu$$

by the monotone convergence theorem. Thus,

$$\frac{1}{\beta_n} ET_n = \frac{1}{\beta_n} \sum_{k=1}^{\beta_n} EY_k \to \mu. \tag{5.5}$$

We need to fill the gaps between β_n and β_{n+1}, Use the fact that $Y_k \geq 0$ to conclude that T_n is monotone increasing. So, for $\beta_n \leq m \leq \beta_{n+1}$,

$$\frac{1}{\beta_{n+1}} T_{\beta_n} \leq \frac{1}{m} T_m \leq \frac{1}{\beta_n} T_{\beta_{n+1}},$$

$$\frac{\beta_n}{\beta_{n+1}} \frac{1}{\beta_n} T_{\beta_n} \leq \frac{1}{m} T_m \leq \frac{\beta_{n+1}}{\beta_n} \frac{1}{\beta_n} T_{\beta_{n+1}},$$

and

$$\liminf_{n\to\infty} \frac{\beta_n}{\beta_{n+1}} \frac{1}{\beta_n} T_{\beta_n} \leq \liminf_{m\to\infty} \frac{1}{m} T_m \leq \limsup_{m\to\infty} \frac{1}{m} T_m \leq \limsup_{n\to\infty} \frac{\beta_{n+1}}{\beta_n} \frac{1}{\beta_n} T_{\beta_{n+1}}.$$

Thus, on the set in which (5.5) holds, we have, for each $\alpha > 1$, that

$$\frac{\mu}{\alpha} \leq \liminf_{m\to\infty} \frac{1}{m} T_m \leq \limsup_{m\to\infty} \frac{1}{m} T_m \leq \alpha\mu.$$

Now consider a decreasing sequence $\alpha_k \to 1$, then

$$\{\lim_{m\to\infty} \frac{1}{m} T_m = \mu\} = \bigcap_{k=1}^{\infty} \{\frac{\mu}{\alpha_k} \leq \liminf_{m\to\infty} \frac{1}{m} T_m \leq \limsup_{m\to\infty} \frac{1}{m} T_m \leq \alpha_k \mu\}.$$

Because this is a countable intersection of probability one events, it also has probability one. □

Proof. (Strong Law of Large Numbers) For general random variables with finite absolute mean, write

$$X_n = X_n^+ - X_n^-.$$

We have shown that each of the events

$$\{\lim_{n\to\infty} \frac{1}{n} \sum_{n=1}^{\infty} X_k^+ = EX_1^+\},\ \{\lim_{n\to\infty} \frac{1}{n} \sum_{n=1}^{\infty} X_k^- = EX_1^-\}$$

has probability 1. Hence, so does their intersection which includes

$$\{\lim_{n\to\infty} \frac{1}{n} \sum_{n=1}^{\infty} X_k = EX_1\}.$$

For the converse, if $\lim_{n\to\infty} \frac{1}{n} S_n$ exists almost surely, then

$$\frac{1}{n} X_n \to^{a.s.} 0 \text{ as } n \to \infty.$$

Therefore $P\{|X_n| > n \text{ i.o.}\} = 0$. Because these events are independent, we can use the second Borel-Cantelli lemma in contraposition to conclude that

$$\infty > \sum_{n=1}^{\infty} P\{|X_n| > n\} = \sum_{n=1}^{\infty} P\{|X_1| > n\} \geq E|X_1| - 1.$$

Thus, $E|X_1| < \infty$. □

Remark 5.23. *Independent and identically distributed integrable random variables are easily seen to be uniformly integrable. Thus, S_n/n is uniformly integrable. Because the limit exists almost surely, and because S_n/n is uniformly integrable, the convergence must also be in L^1.*

5.5 Applications

Example 5.24 (Monte Carlo integration)**.** *Let $X_1, X_2, \ldots$ be independent random variables uniformally distributed on the interval $[0, 1]$. Then*

$$\overline{g(X)}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \to \int_0^1 g(x)\ dx = I(g)$$

with probability 1 as $n \to \infty$. The error in the estimate of the integral is supplied by the variance

$$Var(\overline{g(X)}_n) = \frac{1}{n} \int_0^1 (g(x) - I(g))^2\ dx = \frac{\sigma^2}{n}.$$

Example 5.25 (importance sampling)**.** *Importance sampling methods begin with the observation that we could perform the Monte Carlo integration above beginning with $Y_1, Y_2, \ldots$ independent random variables with common density f_Y with respect to Lebesgue measure on $[0, 1]$. Define the* importance sampling weights

$$w(y) = \frac{g(y)}{f_Y(y)}.$$

Then

$$\overline{w(Y)}_n = \frac{1}{n} \sum_{i=1}^{n} w(Y_i) \to \int_0^1 w(y) f_Y(y)\ dy = \int_0^1 \frac{g(y)}{f_Y(y)} f_Y(y)\ dy = I(g).$$

This is an improvement if the variance in the estimator decreases, i.e.,

$$\int_0^1 (w(x) - I(g))^2 f_Y(y)\ dx = \sigma_f^2 << \sigma^2.$$

The density f_Y is called the importance sampling function *or the* proposal density*. For the case in which g is a non-negative function, the optimal proposal distribution is a constant times g. Knowing this constant is equivalent to solving the original numerical integration problem.*

Example 5.26 (Weierstrass approximation theorem)**.** *Let $\{X_n; n \geq 1\}$ be independent $Ber(p)$ random variables and left $f : [0,1] \to \mathbb{R}$ be continuous. The sum $S_n = X_1 + \cdots + X_n$ is a $Bin(n,p)$ random variable and consequently*

$$Ef\left(\frac{1}{n}S_n\right) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) P\{S_n = k\} = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k}.$$

This is known as the Bernstein polynomial *of degree n.*

By the strong law of large numbers $S_n/n \to p$ almost surely. Thus, by the bounded convergence theorem

$$f(p) = \lim_{n\to\infty} \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} p^k (1-p)^{n-k}.$$

To check that the convergence is uniform, let $\epsilon > 0$. Because f is uniformly continuous, there exists $\delta > 0$ so that $|p - \tilde{p}| < \delta$ implies $|f(p) - f(\tilde{p})| < \epsilon/2$. Therefore,

$$\begin{aligned}
\left|Ef\left(\frac{1}{n}S_n\right) - f(p)\right| &\leq E\left[\left|f\left(\frac{1}{n}S_n\right) - f(p)\right|; \left\{\left|\frac{1}{n}S_n - p\right| < \delta\right\}\right] \\
&\quad + E\left[\left|f\left(\frac{1}{n}S_n\right) - f(p)\right|; \left\{\left|\frac{1}{n}S_n - p\right| \geq \delta\right\}\right] \\
&\leq \frac{\epsilon}{2} + ||f||_\infty P\left\{\left|\frac{1}{n}S_n - p\right| \geq \delta\right\}.
\end{aligned}$$

By Chebyshev's inequality, the second term in the previous line is bounded above by

$$\frac{||f||_\infty}{\delta^2 n} Var(X_1) = \frac{||f||_\infty}{\delta^2 n} p(1-p) \leq \frac{||f||_\infty}{4\delta^2 n} < \frac{\epsilon}{2}$$

whenever $n > ||f||_\infty/(2\delta^2\epsilon)$.

Exercise 5.27. *Generalize and prove the Weierstrass approximation theorem for continuous $f : [0,1]^d \to \mathbb{R}$ for $d > 1$.*

Example 5.28 (Shannon's theorem)**.** *Let $X_1, X_2, \ldots$ be independent random vaariables taking values in a finite* alphabet *S. Define $p(x) = P\{X_1 = x\}$. For the observation $X_1(\omega), X_2(\omega), \ldots$, the random variable*

$$\pi_n(\omega) = p(X_1(\omega)) \cdots p(X_n(\omega))$$

give the probability of that observation. Then

$$\log \pi_n = \log p(X_1) + \cdots + \log p(X_n).$$

By the strong law of large numbers

$$\lim_{n\to\infty} -\frac{1}{n}\pi_n = -\sum_{x\in S} p(x) \log p(x) \quad \textit{almost surely}$$

This sum, often denote H, is called the (Shannon) entropy *of the source and*

$$\pi_n \approx \exp(-nH)$$

The strong law of large numbers stated in this context is called the asymptotic equipartition property.

Exercise 5.29. *Show that the Shannon entropy takes values between 0 and* $\log n$*. Describe the cases that gives these extreme values.*

Definition 5.30. *Let* $X_1, X_2, \ldots$ *be independent with common distribution* F*, then call*

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} I_{(-\infty,x]}(X_k)$$

the empirical distribution function*. This is the fraction of the first* n *observations that fall below* x*.*

Theorem 5.31 (Glivenko-Cantelli)**.** *The empirical distribution functions* F_n *for* $X_1, X_2, \ldots$ *be independent and identically distributed random variables converge uniformly almost surely to the distribution function as* $n \to \infty$*.*

Proof. Let the X_n have common distribution function F. We must show that

$$P\{\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0\} = 1.$$

Call $D_n = \sup_x |F_n(x) - F(x)|$. By the right continuity of F_n and F, this supremum is achieved by restricting the supremum to rational numbers. Thus, in particular, D_n is a random variable.

For fixed x, the strong law of large numbers states that

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} I_{(-\infty,x]}(X_k) \to E[I_{(-\infty,x]}(X_1)] = F(x)$$

on a set R_x having probability 1. Similarly,

$$F_n(x-) = \frac{1}{n}\sum_{k=1}^{n} I_{(-\infty,x)}(X_k) \to F(x-)$$

on a set L_x having probability 1. Define

$$H(t) = \inf\{x; t \le F(x)\}$$

Check that

$$F(H(t)-) \le t \le F(H(t)).$$

Now, define the doubly indexed sequence $x_{m,k} = H(k/m)$. Hence,

$$F(x_{m,k}-) - F(x_{m,k-1}) \le \frac{1}{m}, \quad 1 - F(x_{m,m}) \le \frac{1}{m}.$$

Set

$$D_{m,n} = \max\{|F_n(x_{m,k}) - F(x_{m,k})|, |F_n(x_{m,k}-) - F(x_{m,k}-)|; k = 1, \ldots, m\}.$$

For $x \in [x_{m,k-1}, x_{m,k})$,

$$F_n(x) \le F_n(x_{m,k}-) \le F(x_{m,k}-) + D_{m,n} \le F(x) + \frac{1}{m} + D_{m,n}$$

and

$$F_n(x) \geq F_n(x_{m,k-1}) \geq F(x_{m,k-1}) - D_{m,n} \geq F(x) - \frac{1}{m} - D_{m,n}$$

Use a similar argument for $x < x_{m,1}$ and $x > x_{m,m}$ to see that

$$D_n \leq D_{m,n} + \frac{1}{m}.$$

Define

$$\Omega_0 = \bigcap_{m,k \geq 1} (L_{k/m} \cap R_{k/m}).$$

Then, $P(\Omega_0) = 1$ and on this set

$$\lim_{n \to \infty} D_{m,n} = 0 \text{ for all } m.$$

Consequently,

$$\lim_{n \to \infty} D_n = 0.$$

with probability 1. □

5.6 Large Deviations

We have seen that the statistical average of independent and identically distributed random variables converges almost surely to their common expected value. We now examine how unlikely this average is to be away from the mean.

To motivate the theory of large deviations, let $\{X_k; k \geq 1\}$ be independent and identically distributed random variables with moment generating function m. Choose $x > \mu$. Then, by Chebyshev's inequality, we have for any $\theta > 0$,

$$P\{\frac{1}{n}\sum_{k=1}^{n} X_k > x\} = P\{\exp\theta(\frac{1}{n}\sum_{k=1}^{n} X_k) > e^{\theta x}\} \leq \frac{E[\exp\theta(\frac{1}{n}\sum_{k=1}^{n} X_k)]}{e^{\theta x}}.$$

In addition,

$$E[\exp\theta(\frac{1}{n}\sum_{k=1}^{n} X_k)] = \prod_{k=1}^{n} E[\exp(\frac{\theta}{n}X_k)] = m(\frac{\theta}{n})^n.$$

Thus,

$$\frac{1}{n}\log P\{\frac{1}{n}\sum_{k=1}^{n} X_k > x\} \leq -\frac{\theta}{n}x + \lambda(\frac{\theta}{n})$$

where λ is the logarithm of the moment generating function. Taking infimum over all choices of $\theta > 0$ we have

$$\frac{1}{n}\log P\{\frac{1}{n}\sum_{k=1}^{n} X_k > x\} \leq -\lambda^*(x).$$

with

$$\lambda^*(x) = \sup_{\theta > 0}\{\theta x - \lambda(\theta)\}.$$

If $\lambda^*(x) > 0$, then

$$P\{\frac{1}{n}\sum_{k=1}^{n} X_k > x\} \leq \exp(-n\lambda^*(x)),$$

a geometric sequence tending to 0.

Definition 5.32. *For an $\mathbb{R}$-valued random variable X, define the* logarithmic moment generating function

$$\lambda(\theta) = \log E[\exp(\theta X)],\ \theta \in \mathbb{R}.$$

The Legendre-Fenchel transform *of a function λ is*

$$\lambda^*(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \lambda(\theta)\}.$$

When λ is the log moment generating function, λ^* is called the rate function.

Exercise 5.33. *Find the Legendre-Fenchel transform of $\lambda(\theta) = \theta^p/p$, $p > 1$.*

Call the domains $\mathcal{D}_\lambda = \{\theta : \lambda(\theta) < \infty\}$ and $\mathcal{D}_{\lambda^*} = \{\theta : \lambda^*(\theta) < \infty\}$.

Let's now explore some properties of λ and λ^*.

1. *λ and λ^* are convex.*

   The convexity of λ follows from Hólder's inequality. For $\alpha \in (0,1)$,

   $$\lambda(\alpha\theta_1+(1-\alpha)\theta_2) = \log E[(e^{\theta_1 X})^\alpha (e^{\theta_2 X})^{(1-\alpha)}] \leq \log\left(E[e^{\theta_1 X}]^\alpha E[e^{\theta_2 X}]^{(1-\alpha)}]\right) = \alpha\lambda(\theta_1)+(1-\alpha)\lambda(\theta_2).$$

   The convexity of λ^* follows from the definition. Again, for $\alpha \in (0,1)$,

   $$\begin{aligned}\alpha\lambda^*(x_1) + (1-\alpha)\lambda^*(x_2) &= \sup_{\theta\in\mathbb{R}}\{\alpha\theta x_1 - \alpha\lambda(\theta)\} + \sup_{\theta\in\mathbb{R}}\{(1-\alpha)\theta x_2 - (1-\alpha)\lambda(\theta)\}\\ &\geq \sup_{\theta\in\mathbb{R}}\{\theta(\alpha x_1 + (1-\alpha)x_2) - \lambda(\theta)\} = \lambda^*(\alpha x_1 + (1-\alpha)x_2)\end{aligned}$$

2. If $\mu \in \mathbb{R}$, then $\lambda^*(x)$ take on the minimum value zero at $x = \mu$.

   $\lambda(0) = \log E[e^{0X}] = 0$. Thus,

   $$\lambda^*(x) \geq 0x - \lambda(0x) = 0.$$

   By Jensen's inequality,

   $$\lambda(\theta) = \log E[e^{\theta X}] \geq E[\log e^{\theta X}] = \theta\mu$$

   and thus

   $$\theta\mu - \lambda(\theta) \leq 0$$

   for all θ. Consequently, $\lambda(\mu) = 0$.

3. *If* $\mathcal{D}_\lambda = \{0\}$, *then* λ^* *is identically* 0.

$$\lambda^*(\theta) = \lambda(0) = 0.$$

4. λ^* *is lower semicontinuous.*

Fix a sequence $x_n \to x$, then

$$\liminf_{n\to\infty} \lambda^*(x_n) \geq \liminf_{n\to\infty} (\theta x_n - \lambda(\theta)) = \theta x - \lambda(x).$$

Thus,

$$\liminf_{n\to\infty} \lambda^*(x_n) \geq \sup_{\theta\in\mathbb{R}} \{\theta x - \lambda(\theta)\} = \lambda^*(x).$$

5. *If* $\lambda(\theta) < \infty$ *for some* $\theta > 0$, *then* $\mu \in [-\infty, \infty)$ *and for all* $x \geq \mu$,

$$\lambda^*(x) = \sup_{\theta\geq 0} \{\theta x - \lambda(\theta)\}$$

is a non-decreasing function on (μ, ∞).

For the positive value of θ guaranteed above,

$$\theta E X^+ = E[\theta X; \{X \geq 0\}] \leq E[e^{\theta X}; \{X \geq 0\}] \leq m(\theta) = \exp \lambda(\theta) < \infty.$$

and $\mu \neq \infty$.
So, if $\mu = -\infty$, then $\lambda(\theta) = \infty$ for $\theta < 0$ thus we can reduce the infimum to the set $\{\theta \geq 0\}$. If $\mu \in \mathbb{R}$, then for any $\theta < 0$,

$$\theta x - \lambda(\theta) \leq \theta\mu - \lambda(\theta) \leq \lambda^*(\mu) = 0$$

and the supremum takes place on the set $\theta \geq 0$. The monotonicity of λ^* on (μ, ∞) follows from the fact that $\theta x - \lambda(\theta)$ is non-decreasing as a function of x provided $\theta \geq 0$.

The corresponding statement holds if $\lambda(\theta) < \infty$ for some $\theta < 0$.

6. In all cases, $\inf_{x\in\mathbb{R}} \lambda^*(x) = 0$.

This property has been established if μ is finite or if $\mathcal{D}_\lambda = \{0\}$. Now consider the case $\mu = -\infty$, $\mathcal{D}_\lambda \neq \{0\}$, noting that the case $\mu = \infty$ can be handled similarly. Choose $\theta > 0$ so that $\lambda(\theta) < \infty$. Then, by Chebyshev's inequality,

$$\log P\{X > x\} \leq \inf_{\theta\geq 0} \log E[e^{\theta(X-x)}] = -\sup_{\theta\geq 0} \{\theta x - \lambda(\theta)\} = -\lambda^*(x).$$

Consequently,

$$\lim_{x\to-\infty} \lambda^*(x) \leq \lim_{x\to-\infty} -\log P\{X > x\} = 0.$$

7. **Exercise.** *λ is differentiable on the interior of $\mathcal{D}_\lambda$ with*

$$\lambda'(\theta) = \frac{1}{m(\theta)} E[Xe^{\theta X}].$$

In addition,

$$\lambda'(\theta) = \tilde{x} \text{ implies } \lambda^*(\tilde{x}) = \theta\tilde{x} - \lambda(\theta).$$

Exercise 5.34. *Show that*

1. *If X is $Pois(\mu)$, $\lambda^*(x) = \mu - x + x\log(x/\mu)$, $x > 0$ and infinite if $x \leq 0$.*
2. *If X is $Ber(p)$, $\lambda^*(x) = x\log(x/p) + (1-x)\log((1-x)/(1-p))$ for $x \in [0,1]$ and infinite otherwise.*
3. *If X is $Exp(\beta)$, $\lambda^*(x) = \beta x - 1 - \log(\beta x)$ $x > 0$ and infinite if $x \leq 0$.*
4. *If X is $N(0,\sigma^2)$, $\lambda^*(x) = x^2/2\sigma^2$*

Theorem 5.35 (Cramér)**.** *Let $\{X_k; k \geq 1\}$ be independent and identically distributed random variables with log moment generating function λ. Let λ^* be the Legendre-Fenchel transform of λ and write $I(A) = \inf_{x\in A} \lambda^*(x)$ and ν_n for the distribution of S_n/n, $S_n = X_1 + \cdots + X_n$, then*

1. *(upper bound) For any closed set $F \subset \mathbb{R}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \nu_n(F) \leq -I(F).$$

2. *(lower bound) or any open set $G \subset \mathbb{R}$,*

$$\liminf_{n\to\infty} \frac{1}{n} \log \nu_n(G) \geq -I(G).$$

Proof. (upper bound) Let F be a non-empty closed set. The theorem holds trivially if $I(F) = 0$, so assume that $I(F) > 0$. Consequently, μ exists (possibly as an extended real number number). By Chebyshev's inequality, we have for every x and every $\theta > 0$,

$$\nu_n[x,\infty) = P\{\frac{1}{n}S_n - x \geq 0\} \leq E[\exp(n\theta(S_n/n - x))] = e^{-n\theta x} \prod_{k=1}^{n} E[e^{\theta X_k}] = \exp(-n(\theta x - \lambda(\theta))).$$

Therefore, if $\mu < \infty$,

$$\nu_n[x,\infty) \leq \exp -n\lambda^*(x) \text{ for all } x > \mu.$$

Similarly, if $\mu > -\infty$,

$$\nu_n(-\infty, x] \leq \exp -n\lambda^*(x) \text{ for all } x < \mu.$$

Case I. μ finite.

$\lambda^*(\mu) = 0$ and because $I(F) > 0$, $\mu \in F^c$. Let (x_-, x_+) be the largest open interval in F^c that contains x. Because $F \neq \emptyset$, at least one of the endpoints is finite.

$$x_- \text{ finite implies } x_- \in F \text{ and consequently } \lambda^*(x_-) \geq I(F).$$

$$x_+ \text{ finite implies } x_+ \in F \text{ and consequently } \lambda^*(x_+) \geq I(F).$$

Note that $F \subset (\infty, x_-] \cap [x_+, \infty)$ we have by the inequality above that

$$\nu_n(F) \leq \nu_n(\infty, x_-] + \nu_n[x_+, \infty) \leq \exp -n\lambda^*(x_-) + \exp -n\lambda^*(x_+) \leq 2\exp -nI(F).$$

Case II. μ is infinite.

We consider the case $\mu = -\infty$. The case $\mu = \infty$ is handled analogously. We have previously shown that $\lim_{x\to-\infty} \lambda^*(x) = 0$. Thus, $I(F) > 0$ implies that x_+, the infimum of the set F is finite. F is closed, so $x_+ \in F$ and $\lambda^*(x_+) \geq I(F)$. In addition, $F \subset [x_+, \infty)$ and so

$$\nu_n(F) \leq \nu_n[x_+, \infty) \leq \exp -n\lambda^*(x) \leq \exp -nI(F).$$

(lower bound) *Claim.* For every $\delta > 0$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \nu_n(-\delta, \delta) \geq \inf_{\theta\in\mathbb{R}} \lambda(\theta) = -\lambda^*(0).$$

Case I. The support of X_1 is compact and both $P\{X_1 > 0\} > 0$ and $P\{X_1 < 0\} > 0$.

The first assumption guarantees that $\mathcal{D}_\lambda = \mathbb{R}$. The second assures that $\lambda(\theta) \to \infty$ as $|\theta| \to \infty$. This guarantees a unique finite global minimum

$$\lambda(\eta) = \inf_{\theta\in\mathbb{R}} \lambda(\theta) \text{ and } \lambda'(\eta) = 0.$$

Define a new measure $\tilde{\nu}$ with density

$$\frac{d\tilde{\nu}}{d\nu_1} = \exp(\eta x - \lambda(\eta)).$$

Note that

$$\tilde{\nu}(\mathbb{R}) = \int_{\mathbb{R}} \frac{d\tilde{\nu}}{d\nu_1}\ \nu_1(dx) = \exp(-\lambda(\eta)) \int_{\mathbb{R}} e^{\eta x}\ \nu_1(dx) = 1$$

and $\tilde{\nu}$ is a probability.

Let $\{\tilde{X}_k; k \geq 1\}$ be random variables with distribution $\tilde{\nu}$ and let $\tilde{\nu}_n$ denote the distribution of $(\tilde{X}_1 + \cdots + \tilde{X}_n)/n$. Note that

$$E\tilde{X}_1 = \exp(-\lambda(\eta)) \int_{\mathbb{R}} xe^{\eta x}\ \nu_1(dx) = \lambda'(\eta) = 0.$$

By the law of large numbers, we have, for any $\tilde{\delta} > 0$,

$$\lim_{n\to\infty} \tilde{\nu}_n(-\tilde{\delta}, \tilde{\delta}) = 1.$$

Let's compare this to

$$\begin{aligned}
\nu_n(-\tilde{\delta}, \tilde{\delta}) &= \int I_{\{|\sum_{k=1}^n x_k| < n\tilde{\delta}\}}\ \nu(dx_1)\cdots\nu(dx_n) \\
&\geq \exp(-n\tilde{\delta}|\eta|) \int I_{\{|\sum_{k=1}^n x_k| < n\tilde{\delta}\}} \exp(\eta \sum_{k=1}^n x_k)\ \nu(dx_1)\cdots\nu(dx_n) \\
&= \exp(-n\tilde{\delta}|\eta|) \exp(n\lambda(\eta))\tilde{\nu}_n(-\tilde{\delta}, \tilde{\delta}).
\end{aligned}$$

Therefore, for every $0 < \tilde{\delta} < \delta$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \nu_n(-\delta, \delta) \geq \liminf_{n\to\infty} \frac{1}{n} \log \nu_n(-\tilde{\delta}, \tilde{\delta}) \geq \lambda(\eta) - \tilde{\delta}|\eta|$$

and the claim follows for case I.

Case II. ν does not necessarily have compact support.

Choose M sufficiently large so that both

$$P\{0 < X_1 \leq M\} > 0 \text{ and } P\{0 > X_1 \geq -M\} > 0.$$

Let $\tilde{\nu}^M(A) = P\{X_1 \in A | |X_1| \leq M\}$ and

$$\tilde{\nu}_n^M(A) = P\{(X_1 + \cdots + X_n)/n \in A | |X_k| \leq M; k = 1, \ldots, n\}.$$

Then,

$$\begin{aligned} \nu^n(-\delta, \delta) &= P\{-\delta < (X_1 + \cdots + X_n)/n < \delta | |X_k| \leq M; k = 1, \ldots, n\} P\{|X_k| \leq M; k = 1, \ldots, n\} \\ &= \tilde{\nu}_n^M(-\delta, \delta) \nu[-M, M])^n. \end{aligned}$$

Now apply case I to $\tilde{\nu}^M$. The log moment generating function for $\tilde{\nu}^M$ is

$$\lambda^M(\theta) = \log \int_{-M}^{M} e^{\theta x} \, \nu(dx) - \log \nu[-M, M] = \tilde{\lambda}^M(\theta) - \log \nu[-M, M].$$

Consequently,

$$\liminf_{n\to\infty} \frac{1}{n} \nu_n(-\delta, \delta) \geq \log \nu[-M, M] + \liminf_{n\to\infty} \frac{1}{n} \tilde{\nu}_n^M(-\delta, \delta) \geq \inf_{\theta\in\mathbb{R}} \tilde{\lambda}^M(\theta).$$

Set

$$I_M = -\inf_{\theta\in\mathbb{R}} \tilde{\lambda}^M(\theta).$$

Because $M \mapsto \tilde{\lambda}^M(\theta)$ is nondecreasing, so is $-I_M$ and

$$\tilde{I} = \lim_{M\to\infty} I_M$$

exists and is finite. Moreover,

$$\liminf_{n\to\infty} \frac{1}{n} \nu_n(-\delta, \delta) \geq -\tilde{I}.$$

Because $-I_M \leq \tilde{\lambda}^M(0) \leq \lambda(0) = 0$ for all M, $-\tilde{I} \leq 0$. Therefore, the level sets

$$\tilde{\lambda}_M^{-1}(-\infty, \tilde{I}] = \{\theta; \tilde{\lambda}_M(\theta) \leq \tilde{I}\}$$

are nonempty, closed, and bounded (hence compact) and nested. Thus, by the finite intersection property, their intersection is non-empty. So, choose θ_0 in the intersection. By the monotone convergence theorem,

$$\lambda(\theta_0) = \lim_{M\to\infty} \tilde{\lambda}_M(\theta_0) \leq -\tilde{I}$$

and the claim holds.

Case III. $\nu(-\infty, 0) = 0$ or $\nu(0, \infty) = 0$,

In this situation, λ is monotone and $\inf_{\theta\in\mathbb{R}} \lambda(\theta) = \log \nu\{0\}$. The claim follows from obvserving that

$$\nu_n(-\delta, \delta) \geq \nu_n\{0\} = \nu\{0\}^n.$$

Now consider the transformation $\tilde{X}_k = X_k - x_0$, then its log moment generating function is

$$\tilde{\lambda}(\theta) = \log E[e^{\theta(X_1 - x_0)}] = \lambda(\theta) - \theta x_0.$$

Its Legendre transform

$$\tilde{\lambda}^*(x) = \sup_{\theta\in\mathbb{R}}\{\theta x - \tilde{\lambda}(\theta)\} = \sup_{\theta\in\mathbb{R}}\{\theta(x + x_0) - \lambda(\theta)\} = \lambda^*(x + x_0).$$

Thus, by the claim, we have for every x_0 and every $\delta > 0$

$$\liminf_{n\to\infty} \frac{1}{n} \log \nu_n(x_0 - \delta, x_0 + \delta) \geq -\lambda^*(x_0).$$

Finally, for any open set G and any $x_0 \in G$, we can choose $\delta > 0$ so that $(x_0 - \delta, x_0 + \delta) \subset G$.

$$\liminf_{n\to\infty} \frac{1}{n} \nu_n(G) \geq \liminf_{n\to\infty} \frac{1}{n} \log \nu_n(x_0 - \delta, x_0 + \delta) \geq -\lambda^*(x_0)$$

and the lower bound follows. □

Remark 5.36. *Note that the proof provides that*

$$\mu_n(F) \leq 2 \exp(-nI(F)).$$

Example 5.37. *For $\{X_k; x \geq 1\}$ independent $Exp(\beta)$ random variables, we have for $x > 1/\beta$,*

$$P\{\frac{1}{n}(X_1 + \cdots + X_n) > x\} \leq e^{-(\beta x - 1)} \left(\frac{1}{\beta x}\right)^n,$$

and for $x < 1/\beta$,

$$P\{\frac{1}{n}(X_1 + \cdots + X_n) < x\} \leq e^{-(\beta x - 1)} \left(\frac{1}{\beta x}\right)^n.$$

6 Convergence of Probability Measures

In ths section, (S, d) is a separable metric space, $C_b(S)$ is space of bounded continuous functions on S. If S is complete, then $C_b(S)$ is a Banach space under the supremum norm $||f|| = \sup_{x\in S} |f(x)|$. In addition, let $\mathcal{P}(S)$ denote the collection of probability measures on S.

6.1 Prohorov Metric

Definition 6.1. *For $\mu, \nu \in \mathcal{P}(S)$, define the* Prohorov metric

$$\rho(\nu, \mu) = \inf\{\epsilon > 0; \mu(F) \leq \nu(F^\epsilon) + \epsilon \text{ for all closed sets } F\}.$$

where

$$F^\epsilon = \{x \in S; \inf_{\tilde{x}\in F} d(x, \tilde{x}) < \epsilon\}.$$

the ϵ neighborhood of F. Note that this set is open.

We next show that ρ deserves the name metric.

Lemma 6.2. *Let $\mu, \nu \in \mathcal{P}(S)$ and $\epsilon, \eta > 0$. If*

$$\mu(F) \leq \nu(F^\epsilon) + \eta.$$

for all closed sets F, then

$$\nu(F) \leq \mu(F^\epsilon) + \eta.$$

for all closed sets F,

Proof. Given a closed set $\tilde{F}$, then $F = S\backslash\tilde{F}^\epsilon$ is closed and $\tilde{F} \subset S\backslash F^\epsilon$. Consequently,

$$\mu(\tilde{F}^\epsilon) = 1 - \mu(F) \geq 1 - \nu(F^\epsilon) - \eta \geq \nu(\tilde{F}) - \eta.$$

□

Exercise 6.3. *For any set A, $\lim_{\epsilon\to 0} A^\epsilon = \bar{A}$.*

Proposition 6.4. *The Prohorov metric is a metric.*

Proof. 1. (identity) If $\rho(\mu, \nu) = 0$, then $\mu(F) = \nu(F)$ for all closed F and hence for all sets in $\mathcal{B}(S)$.

2. (symmetry) This follows from the lemma above.

3. (triangle inequality) Let $\kappa, \mu, \nu \in \mathcal{P}(S)$ with

$$\rho(\kappa, \mu) > \epsilon_1, \quad \rho(\mu, \nu) > \epsilon_2.$$

Then, for any closed set

$$\kappa(F) \leq \mu(F^{\epsilon_1}) + \epsilon_1 \leq \mu(\overline{F^{\epsilon_1}}) + \epsilon_1 \leq \nu(\overline{F^{\epsilon_1}}^{\epsilon_2}) + \epsilon_1 + \epsilon_2 \leq \nu(F^{\epsilon_1+\epsilon_2}) + \epsilon_1 + \epsilon_2.$$

So,

$$\rho(\kappa, \nu) \leq \epsilon_1 + \epsilon_2$$

□

Exercise 6.5. *Let $S = \mathbb{R}$, by considering the closed sets $(-\infty, x]$ and the Prohorov metric, we obtain the* Lévy metric *for distribution function on $\mathbb{R}$. For two distributions F and G, define*

$$\rho_L(F, G) = \inf\{\epsilon > 0; G(x - \epsilon) - \epsilon \leq F(x) \leq G(x + \epsilon) + \epsilon\}.$$

1. *Verify that ρ_L is a metric.*
2. *Show that the sequence of distribution F_n converges to F in the Lévy metric if and only if*

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

for all x which are continuity points of F

Exercise 6.6. *If $\{x_k; k \geq 1\}$ is a dense subset of (S, d), then*

$$\{\sum_{k\in A} \alpha_k \delta_{x_k}; A \text{ is finite}, \alpha_k \in \mathbb{Q}^+, \sum_{k\in A} \alpha_k = 1\}.$$

is a dense subset of $(\mathcal{P}(S), \rho)$. This, if (S, d) is separable, so is $(\mathcal{P}(S), \rho)$

With some extra work, we can show that if (S, d) is complete, then so is $(\mathcal{P}(S), \rho)$.

6.2 Weak Convergence

Recall the definition:

Definition 6.7. *A sequence $\{\nu_n; n \geq 1\} \subset \mathcal{P}(S)$ is said to* converge weakly *to $\nu \in \mathcal{P}(S)$ ($\nu_n \Rightarrow \nu$) if*

$$\lim_{n\to\infty} \int_S f(x)\ \nu_n(dx) = \int_S f(x)\ \nu(dx) \text{ for all } f \in C_b(S).$$

A sequence $\{X_n; n \geq 1\}$ of S-valued random variables is said to converge in distribution *to X if*

$$\lim_{n\to\infty} E[f(X_n)] = E[f(X)] \text{ for all } f \in C_b(S).$$

Thus, X_n converges in distribution to X if and only if the distribution of X_n converges weakly to the distribution of X.

Exercise 6.8. *Let $S = [0, 1]$and define $\nu_n\{x\} = 1/n$, $x = k/n$, $k = 0, \ldots, n-1$. Thus, $\nu_n \Rightarrow \nu$, the uniform distribution on $[0, 1]$. Note that $\nu_n(\mathbb{Q} \cap [0, 1]) = 1$ but $\nu(\mathbb{Q} \cap [0, 1]) = 0$*

Definition 6.9. *Recall that the* boundary *of a set $A \subset S$ is given by $\partial A = \bar{A} \cap \overline{A^c}$. A is called a ν-continuity set if $\nu \in \mathcal{P}(S)$, $A \in \mathcal{B}(S)$, and*

$$\nu(\partial A) = 0,$$

Theorem 6.10 (portmanteau)**.** *Let (S, d) be separable and let $\{\nu_k; k \geq 1\} \cup \nu \subset \mathcal{P}(S)$. Then the following are equivalent.*

1. $\lim_{k\to\infty} \rho(\nu_k, \nu) = 0$.

2. $\nu_k \Rightarrow \nu$ *as* $k \to \infty$.

3. $\lim_{k\to\infty} \int_S h(x)\ \nu_k(dx) = \int_S h(x)\ \nu(dx)$ *for all uniformly continuous* $h \in C_b(S)$.

4. $\limsup_{k\to\infty} \nu_k(F) \le \nu(F)$ *for all closed sets* $F \subset S$.

5. $\liminf_{k\to\infty} \nu_k(G) \ge \nu(G)$ *for all open sets* $G \subset S$.

6. $\lim_{k\to\infty} \nu_k(A) = \nu(A)$ *for all* ν*-continuity sets* $A \subset S$.

Proof. $(1 \to 2)$ Let $\epsilon_k = \rho(\nu_k, \nu) + 1/k$ and choose a nonnegative $h \in C_b(S)$. Then for every k,

$$\int h\ d\nu_k = \int_0^{||h||} \nu_k\{h \ge t\}\ dt \le \int_0^{||h||} \nu\{h \ge t\}^{\epsilon_k}\ dt + \epsilon_k||h||$$

Noting that $\{h \ge t\}$ is a closed set.

$$\limsup_{k\to\infty} \int h\ d\nu_k \le \lim_{k\to\infty} \int_0^{||h||} \nu\{h \ge t\}^{\epsilon_k}\ dt = \int_0^{||h||} \nu\{h \ge t\}\ dt = \int h\ d\nu.$$

Apply this inequality to $||h|| + h$ and $||h|| - h$ to obtain

$$\limsup_{k\to\infty} \int (||h|| + h)\ d\nu_k \le \int (||h|| + h)\ d\nu,\ \limsup_{k\to\infty} \int (||h|| - h)\ d\nu_k \le \int (||h|| - h)\ d\nu.$$

Now, combine these two inequalities to obtain 2.

$(2 \to 3)$ is immediate.

$(3 \to 4)$ For F closed, define $d(x, F) = \inf_{\tilde{x}\in F} d(\tilde{x}, x)$ and

$$h_\epsilon(x) = \max\{\left(1 - \frac{d(x,F)}{\epsilon}\right), 0\}.$$

Then h_ϵ is uniformly continuous, $h_\epsilon \ge I_F$, and because F is closed,

$$\lim_{\epsilon\to 0} h_\epsilon(x) = I_F(x).$$

Thus, for each $\epsilon > 0$,

$$\limsup_{k\to\infty} \nu_k(F) \le \lim_{k\to\infty} \int h_\epsilon\ d\nu_k = \int h_\epsilon\ d\nu$$

and, therefore,

$$\limsup_{k\to\infty} \nu_k(F) \le \lim_{\epsilon\to 0} \int h_\epsilon\ d\nu = \nu(F).$$

$(4 \to 5)$ For every open set $G \subset S$,

$$\liminf_{k\to\infty} \nu_k(G) = 1 - \limsup_{k\to\infty} \nu_k(G^c) \ge 1 - \nu(G^c) = \nu(G).$$

$(5 \to 6)$ Note that $\text{int}A = A\backslash\partial A$ and $\bar{A} = A \cup \partial A$. Then

$$\limsup_{k\to\infty} \nu_k(A) \le \limsup_{k\to\infty} \nu_k(\bar{A}) = 1 - \liminf_{k\to\infty} \nu_k((\bar{A})^c) \le 1 - \nu((\bar{A})^c) = \nu(\bar{A}) = \nu(A)$$

and

$$\liminf_{k\to\infty} \nu_k(A) \ge \liminf_{k\to\infty} \nu_k(\text{int}(A)) \ge \nu(\text{int}(A)) = \nu(A).$$

$(6 \to 2)$ Choose a non-negative function $h \in C_b(S)$. Then $\partial\{h \ge t\} \subset \{h = t\}$. So $\{h \ge t\}$ is a ν-continuity set for all but at most countably many $t \ge 0$. Therefore, $\nu_k\{h \ge t\} \to \nu\{h \ge t\}$ as $t \to \infty$ for (Lebesgue) almost all t.

$$\lim_{k\to\infty} \int h \, d\nu_k = \lim_{k\to\infty} \int_0^{||h||} \nu_k\{h \ge t\} \, dt = \int_0^{||h||} \nu\{h \ge t\} dt = \int h \, d\nu.$$

Now consider the positive and negative parts of an arbitrary function in $C_b(S)$.

$(5 \to 1)$ Let $\epsilon > 0$ and choose a countable partition $\{A_j; j \ge 1\}$ of Borel sets whose diameter is at most $\epsilon/2$. Let J be the least integer satisfying

$$\nu(\bigcup_{j=1}^{J} A_j) > 1 - \frac{\epsilon}{2}$$

and let

$$\mathcal{G}_\epsilon = \{(\bigcup_{j\in C} A_j)^{\epsilon/2}; C \subset \{1, \cdots, J\}\}.$$

Note that $\mathcal{G}_\epsilon$ is a finite collection of open sets, Whenever 5 holds, there exists an integer K so that

$$\nu(G) \le \nu_k(G) + \frac{\epsilon}{2}, \text{ for all } k \ge K \text{ and for all } G \in \mathcal{G}_\epsilon.$$

Now choose a closed set F and define

$$F_0 = \bigcup\{A_j; 1 \le j \le J, A_j \cap F \ne \emptyset\}.$$

Then $F_0^{\epsilon/2} \in \mathcal{G}_\epsilon$, $F \subset F_0^{\epsilon/2} \cup \left(S\backslash(\bigcup_{j=1}^{J} A_j)\right)$, and

$$\nu(F) \le \nu(F_0^{\epsilon/2}) + \frac{\epsilon}{2} \le \nu_k(F_0^{\epsilon/2}) + \epsilon \le \nu_k(F^\epsilon) + \epsilon$$

for all $k \ge K$. Hence $\rho(\nu_k, \nu) \le \epsilon$ for all $k \ge K$. □

Exercise 6.11. *1. (continuous mapping theorem) Let h be a measurable function and let D_h be the discontinuity set of h. If $X_n \to^{\mathcal{D}} X$ and if $P\{X \in D_h\} = 0$, then $h(X_n) \to^{\mathcal{D}} h(X)$.*

2. If the distribution functions F_n on $\mathbb{R}$ converge to F for all continuity points of F, and $h \in C_b(\mathbb{R})$ then

$$\lim_{n\to\infty} \int h(x) \, dF_n(x) = \int h(x) \, dF(x).$$

3. *If $F_n, n \geq 1$ and F are distribution functions and $F_n(x) \to F(x)$ for all x. Then F continuous implies*

$$\lim_{n\to\infty} \sup_x |F_n(x) - F(x)| = 0.$$

4. *If $\{X_n; n \geq 1\}$ take values on a discrete set D, then $X_n \to^{\mathcal{D}} X$ if and only if*

$$\lim_{n\to\infty} P\{X_n = x\} = P\{X = x\} \text{ for all } x \in D.$$

5. *If $X_n \to^{\mathcal{D}} c$ for some constant c, then $X_n \to^{P} c$*

6. *Assume that $\nu_n \Rightarrow \nu$ and let $h, g : S \to \mathbb{R}$ be continuous functions satisfying*

$$\lim_{x\to\pm\infty} |g(x)| = \infty, \quad \lim_{x\to\pm\infty} \left|\frac{h(x)}{g(x)}\right| = 0.$$

Show that

$$\limsup_{n\to\infty} \int |g(x)|\ \nu_n(dx) < \infty \text{ implies } \lim_{n\to\infty} \int h(x)\ \nu_n(dx) = \int h(x)\ \nu(dx).$$

Consider of the families of discrete random variables and let $\{\nu_{\theta_n}; n \geq 1\}$ be a collection of distributions from that family. Then $\nu_{\theta_n} \Rightarrow \nu_\theta$ if and only if $\theta_n \to \theta$. For the families of continuous random variables, we have the following.

Theorem 6.12. *Assume that the probability measures $\{\nu_n; n \geq 1\}$ are mutually absolutely continuous with respect to a σ-finite measure μ with respective densities $\{f_n; n \geq 1\}$. If $f_n \to f$, μ-almost everywhere, then $\nu_n \Rightarrow \nu$.*

Proof. Let G be open, then by Fatou's lemma,

$$\liminf_{k\to\infty} \nu_k(G) = \liminf_{k\to\infty} \int_G f_k\ d\mu_k \geq \int_G f\ d\mu = \nu(G)$$

□

Exercise 6.13. *Assume that $c_k \to 0$ and $a_k \to \infty$ and that $a_k c_k \to \lambda$, then $(1 + c_k)^{a_k} \to \exp \lambda$*

Example 6.14.
1. *Let T_n have a $t(0, 1)$-distribution with n degrees of freedom. Then the densities of T_n converge to the density of a standard normal random variable. Consequently, the T_n converge in distribution to a standard normal.*

2. *(waiting for rare events) Let X_p be $Geo(p)$. Then $P\{X > n\} = (1 - p)^n$ Then*

$$P\{pX_p > x\} = (1 - p)^{[x/p]}.$$

Therefore pX_p converges in distribution to an $Exp(1)$ random variable.

Exercise 6.15.
1. *Let X_n be $Bio(n, p)$ with $np = \lambda$. Then X_n converges in distribution to a $Pois(\lambda)$ random variable.*

2. *If $X_n \to^{\mathcal{D}} X$ and $Y_n \to^{\mathcal{D}} c$ where c is a constant, then $X_n + Y_n \to^{\mathcal{D}} X + c$. A corollary is that if $X_n \to^{\mathcal{D}} X$ and $Z_n - X_n \to^{\mathcal{D}} 0$, then $Z_n \to^{\mathcal{D}} X$.*

3. *If $X_n \to^{\mathcal{D}} X$ and $Y_n \to^{\mathcal{D}} c$ where c is a constant, then $X_n Y_n \to^{\mathcal{D}} cX$.*

Example 6.16. *1. (birthday problem) Let $X_1, X_2, \ldots$ be independent and uniform on $\{1, \ldots, N\}$. Let $T_N = \min\{n : X_n = X_m \text{ for some } m < n\}$. Then*

$$P\{T_N > n\} = \prod_{m=2}^{n} \left(1 - \frac{m-1}{N}\right).$$

By the exercise above,

$$\lim_{N\to\infty} P\{\frac{T_N}{\sqrt{N}} > x\} = \exp\left(-\frac{x^2}{2}\right).$$

For the case $N = 365$,

$$P\{T_N > n\} \approx \exp\left(-\frac{n^2}{730}\right).$$

The choice $n = 22$ gives probability 0.515. An exact computation gives 0.524.

2. *(central order statistics) For $2n+1$ observations of independent $U(0,1)$ random variables, $X_{(n+1)}$ the one in the middle is $Beta(n,n)$ and thus has density*

$$(2n+1)\binom{2n}{n} x^n (1-x)^n$$

with respect to Lebesgue measure on $(0,1)$. This density is concentrating around 1/2 with variance

$$\frac{n^2}{(2n)^2(2n+1)} \approx \frac{1}{8n}$$

Thus we look at

$$Z_n = (X_{(n+1)} - \frac{1}{2})\sqrt{8n}$$

which have mean 0 and variance near to one. Then Z_n has density

$$(2n+1)\binom{2n}{n}\left(\frac{1}{2} + \frac{z}{\sqrt{8n}}\right)^n \left(\frac{1}{2} - \frac{z}{\sqrt{8n}}\right)^n \frac{1}{\sqrt{8n}} = \binom{2n}{n} 2^{-2n} \left(1 - \frac{z^2}{2n}\right)^n \frac{2n+1}{2n}\sqrt{\frac{n}{2}}.$$

Now use Sterling's formula to see that this converges to

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

6.3 Prohorov's Theorem

If (S,d) is a complete and separable metric space, then $\mathcal{P}(S)$ is a complete and separable metric space under the Prohorov metric ρ. One common approach to proving the metric convergence $\nu_n \Rightarrow \nu$ is first to verify that $\{\nu_k; k \geq 1\}$ is a *relatively compact* set, i.e., a set whose closure is compact, then this sequence has limit points. Thus, we can obtain convergence by showing that this set has at most one limit point.

In the case of complete and separable metric spaces, we will use that at set C is compact if and only it is closed and *totally bounded*, i.e., for every $\epsilon > 0$ there exists a finite number of points $\nu_1, \ldots, \nu_n \in C$ so that

$$C \subset \bigcup_{k=1}^{n} B_\rho(\nu_k, \epsilon).$$

Definition 6.17. *A collection $\mathcal{A}$ of probabilities on a topological space S is* tight *if for each $\epsilon > 0$, then exists a compact set $K \subset S$*

$$\nu(K) \geq 1 - \epsilon, \text{ for all } \nu \in \mathcal{A}.$$

Lemma 6.18. *If (S,d) is complete and separable then any one point set $\{\nu\} \subset \mathcal{P}(S)$ is tight.*

Proof. Choose $\{x_k; k \geq 1\}$ dense in S. Given $\epsilon > 0$, choose integers $N_1, N_2, \ldots$ so that for all n,

$$\nu(\bigcup_{k=1}^{N_n} B_d(x_k, \frac{1}{n})) \geq 1 - \frac{\epsilon}{2^k}.$$

Define K to be the closure of

$$\bigcap_{n=1}^{\infty} \bigcup_{k=1}^{N_n} B_d(x_k, \frac{1}{n}).$$

Then K is totally bounded and hence compact. In addition,

$$\nu(K) \geq 1 - \sum_{n=1}^{\infty} \frac{\epsilon}{2^n} = 1 - \epsilon.$$

□

Exercise 6.19. *A sequence $\{\nu_m; n \geq 1\} \subset \mathcal{P}(\mathcal{S})$ is tight if and only if for every $\epsilon > 0$, there exists a compact set K so that*

$$\liminf_{n \to \infty} \nu_n(K) > 1 - \epsilon.$$

Exercise 6.20. *Assume that $h : \mathbb{R}^+ \to \mathbb{R}$ satisfies*

$$\lim_{s \to \infty} h(s) = \infty.$$

Let $\{\nu_\lambda; \lambda \in \Lambda\}$ be a collection probabilities on $\mathbb{R}^d$ satisfying

$$\sup\{\int h(|x|)\ \nu_\lambda(dx); \lambda \in \Lambda\} < \infty.$$

Then, $\{\nu_\lambda; \lambda \in \Lambda\}$ is tight.

Theorem 6.21 (Prohorov)**.** *Let (S,d) be complete and separable and let $\mathcal{A} \subset \mathcal{P}(S)$. Then the following are equivalent:*

1. *$\mathcal{A}$ is tight.*
2. *For each $\epsilon > 0$, then exists a compact set $K \subset S$*

$$\nu(K^\epsilon) \geq 1 - \epsilon, \text{ for all } \nu \in \mathcal{A}.$$

3. *$\mathcal{A}$ is relatively compact.*

Proof. $(1 \to 2)$ is immediate.

$(2 \to 3)$ We show that $\mathcal{A}$ is totally bounded. So, given $\eta > 0$, we must find a finite set $\mathcal{N} \subset \mathcal{P}(S)$ so that

$$\mathcal{A} \subset \{\mu : \rho(\nu,\mu) < \eta \text{ for some } \nu \in \mathcal{N}\} = \bigcup_{\nu \in \mathcal{N}} B_\rho(\mu, \eta).$$

Fix $\epsilon \in (0, \eta/2)$ and choose a compact set K satisfying 2. Then choose $\{x_1, \ldots, x_n\} \subset K$ such that

$$K^\epsilon \subset \bigcup_{k=1}^{n} B_d(x_k, 2\epsilon).$$

Fix $x_0 \in S$ and $M \geq n/\epsilon$ and let

$$\mathcal{N} = \{\nu = \sum_{j=0}^{n} \frac{m_i}{M} \delta_{x_j}; 0 \leq m_j, \sum_{j=0}^{n} m_j = M\}.$$

To show that every $\mu \in \mathcal{A}$ is close to some probability in $\mathcal{N}$, Define,

$$A_j = B_d(x_j, 2\epsilon) \setminus \bigcap_{k=1}^{j-1} B_d(x_k, 2\epsilon), \; k_j = [M\mu(A_j)], \; k_0 = M - \sum_{j=1}^{n} m_j$$

and use this to choose $\nu \in \mathcal{N}$. Then, for any closed set F,

$$\mu(F) \leq \mu\left(\bigcup \{A_j : F \cap A_j \neq \emptyset\}\right) + \epsilon \leq \sum_{\{j : F \cap A_j \neq \emptyset\}} \frac{[M\mu(A_j)] + 1}{M} + \epsilon \leq \nu(F^{2\epsilon}) + 2\epsilon.$$

Thus $\rho(\nu, \mu) < 2\epsilon < \eta$.

$(3 \to 1)$ Because $\mathcal{A}$ is totally bounded, there exists, for each $n \in \mathbb{N}$, a finite set $\mathcal{N}_n$ such that

$$\mathcal{A} \subset \{\mu : \rho(\nu,\mu) < \frac{\epsilon}{2^{n+1}} \text{ for some } \nu \in \mathcal{N}_n\}.$$

By the lemma, choose a compact set K_n so that

$$\nu(K_n) \geq 1 - \frac{\epsilon}{2^n} \text{ for all } \nu \in \mathcal{A}.$$

Given $\mu \in \mathcal{A}$, there exists $\nu_n \in \mathcal{N}_n$ so that

$$\mu(K_n^{\epsilon/2^{n+1}}) \geq \nu_n(K_n) - \frac{\epsilon}{2^{n+1}} \geq 1 - \frac{\epsilon}{2^n}.$$

Now, note that K, the closure of

$$\bigcap_{n=1}^{\infty} K_n^{\epsilon/2^{n+1}}$$

is compact and that

$$\mu(K) \geq 1 - \sum_{n=1}^{\infty} \frac{\epsilon}{2^n} = 1 - \epsilon.$$

□

Of course, it is the case $1 \to 3$ that will attract the most attention.

6.4 Separating and Convergence Determining Sets

We now use the tightness criterion based on the Prohorov metric to give us assistance in determining weak limits. The goal in this section is to reduce the number of test functions needed for convergence. We begin with two definitions.

Definition 6.22. *1. A set $H \subset C_b(S)$ is called* separating *if for any $\mu, \nu \in \mathcal{P}(S)$,*

$$\int h \, d\mu = \int h \, d\nu \text{ for all } h \in H$$

implies $\mu = \nu$.

2. A set $H \subset C_b(S)$ is called convergence determining *if for any sequence $\{\nu_n; n \geq 1\} \subset \mathcal{P}(S)$ and $\nu \in \mathcal{P}(S)$,*

$$\lim_{n\to\infty} \int h \, d\nu_n = \int h \, d\nu \text{ for all } h \in H$$

implies $\nu_n \Rightarrow \nu$.

Example 6.23. *If $S = \mathbb{N}$, then by the uniqueness of power series, the collection $\{z^x;, 0 \leq z \leq 1\}$ is separating. Take $\nu_k = \delta_k$ to see that it is not convergence determining.*

Exercise 6.24. *1. $C_b(S)$ is convergence detemining.*

2. Convergence determining sets are separating.

For a converse in the case of tightness, we have:

Proposition 6.25. *Let $\{\nu_n; n \geq 1\} \subset \mathcal{P}(S)$ be relatively compact and let $H \subset C_b(S)$ be separating. Then $\nu \Rightarrow \nu$ if and only if*

$$\lim_{n\to\infty} \int h \, d\nu_n$$

exists for all $h \in H$. In this case, the limit is $\int h \, d\nu$.

Proof. Let $\tilde{\nu}$ and $\tilde{\mu}$ be a weak limits of $\{\nu_n; n \geq 1\}$, then for some subsequences $\{n_k; k \geq 1\}$, and $\{m_k; k \geq 1\}$,

$$\lim_{k\to\infty} \int h \, d\nu_{n_k} = \int h \, d\tilde{\nu} \text{ and } \lim_{k\to\infty} \int h \, d\nu_{m_k} = \int h \, d\tilde{\mu} \text{ for all } h \in H$$

Thus,

$$\int h \, d\tilde{\nu} = \int h \, d\tilde{\mu} \text{ for all } h \in H$$

and because H is separating $\tilde{\nu} = \tilde{\mu}$, and $\nu_n \Rightarrow \tilde{\nu}$. □

Exercise 6.26. *Let K be a compact metric space set $f_n : K \to \mathbb{R}$ be continuous. If, for all $z \in K$,*

$$\lim_{n\to\infty} f_n(z) = f(z),$$

a continuous function, then the convergence is uniform.

Theorem 6.27. *Let $\{X_n; n \geq 1\}$ be $\mathbb{N}$-valued random variables having respective generating function $g_n(z) = Ez^{X_n}$. If*

$$\lim_{n\to\infty} g_n(z) = g(z),$$

and g is continuous at 1, then X_n converges in distribution to a random variable X with generating function g.

Proof. Let $z \in [0, 1)$ and choose $\tilde{z} \in (z, 1)$. Then for each n and k

$$P\{X_n = k\}z^k < \tilde{z}^k.$$

Thus, by the Weierstrass M-test, g_n converges uniformly to g on $[0, \tilde{z}]$ and thus g is continuous at z. Thus, by hypothesis, g is an analytic function on $[0, 1]$.

$$\begin{aligned}
\lim_{n\to\infty} P\{X_n > x\} &= \lim_{n\to\infty} \lim_{z\to 1} \left(g_n(z) - \sum_{k=1}^{x} P\{X_n = k\}z^k \right) \\
&= \lim_{z\to 1} \lim_{n\to\infty} \left(g_n(z) - \sum_{k=1}^{x} P\{X_n = k\}z^k \right) = \lim_{z\to 1} \left(g(z) - \sum_{k=1}^{x} g^{(k)}(0)z^k \right) \\
&= g(1) - \sum_{k=1}^{x} g^{(k)}(0) < \epsilon
\end{aligned}$$

by choosing x sufficiently large. Thus, we have that $\{X_n; n \geq 1\}$ is tight and hence relatively compact. Because $\{z^x; , 0 \leq z \leq 1\}$ is separating, we have the theorem. □

Example 6.28. *Let X_n be a Bin(n, p) random variable. Then*

$$Ez^{X_n} = ((1 - p) + pz)^n$$

Set $\lambda = np$, then

$$\lim_{n\to\infty} Ez^{X_n} = \lim_{n\to\infty} (1 + \frac{\lambda}{n}(z - 1))^n = \exp \lambda(z - 1),$$

the generating function of a Poisson random variable. The convergence of the distributions of $\{X_n; n \geq 1\}$ follows from the fact that the limiting function is continuous at $z = 1$.

We will now go on to show that if H separates points then it is separating. We recall a definition,

Definition 6.29. *A collection of functions $H \subset C_b(S)$ is said to* separate points *if for every distinct pair of points $x_1, x_2 \in S$, there exists $h \in H$ such that $h(x_1) \neq h(x_2)$.*

. . . and a generalization of the Weierstrass approximation theorem.

Theorem 6.30 (Stone-Weierstrass)**.** *Assume that S is compact. Then $C(S)$ is an algebra of functions under pointwise addition and multiplication. Let A be a sub-algebra of $C(S)$ that contains the constant functions and separates points then A is dense in $C(S)$ under the topology of uniform convergence.*

Theorem 6.31. *Let (S, d) be complete and separable and let $H \subset C_b(S)$ be an algebra. If H separates points, the H is separating.*

Proof. Let $\mu, \nu \in \mathcal{P}(S)$ and define

$$M = \{h \in C_b(S); \int h \, d\mu = \int h \, d\nu\}.$$

If $H \subset M$, then the closure of the algebra $\tilde{H} = \{a + h; h \in H, a \in \mathbb{R}\}$ is contained in M.

Let $h \in C_b(S)$ and let $\epsilon > 0$. By a previous lemma, the set $\{\mu, \nu\}$ is tight. Choose K compact so that

$$\mu(K) \geq 1 - \epsilon, \quad \nu(K) \geq 1 - \epsilon.$$

By the Stone-Weierstrass theorem, there exists a sequence $\{h_n; n \geq 1\} \subset \tilde{H}$ such that

$$\lim_{n \to \infty} \sup_{x \in K} |h_n(x) - h(x)| = 0.$$

Because h_n may not be bounded on K^c we replace it with $h_{n,\epsilon}(x) = h_n(x) \exp(-\epsilon h_n(x)^2)$. Note that $h_{n,\epsilon}$ is in the closure of $\tilde{H}$ Define h_ϵ similarly.

Now observe that for each n

$$\begin{aligned}
\Big| \int_S h_n \, d\mu - \int_S h_n \, d\nu \Big| \quad \leq \quad & \Big| \int_S h_n \, d\mu - \int_K h_n d\mu \Big| + \Big| \int_K h_n \, d\mu - \int_K h_{n,\epsilon} \, d\mu \Big| + \Big| \int_K h_{n,\epsilon} \, d\mu - \int_S h_{n,\epsilon} \, d\mu \Big| \\
+ \quad & \Big| \int_S h_{n,\epsilon} \, d\mu - \int_S h_{n,\epsilon} \, d\nu \Big| \\
+ \quad & \Big| \int_S h_{n,\epsilon} \, d\nu - \int_K h_{n,\epsilon} \, d\nu \Big| + \Big| \int_K h_{n,\epsilon} \, d\nu - \int_K h_n \, d\nu \Big| + \Big| \int_K h_n \, d\nu - \int_S h_n \, d\nu \Big|
\end{aligned}$$

For the seven terms, note that:

- The fourth term is zero because $h_{\epsilon,n}$ is in the closure of $\tilde{H}$.
- The second and sixth terms tend to zero as $n \to \infty$ by the uniform convergence of $h_{n\epsilon}$ to h_ϵ.
- The remaining terms are integrals over $S \backslash K$, a set that has both ν and μ measure at most ϵ. The integrands are bounded by $1/\sqrt{2e\epsilon}$.

Thus, letting $\epsilon \to 0$ we obtain that $M = C_b(S)$. □

This creates for us an easy method of generating separating classes. So, for example, polynomials (for compact spaces), trigonometric polynomials, n-times continuously differentiable and bounded functions are separating classes.

6.5 Characteristic Functions

Recall that the characteristic function for a probability measure on $\mathbb{R}^d$ is

$$\phi(\theta) = \int e^{i\langle\theta,x\rangle}\ \nu(dx) = Ee^{i\langle\theta,X\rangle}$$

if X is a random variable with distribution ν. Sometimes we shall write ϕ_ν of ϕ_X if more than one characteristics function is under discussion.

Because the functions $\{e^{i\langle\theta,x\rangle}; \theta \in \mathbb{R}^d\}$ for an algebra that separates points, this set is separating. This is just another way to say that the Fourier transform is one-to-one.

Some additional properties of the characteristic function are:

1. *For all* $\theta \in \mathbb{R}^d$,
$$|\phi(\theta)| \le 1 = \phi(0).$$

2. *For all* $\theta \in \mathbb{R}^d$,
$$\phi(-\theta) = \overline{\phi(\theta)}.$$

3. *The characteristic function* ϕ *is uniformly continuous in* $\mathbb{R}^d$.
For all $\theta, h \in \mathbb{R}^d$,
$$\phi(\theta+h) - \phi(\theta) = \int (e^{i\langle\theta+h,x\rangle} - e^{i\langle\theta,x\rangle})\ \nu(dx) = \int e^{i\langle\theta,x\rangle}(e^{i\langle h,x\rangle} - 1)\ \nu(dx).$$
Therefore,
$$|\phi(\theta+h) - \phi(\theta)| \le \int |e^{i\langle h,x\rangle} - 1|\ \nu(dx).$$
This last integrand is bounded by 2 and has limit 0 as $h \to 0$ for each $x \in \mathbb{R}^d$. Thus, by the bounded convergence theorem, the integral has limit 0 as $h \to 0$. Because the limit does not involve θ, it is uniform.

4. *Let* $a \in \mathbb{R}$ *and* $b \in \mathbb{R}^d$, *then*
$$\phi_{aX+b}(\theta) = \phi(a\theta)e^{i\langle\theta,b\rangle}.$$
Note that
$$Ee^{i\langle\theta,aX+b\rangle} = e^{i\langle\theta,b\rangle}Ee^{i\langle a\theta,X\rangle}.$$

5. $\phi_{-X}(\theta) = \overline{\phi_X(\theta)}$. *Consequently,* X *has a symmetric distribution if and only if its characteristic function is real.*

6. *If* $\{\phi_j; j \ge 1\}$ *are characteristic functions and* $\lambda_j \ge 0$, $\sum_{j=1}^\infty \lambda_j = 1$, *then the mixture*
$$\sum_{j=1}^\infty \lambda_j \phi_j$$
is a characteristic function.
If ν_j has characteristic function ϕ_j, then $\sum_{j=1}^\infty \lambda_j \nu_j$ is a probability measure with characteristic function $\sum_{j=1}^\infty \lambda_j \phi_j$.

7. *If $\{\phi_j; n \geq j \geq 1\}$ are characteristic functions, then*

$$\prod_{j=1}^{n} \phi_j$$

is a characteristic function.

If the ϕ_j are the characteristic functions for independent random variable X_j, then the product above is the characteristic function for their sum.

Exercise 6.32. *If ϕ is a characteristic function, then so is $|\phi|^2$.*

Exercise 6.33.

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

Hint: *Write the error term in Taylor's theorem in two ways:*

$$\frac{i^n}{n!} \int_0^x (x-t)^n e^{it}\ dt = \frac{i^{n+1}}{(n-1)!} \int_0^x (x-t)^{n-1}(e^{it}-1)\ dt.$$

One immediate consequence of this is that

$$|Ee^{i\theta X} - (1 + i\theta EX - \frac{\theta}{2} EX^2)| \leq \frac{\theta^2}{6} E\left[\min\{|\theta||X|^3, 6|X|^2\}\right].$$

Note in addition, that the dominated convergence theorem implies that the expectation on the right tends to 0 as $\theta \to 0$.

Exercise.

1. Let X_i, $i = 1, 2$ be independent $Cau(\mu_i, 0)$, then $X_1 + X_2$ is $Cau(\mu_1 + \mu_2, 0)$.
2. Let X_i, $i = 1, 2$ be independent $\chi^2_{a_1}$, then $X_1 + X_2$ is $\chi^2_{a_1+a_2}$.
3. Let X_i, $i = 1, 2$ be independent $\Gamma(\alpha_i, \beta)$, then $X_1 + X_2$ is $\Gamma(\alpha_1 + \alpha_1, \beta)$.
4. Let X_i, $i = 1, 2$ be independent $N(\mu_i, \sigma_i^2)$, then $X_1 + X_2$ is $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Example 6.34 (t-distribution)**.** *Let $\{X_j; 1 \leq j \leq n\}$ be independent $N(\mu, \sigma^2)$ random variable. Set*

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j,\ S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2.$$

Check that

$$E\bar{X} = \mu,\ ES^2 = \sigma^2.$$

As before, define

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Check that the distribution of T is independent of affine transformations and thus we take the case $\mu = 0$, $\sigma^2 = 1$. We have seen that $\bar{X}$ is $N(0, 1/n)$ and is independent of S^2. We have the identity

$$\sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n} (X_j - \bar{X} + \bar{X})^2 = (n-1)S^2 + n\bar{X}^2.$$

(The cross term is 0.) Now

- *the characteristic function of the left equals the characteristic function of the right,*
- *the left is a χ_n^2 random variable,*
- *the terms on the right are independent, and*
- *the second term is χ_1^2.*

Thus, by taking characteristic functions, we have that

$$(1 - 2i\theta)^{-n/2} = \phi_{(n-1)S^2}(\theta)(1 - 2i\theta)^{-1/2}.$$

Now, divide to see that $(n-1)S^2$ is χ_{n-1}^2.

We now relate characteristic functions to convergence in distribution. First in dimension 1.

Theorem 6.35 (continuity theorem)**.** *Let $\{\nu_n; n \geq 1\}$ be probability measures on $\mathbb{R}$ with corresponding characteristic function $\{\phi_n; n \geq 1\}$ satisfying*

1. *$\lim_{n\to\infty} \phi_n(\theta)$ exists for all $\theta \in \mathbb{R}$, and*
2. *$\lim_{n\to\infty} \phi_n(\theta) = \phi(\theta)$ is continuous at zero. Then there exists $\nu \in \mathcal{P}(\mathbb{R})$ with characteristic function ϕ and $\nu_n \Rightarrow \nu$.*

Proof. All that needs to be shown is that the continuity of ϕ at 0 implies that $\{\nu_n; n \geq 1\}$ is tight. This can be seen from the following argument.

Note that

$$\int_{-t}^{t} (1 - e^{i\theta x})\, d\theta = 2t - \frac{e^{itx} - e^{-itx}}{ix} = 2t - \frac{2\sin tx}{x}.$$

Consequently,

$$\begin{aligned}
\frac{1}{t}\int_{-t}^{t} (1 - \phi_n(\theta))\, d\theta &= \frac{1}{t}\int_{-t}^{t}\int (1 - e^{i\theta x})\, \nu_n(dx)\, d\theta \\
&= \int \frac{1}{t}\int_{-t}^{t} (1 - e^{i\theta x})\, d\theta\, \nu_n(dx) = 2\int (1 - \frac{\sin tx}{tx})\, \nu_n(dx) \\
&\geq 2\int_{|x|\geq 2/t} \left(1 - \frac{1}{|tx|}\right)\, \nu_n(dx) \geq \nu_n\left\{x; |x| > \frac{2}{t}\right\}
\end{aligned}$$

Let $\epsilon > 0$. By the continuity of ϕ at 0, we can choose t so that

$$\frac{1}{t}\int_{-t}^{t} (1 - \phi(\theta))\, d\theta < \frac{\epsilon}{2}.$$

By the bounded convergence theorem, there exists N so that for all $n \geq N$,

$$\epsilon > \frac{1}{t}\int_{-t}^{t}(1-\phi_n(\theta))\,d\theta \geq \nu_n\{x;|x|>\frac{2}{t}\}$$

and $\{\nu_n; n \geq 1\}$ is tight. □

Now, we use can use the following to set the theorem in multidimensions.

Theorem 6.36 (Cramér-Wold devise)**.** *Let $\{X_n; n \geq 1\}$ be $\mathbb{R}^d$-valued random vectors. Then $X_n \to^{\mathcal{D}} X$ if and only if $\langle\theta, X_n\rangle \to^{\mathcal{D}} \langle\theta, X\rangle$ for all $\theta \in \mathbb{R}^d$.*

Proof. The necessity follows by considering the bounded continuous functions $h_\theta(x) = h(\langle\theta, x\rangle)$, $h \in C_b(S)$.

If $\langle\theta, X_n\rangle \to^{\mathcal{D}} \langle\theta, X\rangle$, then $\langle\theta, X_n\rangle$ is tight. Now take θ to be the standard basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_d$ and choose M_k so that

$$P\{-M_k \leq \langle\mathbf{e}_k, X_n\rangle \leq M_k\} \geq 1 - \frac{\epsilon}{d}.$$

Then the compact set $K = [-M_1, M_1] \times \cdots \times [-M_n, M_n]$ satisfies

$$P\{X_n \in K\} \geq 1 - \epsilon.$$

Consequently, $\{X_n; n \geq 1\}$ is tight.

Also, $\langle\theta, X_n\rangle \to^{\mathcal{D}} \langle\theta, X\rangle$ implies that

$$\lim_{n\to\infty} E[e^{is\langle\theta,X_n\rangle}] = E[e^{is\langle\theta,X\rangle}].$$

To complete the proof, take $s = 1$ and note that $\{\exp i\langle\theta, x\rangle; \theta \in \mathbb{R}^d\}$ is separating. □

7 Central Limit Theorems

7.1 The Classical Central Limit Theorem

Theorem 7.1. *Let $\{X_n; n \geq 1\}$ be and independent and identically distributed sequence of random variables having common mean μ and common variance σ^2. Write $S_n = X_1 + \cdots + X_n$, then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to^{\mathcal{D}} Z$$

where Z is a $N(0,1)$ random variable.

With the use of characteristc functions, the proof is now easy. First replace X_n with $X_n - \mu$ to reduce to the case of mean 0. Then note that if the X_n have characteristic function ϕ, then

$$\frac{S_n}{\sigma\sqrt{n}} \text{ has characteristic function } \phi\left(\frac{\theta}{\sigma\sqrt{n}}\right)^n$$

Note that

$$\phi\left(\frac{\theta}{\sigma\sqrt{n}}\right)^n = \left(1 - \frac{\theta^2}{2n} + \epsilon\left(\frac{\theta}{\sigma\sqrt{n}}\right)\right)^n$$

where $\epsilon(t)/t^2 \to 0$ as $t \to 0$. Thus,

$$\phi\left(\frac{\theta}{\sigma\sqrt{n}}\right)^n \to e^{-\theta^2/2}$$

and the theorem follows from the continuity theorem. This limit is true for real numbers. Because the exponential is not one-to-one on the complex plane, this argument needs some further refinement for complex numbers

Proposition 7.2. *Let $c \in \mathbb{C}$. Then*

$$\lim_{n\to\infty} c_n = c \quad \textit{implies} \quad \lim_{n\to\infty}\left(1 + \frac{c_n}{n}\right)^n = e^c.$$

Proof. We show first quickly establish two claims.

Claim I. Let $z_1, \ldots, z_n$ and $w_1, \ldots, w_n$ be complex numbers whose modulus is bounded above by M. Then

$$|z_1 \cdots z_n - w_1 \cdots w_n| \leq M^{n-1} \sum_{j=1}^{n} |z_j - w_j|. \tag{7.1}$$

For a proof by induction, note that the claim holds for $n = 1$. For $n > 1$, observe that

$$\begin{aligned} |z_1 \cdots z_n - w_1 \cdots w_n| &\leq |z_1 \cdots z_n - z_1 w_2 \cdots w_n| + |z_1 w_2 \cdots w_n - w_1 \cdots w_n| \\ &\leq M|z_2 \cdots z_n - w_2 \cdots w_n| + M^{n-1}|z_1 - w_1|. \end{aligned}$$

Claim II. For $w \in \mathbb{C}$, $|w| \leq 1$, $|e^w - (1 + w)| \leq |w|^2$.

$$e^w - (1+w) = \frac{w^2}{2!} + \frac{w^3}{3!} + \frac{w^4}{4!} + \cdots .$$

Therefore,

$$|e^w - (1+w)| \le \frac{|w|^2}{2}(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots) = |w|^2. \tag{7.2}$$

Now, choose $z_k = (1 + c_n/n)$ and $w_k = \exp(c_n/n)$, $k = 1, \ldots, n$. Let $\gamma = \sup\{|c_n|; n \ge 1\}$, then $\sup\{(1 + |c_n|/n), \exp(|c_n|/n); n \ge 1\} \le \exp \gamma/n$. Thus, as soon as $|c_n|/n \le 1$,

$$|\left(1 + \frac{c_n}{n}\right)^n - \exp c_n| \le (\exp \frac{\gamma}{n})^{n-1} n \left|\frac{c_n}{n}\right|^2 \le e^\gamma \frac{\gamma^2}{n}.$$

Now let $n \to \infty$. □

Exercise 7.3. *For $w \in \mathbb{C}$, $|w| \le 2$, $|e^w - (1+w)| \le 2|w|^2$.*

7.2 Infinitely Divisible Distributions

We have now seen two types of distributions be the limit of sums S_n of triangular arrays

$$\{X_{n,k}; n \ge 1, 1 \le k \le k_n\}$$

of independent random variables with $\lim_{n\to\infty} k_n = \infty$.

In the first, we chose $k_n = n$, $X_{n,k}$ to be $Ber(\lambda/n)$ and found the sum

$$S_n \to^{\mathcal{D}} Y$$

where Y is $Pois(\lambda)$.

In the second, we chose $k_n = n$, $X_{n,k}$ to be $X_k/\sqrt{n}$ with X_k having mean 0 and variance one and found the sum

$$S_n \to^{\mathcal{D}} Z$$

where Z is $N(0, 1)$.

The question arises: Can we see any other convergences and what trianguler arrays have sums that realize this convergence?

Definition 7.4. *Call a random variable X* infinitely divisible *if for each n, there exists independent and identically distributed sequence $\{X_{n,k}; 1 \le k \le n\}$ so that the sum $S_n = X_{n,1} + \cdots + X_{n,n}$ has the same distribution as X.*

Exercise 7.5. *Show that normal, Poisson, Cauchy, and gamma random variable are infinitely divisible.*

Theorem 7.6. *A random variable S is the weak limit of sums of a triangular array with each row $\{X_{n,k}; 1 \le k \le k_n\}$ independent and identically distributed if and only if S is infinitely divisible.*

Proof. Sufficiency follows directly from the definition,

To establish necessity, first, fix an integer K. Because each individual term in the triangular array converges in distribution to 0 as $n \to \infty$, we can assume that k_n is a multiple of K. Now, write

$$S_n = Y_{n,1} + \cdots + Y_{n,K}$$

where $Y_{j,n} = X_{(j-1)k_n/K+1,n} + \cdots + X_{jk_n/K,n}$ are independent and identically distributed.

Note that for $y > 0$,

$$P\{Y_{n,1} > y\}^K = \prod_{j=1}^{K} P\{Y_{n,j} > y\} \leq P\{S_n > Ky\}$$

and

$$P\{Y_{n,1} < -y\}^K \leq P\{S_n < -Ky\}.$$

Because the S_n have a weak limit, the sequence is tight. Consequently, $\{Y_{n,j}; n \geq 1\}$ are tight and has a weak limit along a subsequence

$$Y_{m_n,j} \to^{\mathcal{D}} Y_j$$

(Note that the same subsequential limit holds for each j.) Thus S has the same distribution as the sum $Y_1 + \cdots + Y_K$ □

7.3 Weak Convergence of Triangular Arrays

We now characterize an important subclass of infinitely divisible distributions and demonstrate how a triangular array converges to one of these distributions. To be precise about the set up:

For $n = 1, 2, \ldots$, let $\{X_{n,1}, \ldots, X_{n,k_n}\}$ be an independent sequence of random variables. Put

$$S_n = X_{1,n} + \cdots + X_{n,k_n}. \tag{7.3}$$

Write

$$\mu_{n,k} = EX_{n,k},\ \mu_n = \sum_{k=1}^{k_n} \mu_{n,k}, \quad \sigma^2_{n,k} = \mathrm{Var}(X_{n,k}),\ \sigma^2_n = \sum_{k=1}^{k_n} \sigma^2_{n,k}.$$

and assume

$$\sup_n \mu_n < \infty, \quad \text{and} \quad \sup_n \sigma^2_n < \infty.$$

To insure that the variation of no single random variable contributes disproportionately to the sum, we require

$$\lim_{n\to\infty} (\sup_{1\leq k\leq k_n} \sigma^2_{n,k}) = 0.$$

First, we begin with the characterization:

Theorem 7.7 (Lévy-Khinchin)**.** *ϕ is the characteristic function of an infinitely divisible distribution if and only if for some finite measure μ and some $b \in \mathbb{R}$,*

$$\phi(\theta) = \exp\left(ib\theta + \int_{\mathbb{R}} (e^{i\theta x} - 1 - i\theta x) \frac{1}{x^2}\ \mu(dx) \right). \tag{7.4}$$

In addition, this distribution has mean b and variance $\mu(\mathbb{R})$.

This formulation is called the the *canonical* or *Lévy-Khinchin representation* of ϕ. The measure μ is called the *canonical* or *Lévy measure.* Check that the integrand is continuous at 0 with value $-\theta^2/2$.

Exercise 7.8. *Verify that the characteristic function above has mean b and variance $\mu(\mathbb{R})$.*

We will need to make several observations before moving on to the proof of this theorem. To begin, we will need to obtain a sense of closeness for Lévy measures.

Definition 7.9. *Let (S, d) be a locally compact, complete and separable metric space and write $C_0(S)$ denote the space of continuous functions that "vanish at infinity" and $\mathcal{M}_F(S)$ the finite Borel measures on S. For $\{\mu_n; n \geq 1\}, \mu \in \mathcal{M}_F(S)$, we say that μ_n* converges vaguely to *μ and write $\mu_m \to^v \mu$ if*

1. *$\sup_n \mu_n(\mathbb{R}) < \infty$, and*
2. *for every $h \in C_0(\mathbb{R})$,*
$$\lim_{n\to\infty} \int_S h(x)\ \mu_n(dx) = \int_S h(x)\ \mu(dx).$$

This is very similar to weak convergence and thus we have analogous properties. For example,

1. Let A be a μ continuity set, then
$$\lim_{n\to\infty} \mu_n(A) = \mu(A).$$
2. $\sup_n \mu_n(\mathbb{R}) < \infty$ implies that $\{\mu_n; n \geq 1\}$ is relatively compact. This is a stronger statement than what is possible under weak convergence. The difference is based on the reduction of the space of test functions from continuous bounded functions to $C_0(S)$.

Write $e_\theta(x) = (e^{i\theta x} - 1 - i\theta x)/x^2, e_\theta(0) = -\theta^2$ Then $e_\theta \in C_0(\mathbb{R})$. Thus, if $b_n \to b$ and $\mu_n \to^v \mu$, then

$$\lim_{n\to\infty} \exp\left(ib_n\theta + \int (e^{i\theta x} - 1 - i\theta x)\frac{1}{x^2}\ \mu_n(dx)\right) = \exp\left(ib\theta + \int (e^{i\theta x} - 1 - i\theta x)\frac{1}{x^2}\ \mu(dx)\right).$$

Example 7.10.
1. *If $\mu = \sigma^2\delta_0$, then $\phi(\theta) = \exp(ib\theta - \sigma^2\theta^2/2)$, the characteristics function for a $N(b, \sigma^2)$ random variable.*
2. *Let N be a $Pois(\lambda)$ random variable, and set $X = x_0N$, then X is infinitely divisible with characteristic function*
$$\phi_X(\theta) = \exp(\lambda(e^{i\theta x_0} - 1)) = \exp(i\theta x_0\lambda + (e^{i\theta x_0} - 1 - i\theta x_0)\lambda).$$
Thus, this infinitely divisible distribution has mean $x_0\lambda$ and Lévy measure $x_0^2\lambda\delta_{x_0}$
3. *More generally consider a* compound Poisson random variable
$$X = \sum_{n=1}^{N} \xi_n$$

where the ξ_n are independent with distribution γ and N is a $Pois(\lambda)$ random variable independent of the X_n. Then

$$\begin{aligned}\phi_X(\theta) &= E[E[e^{i\theta X}|N]] = \sum_{n=0}^{\infty} E[\exp i\theta(\xi_1+\cdots+\xi_n)|N=n]P\{N=n\} = \sum_{n=0}^{\infty}\phi_\gamma(\theta)^n\frac{\lambda^n}{n!}e^{-\lambda}\\ &= \exp\lambda(\phi_\gamma(\theta)-1) = \exp(i\theta\lambda\mu_\gamma - \lambda\int(e^{i\theta x}-1-i\theta x)\ \gamma(dx).\end{aligned}$$

where $\mu_\gamma = \int x\ \gamma(dx)$. This gives the canonical form for the characteristic function with Leévy measure $\mu(dx) = \lambda x^2\gamma(dx)$. Note that by the conditional variance formula and Wald's identities:

$$\begin{aligned}Var(X) &= E[\,Var(X|N)] + Var(E[X|N]) = EN\sigma_\gamma^2 + Var(N\mu_\gamma)\\ &= \lambda(\sigma_\gamma^2+\mu_\gamma^2) = \lambda\int x^2\ \gamma(dx) = \mu(\mathbb{R}).\end{aligned}$$

4. *For $j=1,\ldots,J$, let ϕ_j be the characteristic function for the canonical form for an infinitely divisible distribution with Lévy measure μ_j and mean b_j. Then $\phi_1(\theta)\cdots\phi_J(\theta)$ is the characteristic function for an infinitely divisible random variable whose canonical representation has*

$$\text{mean } b = \sum_{j=1}^{J} b_j, \quad \text{and} \quad \text{Lévy measure } \mu = \sum_{j=1}^{J}\mu_j.$$

Exercise 7.11. 1. *Show that the Lévy measure for $Exp(1)$ has density xe^{-x} with respect to Lebesgue measure.*

2. *Show that the Lévy measure for $\Gamma(\alpha,1)$ has density $e^{-x}x^{\alpha+1}/(\Gamma(\alpha))$ with respect to Lebesgue measure.*

3. *Show that the uniform distribution is not infinitely divisible.*

Now we are in a position to show that the representation above is the characteristic function of an infinitely divisible distribution.

Proof. (Lévy-Khinchin). Define the discrete measures

$$\mu_n\{\frac{j}{2^n}\} = \mu(\frac{j}{2^n},\frac{j+1}{2^n}] \text{ for } j=-2^{2n},-2^{2n}+1,\ldots,-1,0,1,\ldots,2^{2n}-1,2^{2n},$$

i.e.,

$$\mu_n = \sum_{j=-2^n}^{2^n}\mu(\frac{j}{2^n},\frac{j+1}{2^n}]\delta_{j/2^n}.$$

We have shown that a point mass Lévy measure gives either a normal random variable or a linear transformation of a Poisson random variable. Thus, by the example above, μ, as the sum of point masses, is the Lévy measure of an infinitely divisible distribution whose characteristic function has the canonical form.

Write $\tilde{\phi}_n$ for the corresponding characteristic function. Note that $\mu_n(\mathbb{R}) \le \mu(\mathbb{R})$. Moreover, by the theory of Riemann-Stieltjes integrals, $\mu_n \to^v \mu$ and consequently,

$$\lim_{n\to\infty}\tilde{\phi}_n(\theta) = \phi(\theta).$$

Thus, by the continuity theorem, the limit is a characteristic function. Now, write ϕ_n to be the characteristic function with Lévy measure μ replaced by μ/n. Then ϕ_n is a characteristics function and $\phi(\theta) = \phi_n(\theta)^n$ and thus ϕ is the characteristic function of an infinitely divisible distribution. □

Let's rewrite the characteristic function as

$$\phi(\theta) = \exp\left(ib\theta - \frac{1}{2}\sigma^2\theta^2 + \lambda \int_{\mathbb{R}\setminus\{0\}} (e^{i\theta x} - 1 - i\theta x)\gamma(dx) \right).$$

where

1. $\sigma^2 = \mu\{0\}$
2. $\lambda = \int_{\mathbb{R}\setminus\{0\}} x^{-2}\mu(dx)$, and
3. $\gamma(A) = \int_{A\setminus\{0\}} x^{-2}\ \mu(dx)/\lambda$.

Thus, we can represent an infinitely divisible random variable X having finite mean and variance as

$$X = b - \lambda\mu_\gamma + \sigma Z + \sum_{n=1}^{N} \xi_n$$

where

1. $b \in \mathbb{R}$,
2. $\sigma \in [0, \infty)$,
3. Z is a standard normal random variable,
4. N is a Poisson random variable, parameter λ,
5. $\{\xi_n; n \geq 1\}$ are independent mean μ_γ random variables with distribution γ, and
6. Z, N, and $\{\xi_n; n \geq 1\}$ are independent.

The following theorem is proves the converse of the theorem above and, at the same time, will help identify the limiting distribution.

Theorem 7.12. *Let ν be the limit law for S_n, the sums of the rows of the triangular array described in (7.3). Then ν has one of the characteristic functions of the infinitely divisible distributions characterized by the Lévy-Khinchin formula (7.4).*

Proof. Let $\phi_{n,k}$ denote the characteristics function of $X_{n,k}$. By considering $X_{n,k} - \mu_{n,k}$, we can assume the random variables in the triangular array have mean 0.

Claim.

$$\lim_{n\to\infty} \left(\prod_{k=1}^{k_n} \phi_{n,k}(\theta) - \exp \sum_{k=1}^{k_n} (\phi_{n,k}(\theta) - 1) \right) = 0.$$

Use the first claim (7.1) in the proof of the classical central limit theorem with $z_k = \phi_{n,k}(\theta)$ and $w_k = \exp(\phi_{n,k}(\theta) - 1)$ and note that each of the z_k and w_k have modulus at most 1. Therefore the absolute value of the terms in the limit above is bound above by

$$\sum_{k=1}^{k_n} |\phi_{n,k}(\theta) - \exp(\phi_{n,k}(\theta) - 1)|.$$

Next, use the exercise (with $w = \phi_{n,k}(\theta) - 1, |w| \leq 2$), the second claim (7.2) in that proof (with $w = \leq^{i\theta x}$) and the fact that $X_{n,k}$ has mean zero to obtain

$$|\phi_{n,k}(\theta) - \exp(\phi_{n,k}(\theta) - 1)| \leq 2|\phi_{n,k}(\theta) - 1|^2 = 2|E[e^{i\theta X_{n,k}} - 1 - i\theta X_{n,k}]| \leq 2(\theta^2 \sigma_{n,k}^2)^2.$$

Thus, the sum above is bound above by a constant times

$$\sum_{k=1}^{k_n} \sigma_{n,k}^4 \leq \left(\sup_{1 \leq k \leq k_n} \sigma_{n,k}^2 \right) \sum_{k=1}^{k_n} \sigma_{n,k}^2 \leq \left(\sup_{1 \leq k \leq k_n} \sigma_{n,k}^2 \right) \left(\sup_{n \geq 1} \sigma_n^2 \right)$$

and this tends to zero as $n \to \infty$ and the claim is established.

Let $\nu_{n,k}$ denote the distribution of $X_{n,k}$, then set

$$\sum_{k=1}^{k_n} (\phi_{n,k}(\theta) - 1) = \sum_{k=1}^{k_n} \int (e^{i\theta x} - 1 - i\theta x) \ \nu_{n,k}(dx) = \int (e^{i\theta x} - 1 - i\theta x) \frac{1}{x^2} \ \mu_n(dx)$$

where μ_n is the measure defined by

$$\mu_n(A) = \sum_{k=1}^{k_n} \int_A x^2 \ \nu_{n,k}(dx).$$

Now set,

$$\phi_n(\theta) = \exp\left(\int (e^{i\theta x} - 1 - i\theta x) \frac{1}{x^2} \ \mu_n(dx) \right).$$

Then, the limit in the claim can be written

$$\lim_{n \to \infty} \phi_{S_n}(\theta) - \phi_n(\theta) = 0.$$

Because $\sup_n \mu_n(\mathbb{R}) = \sup_n \sigma_n^2 < \infty$, some subsequence $\{\mu_{n_j}; j \geq 1\}$ converges vaguely to a finite measure μ and

$$\lim_{j \to \infty} \phi_{n_j}(\theta) = \exp\left(\int (e^{i\theta x} - 1 - i\theta x) \frac{1}{x^2} \ \mu(dx) \right).$$

However,

$$\lim_{n \to \infty} \phi_{S_n}(\theta) \text{ exists.}$$

and the characteristic function has the canonical form given above. □

Thus, the vague convergence of μ_n is sufficient for weak convergences of S_n. We now prove that it is necessary.

Theorem 7.13. *Let S_n be the row sums of mean zero bounded variance triangular arrays. Then the distribution of S_n converges to infinitely divisible distribution with Lévy measure μ if and only if $\mu_n \to^v \mu$ where*

$$\mu_n(A) = \sum_{k=1}^{k_n} \int_A x^2 \ \nu_{n,k}(dx) = \sum_{k=1}^{k_n} E[X_{n,k}^2; \{X_{n,k} \in A\}]$$

and $\nu_{n,k}(A) = P\{X_{n,k} \in A\}$.

Proof. All that remains to the shown is the necessity of the vague convergence. Now suppose that

$$\lim_{n\to\infty} \phi_n(\theta) = \phi(\theta)$$

where ϕ_n is the characteristic function of an infinitely divisible distribution with Lévy measure μ_n. Because $\sup_n \mu_n(\mathbb{R}) < \infty$, every subsequence $\{\mu_{n_j}; j \geq 1\}$ contains a further subsequence $\{\mu_{n_{j(\ell)}}; \ell \geq 1\}$ that converges vaguely to some $\tilde{\mu}$. Set

$$\tilde{\phi}(\theta) = \exp\left(\int (e^{i\theta x} - 1 - i\theta x)\frac{1}{x^2} \ \tilde{\mu}(dx)\right).$$

Because $\phi = \tilde{\phi}$, we have that $\phi' = \tilde{\phi}'$ or

$$i\phi(\theta)\int (e^{i\theta x} - 1)\frac{1}{x} \ \mu(dx) = i\tilde{\phi}(\theta)\int (e^{i\theta x} - 1)\frac{1}{x} \ \tilde{\mu}(dx)$$

Use the fact that ϕ and $\tilde{\phi}$ are never 0 to see that

$$\int (e^{i\theta x} - 1)\frac{1}{x} \ \mu(dx) = \int (e^{i\theta x} - 1)\frac{1}{x} \ \tilde{\mu}(dx).$$

Differentiate again with respect to θ to obtain

$$\int ie^{i\theta x} \ \mu(dx) = \int ie^{i\theta x}\tilde{\mu}(dx).$$

Thus $\sigma^2 = \mu(\mathbb{R}) = \tilde{\mu}(\mathbb{R})$. Now, divide the equation above by σ^2/i and use the fact that characteristic functions uniquely determine the probability measure to show that $\mu = \tilde{\mu}$. □

7.4 Applications of the Lévy-Khinchin Formula

Example 7.14. *Let N_λ be Pois(λ), then*

$$Z_\lambda = \frac{N_\lambda - \lambda}{\sqrt{\lambda}}$$

has mean zero and variance one and is infinitely divisible with Lévy measure $\delta_{1/\lambda}$. Because

$$\delta_{1/\lambda} \to^v \delta_0 \quad as \quad \lambda \to \infty,$$

we see that

$$Z_\lambda \Rightarrow Z,$$

a standard normal random variable.

We can use the theorem to give necessary and sufficient conditions for a triangular array to converge to a normal random variable.

Theorem 7.15 (Lindeberg-Feller)**.** *For the triangular array above,*

$$\frac{S_n}{\sigma_n} \to^{\mathcal{D}} Z,$$

a standard normal random variable if and only if for every $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{k=1}^{k_n} E[X_{n,k}^2; \{|X_{n,k}| \geq \epsilon\sigma_n\}] = 0.$$

Proof. Define

$$\mu_n(A) = \frac{1}{\sigma_n^2} \sum_{k=1}^{k_n} E[X_{n,k}^2; \{X_{n,k} \in A\}]$$

Then the theorem holds if and only if $\mu_n \to^v \delta_0$. Each μ_n has total mass 1. Thus, it suffices to show that for every $\epsilon > 0$

$$\lim_{n\to\infty} \mu_n([-\epsilon, \epsilon]^c) = 0$$

This is exactly the condition above. □

The sufficiency of this condition is due to Lindeberg and is typically called the *Lindeberg condition*. The necessity of the condition is due to Feller.

Exercise 7.16. *Show that the classical central limit theorem follows from the Lindeberg-Feller central limit theorem.*

Example 7.17. *Consider the sample space* Ω *that conists of the* $n!$ *permutations of the integers* $\{1, \ldots, n\}$ *and define a probability that assigns* $1/n!$ *to each of the outcomes in* Ω*.*

Define $Y_{n,j}(\omega)$ *to be the number of inversions caused by* j *in a given permutation* ω*. In other words,* $Y_{n,j}(\omega) = k$ *if and only if* j *precedes exactly* k *of the integers* $1, \ldots, j-1$ *in* ω*.*

Claim. *For each* n*,* $\{Y_{n,j}; 1 \leq j \leq n\}$ *are independent and satisfy*

$$P\{Y_{n,j} = k\} = \frac{1}{j}, \text{ for } 0 \leq k \leq j-1.$$

Note that the values of $Y_{n,1}, \ldots, Y_{n,j}$ *are determined as soon as the positions of the integers* $1, \ldots, j$ *are known. Given any* j *designated positions among the* n *ordered slots, the number of permutations in which* $1, \ldots, j$ *occupy these positions in some order is* $j!(n-j)!$*. Among these pemutations, the number in which* j *occupies the* k*-th position is* $(j-1)!(n-j)!$*. The remaining values* $1, \ldots, j-1$ *can occupy the remaining positions in* $(j-1)!$ *distinct ways. Each of these choice corresponds uniquely to a possible value of the random vector*

$$(Y_{n,1}, \ldots, Y_{n,j-1}).$$

On the other hand, the number of possible values is $1 \times 2 \times \cdots \times (j-1) = (j-1)!$ *and the mapping between permutations and the possible values of the* j*-tuple above is a one-to-one correspondence.*

In summary, for any distinct value $(i_1, \ldots, i_{j-1})$*, the number of permutations* ω *in which*

1. $1, \ldots, j$ *occupy the given positions, and*

2. $Y_{n,1}(\omega) = i_1, \ldots, Y_{n,j-1}(\omega) = i_{j-1}, Y_{n,j}(\omega) = k$

is equal to $(n-j)!$*. Hence the number of permutations satisfying the second condition alone is equal to*

$$\binom{n}{j}(n-j)! = \frac{n!}{j!}.$$

Sum this over the values of $k = 0, \ldots, j-1$*, we obtain that the number of permutations satisfying*

$$Y_{n,1}(\omega) = i_1, \ldots, Y_{n,j-1}(\omega) = i_{j-1}$$

is

$$j\frac{n!}{j!} = \frac{n!}{(j-1)!}.$$

Therefore,

$$P\{\omega; Y_{n,j}(\omega) = k | Y_{n,1}(\omega) = i_1, \ldots, Y_{n,j-1}(\omega) = i_{j-1}\} = \frac{\frac{n!}{j!}}{\frac{n!}{(j-1)!}} = \frac{1}{j},$$

proving the claim.

This gives

$$EY_{n,j} = \frac{j-1}{2}, \quad Var(Y_{n,j}) = \frac{j^2-1}{12},$$

and letting T_n *denote the sum of the* n*-th row,*

$$ET_n \approx \frac{n^2}{2}, \quad Var(T_n) \approx \frac{n^3}{36}.$$

Note that for any $\epsilon > 0$*, we have for sufficiently large* n

$$|Y_{n,j} - EY_{n,j}| \leq j - 1 \leq n - 1 \leq \epsilon\sqrt{Var(T_n)}$$

Set

$$X_{n,k} = \frac{Y_{n,j} - EY_{n,j}}{\sqrt{Var(T_n)}}.$$

Then $\sigma_n^2 = 1$ *and the Lindeberg condition applies. Thus*

$$\frac{T_n - \frac{n^2}{4}}{\frac{n^{3/2}}{6}} \rightarrow^{\mathcal{D}} Z,$$

a standard normal.

A typical sufficient condition for the central limit theorem is the Lyapounov condition given below.

Theorem 7.18 (Lyapounov)**.** *For the triangular array above, suppose that*

$$\lim_{n\to\infty} \frac{1}{\sigma_n^{2+\delta}} \sum_{k=1}^{k_n} E[|X_{n,k}|^{2+\delta}] = 0.$$

Then

$$\frac{S_n}{\sigma_n} \to^{\mathcal{D}} Z,$$

a standard normal random variable.

Proof. We show that the Lyapounov condition implies the Lindeberg condition by showing that a fixed multiple of each term in the Lyapounov condition is larger than the corresponding term in the Lindeberg condition.

$$\frac{1}{\sigma_n^2} E[X_{n,k}^2; \{|X_{n,k}| \geq \epsilon\sigma_n\}] \leq \frac{1}{\sigma_n^2} E[X_{n,k}^2 \left(\frac{|X_{n,k}|}{\epsilon\sigma_n}\right)^{\delta}; \{|X_{n,k}| \geq \epsilon\sigma_n\}] \leq \frac{1}{\epsilon^{\delta}\sigma_n^{2+\delta}} E[|X_{n,k}|^{2+\delta}].$$

□

Example 7.19. *Let $\{X_k; k \geq 1\}$ be independent random variables, X_k is $Ber(p_k)$. Assume that*

$$a_n^2 = \sum_{k=1}^{n} Var(X_k) = \sum_{k=1}^{n} p_k(1-p_k)$$

has an infinite limit. Consider the triangular array with $X_{n,k} = (X_k - p_k)/a_n$ and write $S_n = X_{n,1} + \cdots + X_{n,n}$. We check Lyapounov's condition with $\delta = 1$.

$$E|X_k - p_k|^3 = (1-p_k)^3 p_k + p_k^3(1-p_k) = p_k(1-p_k)((1-p_k)^2 + p_k^2) \leq 2p_k(1-p_k).$$

Then, $\sigma_n^2 = 1$ for all n and

$$\frac{1}{\sigma_n^3} \sum_{k=1}^{n} E[|X_{n,k}|^3] \leq \frac{2}{a_n^3} \sum_{k=1}^{n} p_k(1-p_k) \leq \frac{2}{a_n}.$$

We can also use the Lévy-Khinchin theorem to give necessary and sufficient conditions for a triangular array to converge to a Poisson random variable. We shall use this in the following example.

Example 7.20. *For each n, let $\{Y_{n,k}; 1 \leq k_n\}$ be independent $Ber(p_{n,k})$ random variables and assume that*

$$\lim_{n\to\infty} \sum_{k=1}^{k_n} p_{n,k} = \lambda$$

and

$$\lim_{n\to\infty} \sup_{1\leq k\leq k_n} p_{n,k} = 0.$$

Note that

$$|\sigma_n^2 - \lambda| \leq |\sum_{k=1}^{k_n} p_{n,k}(1-p_{n,k}) - \sum_{k=1}^{k_n} p_{n,k}| + |\sum_{k=1}^{k_n} p_{n,k} - \lambda|.$$

Now the first term is equal to

$$\sum_{k=1}^{k_n} p_{n,k}^2 \leq \left(\sup_{1 \leq k \leq k_n} p_{n,k} \right) \sum_{k=1}^{k_n} p_{n,k} \tag{7.5}$$

which has limit zero.

The second term has limit zero by hypothesis. Thus,

$$\lim_{n \to \infty} \sigma_n^2 = \lambda$$

Set

$$S_n = \sum_{k=1}^{k_n} Y_{n,k}.$$

Then

$$S_n \to^{\mathcal{D}} N,$$

a Pois(λ)-random variable if and only if the measures

$$\mu_n(A) = \sum_{k=1}^{k_n} E[(Y_{n,k} - p_{n,k})^2; \{(Y_{n,k} - p_{n,k}) \in A\}]$$

converges vaguely to $\lambda \delta_1$.

We have that

$$\lim_{n \to \infty} \mu_n(\mathbb{R}) = \lim_{n \to \infty} \sigma_n^2 = \lambda.$$

Thus, all that is left to show is that

$$\lim_{n \to \infty} \mu_n([1 - \epsilon, 1 + \epsilon]^c) = 0.$$

So, given $\epsilon > 0$, choose N so that $\sup_{1 \leq k \leq k_n} p_{n,k} < \epsilon$ for all $n > N$. Then

$$\{|Y_{n,k} - p_{n,k} - 1| > \epsilon\} = \{Y_{n,k} = 0\}$$

Thus,

$$\mu_n([1 - \epsilon, 1 + \epsilon]^c) = \sum_{k=1}^{k_n} E[(Y_{n,k} - p_{n,k})^2; \{Y_{n,k} = 0\}] = \sum_{k=1}^{k_n} E[p_{n,k}^2; \{Y_{n,k} = 0\}] \leq \sum_{k=1}^{k_n} p_{n,k}^2.$$

We have previously shown in (7.5) that the limit as $n \to \infty$ and the desired vague convergence holds.