University of Regina

Statistics 851 – Probability

Lecture Notes

Winter 2008

Michael Kozdron

kozdron@stat.math.uregina.ca
http://stat.math.uregina.ca/∼kozdron

References

[1] Jean Jacod and Philip Protter. *Probability Essentials*, second edition. Springer, Heidelberg, Germany, 2004.

[2] Jeffrey S. Rosenthal. *A First Look at Rigorous Probability Theory.* World Scientific, Singapore, 2000.

Preface

Statistics 851 is the first graduate course in measure-theoretic probability at the University of Regina. There are no formal prerequisites other than graduate student status, although it is expected that students encountered basic probability concepts such as discrete and continuous random variables at some point in their undergraduate careers. It is also expected that students are familiar with some basic ideas of set theory including proof by induction and countability.

The primary textbook for the course is [1] and references to chapter, theorem, and exercise numbers are for that book. Lectures #1 and #9, however, are based on [2] which is one of the supplemental course books.

Michael Kozdron
Regina, SK
January 2008

List of Lectures and Handouts

Statistics 851 (Winter 2008) January 7, 2008
Prof. Michael Kozdron

Lecture #1: The Need for Measure Theory

Reference. Today's notes and §1.1 of [2]

In order to study probability theory in a mathematically rigorous way, it is necessary to learn the appropriate language which is "measure theory."

Theoretical probability is used in a variety of diverse fields such as

- statistics,
- economics,
- management,
- finance,
- computer science,
- engineering,
- operations research.

More will be said about the applications as the course progresses.

For now, however, we will begin with some "easy" examples from undergraduate probability (STAT 251/STAT 351).

Example. Let X be a Poisson(5) random variable. What does this mean?

Solution. It means that X takes on a "random" non-negative integer k ($k \geq 0$) according to the *probability mass fuction*

$$f_X(k) := P\{X = k\} = \frac{e^{-5}5^k}{k!}, \quad k \geq 0.$$

We can then compute things like the expected value of X^2, i.e.,

$$\mathbb{E}(X^2) = \sum_{k=0}^{\infty} k^2 P\{X = k\} = \sum_{k=0}^{\infty} \frac{k^2 e^{-5}5^k}{k!}.$$

Note that X is an example of a *discrete random variable*.

Example. Let Y be a Normal$(0,1)$ random variable. What does this mean?

Solution. It means that the probability that Y lies between two real numbers a and b (with $a \leq b$) is given by

$$P\{a \leq Y \leq b\} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, \mathrm{d}y.$$

Of course, for any real number y, we have $P\{Y = y\} = 0$. (How come?) We say that the *probability density function* for Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \quad -\infty < y < \infty.$$

We can then compute things like the expected value of Y^2, i.e.,

$$\mathbb{E}(Y^2) = \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, \mathrm{d}y.$$

Note that Y is an example of a *(absolutely) continuous random variable.*

Example. Introduce a new random variable Z as follows. Flip a fair coin (independently), and set $Z = X$ if it comes up heads and set $Z = Y$ if it comes up tails. That is,

$$P\{Z = X\} = P\{Z = Y\} = 1/2.$$

What kind of random variable is Z?

Solution. We can formally define Z to be the random variable

$$Z = XW + Y(1 - W)$$

where the random variable W satisfies

$$P\{W = 1\} = P\{W = 0\} = 1/2$$

and is independent of X and Y. Clearly Z is not discrete since it can take on uncountably many values, and it is not absolutely continuous since $P\{Z = z\} > 0$ for certain values of z (namely when z is a non-negative integer).

Question. How do we study Z? How could you compute things like $\mathbb{E}(Z^2)$? (Hint: If you tried to compute $\mathbb{E}(Z^2)$ using undergraduate probability then you would be conditioning on an event of probability 0. That is not allowed.)

Answer. The distinction between discrete and absolutely continuous is artificial. Measure theory gives a common definition to expected value which applies equally well to discrete random variables (like X), absolutely continuous random variables (like Y), combinations (like Z), and ones not yet imagined.

Lecture #2: Introduction

Reference. Chapter 1 pages 1–5

Intuitively, a *probability* is a measure of the likelihood of an *event*.

Example. There is a 60% chance of snow tomorrow:

$$P\{\text{snow tomorrow}\} = 0.60.$$

Example. There is a probability of 1/2 that this coin will land heads:

$$P\{\text{heads}\} = 0.50.$$

The method of assigning probabilities to events is a philosophical question.

- What does it mean for there to be a 60% chance of snow tomorrow?
- What does it mean for a fair coin to have a 50% chance of landing heads?

There are roughly two paradigms for assigning probabilities. One is the *frequentist approach* and the other is the *Bayesian approach*.

The frequentist approach is based on repeated experimentation and long-run averages. If we flip a coin many, many times, we expect to see a head about half the time. Therefore, we conclude that the probability of heads on a single toss is 0.50.

In contrast, we cannot repeat "tomorrow." There will only be one January 10, 2008. Either it will snow or it will not. The weather forecaster can predict snow with probability 0.60 based on weather patterns, historical data, and intuition. This is a Bayesian, or subjectivist, approach to assigning probabilities.

Instead of entering into a philosophical discussion, we will assume that we have a reasonable way of assigning probabilities to events—either frequentist, Bayesian, or a combination of the two.

Question. What are some properties that a probability must have?

Answer. Assign a number between 0 and 1 to each event.

- The impossible event will have probability 0.
- The certain event will have probability 1.
- All events will have probability between 0 and 1. That is, if A is an event, then $0 \leq P\{A\} \leq 1$.

However, an event is not the most basic outcome of an experiment. An event may be a collection of outcomes.

Example. Roll a fair die. Let A be the event that an even number is rolled; that is, $A = \{\text{even number}\} = \{2 \text{ or } 4 \text{ or } 6\}$. Intuitively, $P\{A\} = 3/6$.

Notation. Let Ω denote the set of all possible outcomes of an experiment. We call Ω the *sample space* which is composed of all individual *outcomes*. These are labelled by ω so that $\Omega = \{\omega \in \Omega\}$.

Example. Model the experiment of rolling a fair die.

Solution. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The individual outcomes $\omega_i = i$, $i = 1, 2, \ldots, 6$, are all equally likely so that

$$P\{\omega_i\} = \frac{1}{6}.$$

Definition. An *event* is a collection of outcomes. Formally, an event A is a subset of the sample space; that is, $A \subseteq \Omega$.

Technical Matter. It will not be as easy as declaring all events as simply all possibly subsets of Ω. (The set of all subsets of Ω is called the *power set* of Ω and denoted by 2^Ω.) If Ω is uncountable, then the power set will be too big!

Example. Model the experiment of flipping a fair coin. Include a list of all possible events.

Solution. The sample space is $\Omega = \{H, T\}$. The list of all possible events is

$$\emptyset, \Omega, \{H\}, \{T\}.$$

Note that if we write $\mathcal{A}$ to denote the set of all possible events, then

$$\mathcal{A} = \{\emptyset, \Omega, \{H\}, \{T\}\}$$

satisfies $|\mathcal{A}| = 2^{|\Omega|} = 2^2 = 4$. Since the individual outcomes are each equally likely we can assign probabilities to events as

$$P\{\emptyset\} = 0, \quad P\{\Omega\} = 1, \quad P\{H\} = P\{T\} = \frac{1}{2}.$$

Instead of worrying about this technical matter right now, we will first study countable spaces. The collection $\mathcal{A}$ in the previous example has a special name.

Definition. A collection $\mathcal{A}$ of subsets $A_i \subseteq \Omega$ is called a *σ-algebra* if

(i) $\emptyset \in \mathcal{A}$,

(ii) $\Omega \in \mathcal{A}$,

(iii) $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$, and

(iv) $A_1, A_2, A_3, \ldots \in \mathcal{A}$ implies

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}.$$

Remark. Condition (iv) is called *countable additivity*.

Example. $\mathcal{A} = \{\emptyset, \Omega\}$ is a σ-algebra (called the *trivial σ-algebra*).

Example. For any subset $A \subset \Omega$,

$$\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$$

is a σ-algebra.

Example. If Ω is any space (whether countable or uncountable), then the power set of Ω is a σ-algebra.

Proposition. *Suppose that $\mathcal{A}$ is a σ-algebra. If $A_1, A_2, A_3, \ldots \in \mathcal{A}$, then*

(v) $A_1 \cup A_2 \cup \cdots \cup A_n \in \mathcal{A}$ *for every* $n < \infty$, *and*

(vi) $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$.

Remark. Condition (v) is called *finite additivity*.

Proof. To show that finite additivity (v) follows from countable additivity simply take A_i, $i > n$, to be $\emptyset$. That is,

$$A_1 \cup A_2 \cup \cdots \cup A_n \cup \emptyset \cup \emptyset \cup \cdots = \bigcup_{i=1}^{n} A_i \in \mathcal{A}$$

since $A_1, A_2, \ldots, A_n, \emptyset \in \mathcal{A}$. Condition (vi) follows from De Morgan's law (Exercise 2.3). Suppose that $A_1, A_2, \ldots \in \mathcal{A}$ so that $A_1^c, A_2^c, \ldots \in \mathcal{A}$ by (iii). By countably additivity, we have

$$\bigcup_{i=1}^{\infty} A_i^c \in \mathcal{A}.$$

Thus, by (iii) again we have

$$\left[\bigcup_{i=1}^{\infty} A_i^c\right]^c \in \mathcal{A}.$$

However,

$$\left[\bigcup_{i=1}^{\infty} A_i^c\right]^c = \bigcap_{i=1}^{\infty} [A_i^c]^c = \bigcap_{i=1}^{\infty} A_i$$

which establishes (vi). □

Definition. A collection $\mathcal{A}$ of events satisfying (i), (ii), (iii), (v) (but not (iv)) is called an *algebra*.

Remark. Some people say *σ-field/field* instead of σ-algebra/algebra.

Proposition. *Suppose that $\mathcal{A}$ is a σ-algebra. If $A_1, A_2, A_3, \ldots \in \mathcal{A}$, then*

(vii) $A_1 \cap A_2 \cap \cdots \cap A_n \in \mathcal{A}$ *for every* $n < \infty$.

Proof. To show that this result follows from (vi) simply simply take A_i, $i > n$, to be Ω. That is,

$$A_1 \cap A_2 \cap \cdots \cap A_n \cap \Omega \cap \Omega \cap \cdots = \bigcap_{i=1}^{n} A_i \in \mathcal{A}$$

since $A_1, A_2, \ldots, A_n, \Omega \in \mathcal{A}$. □

Definition. A *probability (measure)* is a set function $P : \mathcal{A} \to [0, 1]$ satisfying

(i) $P\{\emptyset\} = 0$, $P\{\Omega\} = 1$, and $0 \leq P\{A\} \leq 1$ for every $A \in \mathcal{A}$, and

(ii) if $A_1, A_2, \ldots \in \mathcal{A}$ are disjoint, then

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\}.$$

Definition. A *probability space* is a triple $(\Omega, \mathcal{A}, P)$ where Ω is a sample space, $\mathcal{A}$ is a σ-algebra of subsets of Ω, and P is a probability measure.

Example. Consider the experiment of tossing a fair coin twice. In this case,

$$\Omega = \{HH, HT, TH, TT\}$$

and $\mathcal{A}$ consists of the $|\mathcal{A}| = 2^{|\Omega|} = 2^4 = 16$ elements

$$\mathcal{A} = \Big\{\emptyset, \Omega, HH, HT, TH, TT, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\},$$
$$\{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\}\Big\}.$$

Example. Consider the experiment of rolling a fair die so that $\Omega = \{1, 2, 3, 4, 5, 6\}$. In this case, $\mathcal{A}$ consists of the $2^6 = 64$ events

$$\mathcal{A} = \{\emptyset, \Omega, 1, 2, 3, 4, 5, 6, 12, 13, 14, 15, 16, 21, \text{etc.}\}.$$

We can then define

$$P\{A\} = \frac{|A|}{6} \quad \text{for every } A \in \mathcal{A}. \qquad (*)$$

For instance, if A is the event $\{\text{roll even number}\} = \{2, 4, 6\}$ and if B is the event $\{\text{roll odd number less than } 4\} = \{1, 3\}$, then

$$P\{A\} = \frac{3}{6} \quad \text{and} \quad P\{B\} = \frac{2}{6}.$$

Furthermore, since A and B are disjoint (that is, $A \cap B = \emptyset$) we can use finite additivity to conclude

$$P\{A \cup B\} = P\{A\} + P\{B\} = \frac{5}{6}. \qquad (**)$$

Since $A \cup B = \{1, 2, 3, 4, 6\} \in \mathcal{A}$ has cardinality $|A \cup B| = 5$, we see that from $(*)$ that

$$P\{A \cup B\} = \frac{|A \cup B|}{6} = \frac{5}{6}$$

which is consistent with $(**)$.

Lecture #3: Axioms of Probability

Reference. Chapter 2 pages 7–11

Example. Construct a probability space to model the experiment of tossing a fair coin twice.

Solution. We must carefully define $(\Omega, \mathcal{A}, P)$. Thus, we take our sample space to be

$$\Omega = \{HH, HT, TH, TT\}$$

and our σ-algebra $\mathcal{A}$ consists of the $|\mathcal{A}| = 2^{|\Omega|} = 2^4 = 16$ elements

$$\mathcal{A} = \Big\{\emptyset, \Omega, HH, HT, TH, TT, \{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\},$$
$$\{TH, TT\}, \{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\}\Big\}.$$

In general, if Ω is finite, then the power set $\mathcal{A} = 2^\Omega$ (i.e., the set of all subsets of Ω) is a σ-algebra with $|\mathcal{A}| = 2^{|\Omega|}$. We take as our probability a set function $P : \mathcal{A} \to [0, 1]$ satisfying certain conditions. Thus,

$$P\{\emptyset\} = 0, \quad P\{\Omega\} = 1,$$

$$P\{HH\} = P\{TT\} = P\{HT\} = P\{TH\} = \frac{1}{4},$$

$$P\{HH, HT\} = P\{HH, TH\} = P\{HH, TT\} = P\{HT, TH\} = P\{HT, TT\} = P\{TH, TT\} = \frac{1}{2},$$

$$P\{HH, HT, TH\} = P\{HH, HT, TT\} = P\{HH, TH, TT\} = P\{HT, TH, TT\} = \frac{3}{4}.$$

That is, if $A \in \mathcal{A}$, let

$$P\{A\} = \frac{|A|}{4}.$$

Random Variables

Definition. A function $X : \Omega \to \mathbb{R}$ given by $\omega \mapsto X(\omega)$ is called a *random variable*.

Remark. A random variable is NOT a variable in the algebraic sense. It is a function in the calculus sense.

Remark. When Ω is more complicated than just a finite sample space we will need to be more careful with this definition. For now, however, it will be fine.

Example. Consider the experiment of tossing a fair coin twice. Let the random variable X denote the number of heads. In order to define the function $X : \Omega \to \mathbb{R}$ we must define $X(\omega)$ for every $\omega \in \Omega$. Thus,

$$X(HH) = 2, \quad X(HT) = 1, \quad X(TH) = 1, \quad X(TT) = 0.$$

Note. Every random variable X on a probability space $(\Omega, \mathcal{A}, P)$ induces a probability measure on $\mathbb{R}$ which is denoted P^X and called the *law* (or *distribution*) of X. It is defined for every $B \in \mathcal{B}$ by

$$P^X(B) := P\{\omega \in \Omega : X(\omega) \in B\} = P\{X \in B\} = P\{X^{-1}(B)\}.$$

In other words, the random variable X transforms the probability space $(\Omega, \mathcal{A}, P)$ into the probability space $(\mathbb{R}, \mathcal{B}, P^X)$:

$$(\Omega, \mathcal{A}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}, P^X).$$

The σ-algebra $\mathcal{B}$ is called the *Borel* σ-algebra on $\mathbb{R}$. We will discuss this in detail shortly.

Example. Toss a coin twice and let X denote the number of heads observed. The law of X is defined by

$$P^X(B) = \frac{\#(i,j) \text{ such that } 1\{i = H\} + 1\{j = H\} \in B}{4}$$

where 1 denotes the indicator function (defined below). For instance,

$$P^X(\{2\}) = P\{\omega \in \Omega : X(\omega) = 2\} = P\{X = 2\} = P\{HH\} = \frac{1}{4},$$

$$P^X(\{1\}) = P\{X = 1\} = P\{HT, TH\} = \frac{1}{2},$$

$$P^X(\{0\}) = P\{X = 0\} = P\{TT\} = \frac{1}{4},$$

and

$$\begin{aligned} P^X\left(\left[-\frac{1}{2}, \frac{3}{2}\right]\right) = P\left\{\omega \in \Omega : X(\omega) \in \left[-\frac{1}{2}, \frac{3}{2}\right]\right\} &= P\left\{-\frac{1}{2} \le X \le \frac{3}{2}\right\} \\ &= P\{X = 0 \text{ or } X = 1\} \\ &= P\{TT, HT, TH\} = \frac{3}{4}. \end{aligned}$$

Later we will see how P^X will be related to F_X, the *distribution function* of X.

Notation. The *indicator function* of the event A is defined by

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

Note. On page 5, it should read: "(for example, $P^X(\{2\}) = P(\{1,1\}) = \frac{1}{36} \ldots$"

Definition. If $\mathcal{C} \subseteq 2^\Omega$, then the *$\sigma$-algebra generated by $\mathcal{C}$* is the smallest σ-algebra containing $\mathcal{C}$. It is denoted by $\sigma(\mathcal{C})$.

Note. By Exercise 2.1, the power set 2^Ω is a σ-algebra. By Exercise 2.2, the intersection of σ-algebras is itself a σ-algebra. Therefore, $\sigma(\mathcal{C})$ always exists and

$$\mathcal{C} \subseteq \sigma(\mathcal{C}) \subseteq 2^\Omega.$$

Definition. The *Borel σ-algebra* on $\mathbb{R}$, written $\mathcal{B}$, is the σ-algebra generated by the open sets. (Equivalently, $\mathcal{B}$ is generated by the closed sets.) That is, if $\mathcal{C}$ denotes the collection of all open sets in $\mathbb{R}$, then $\mathcal{B} = \sigma(\mathcal{C})$.

In fact, to understand $\mathcal{B}$ it is enough to consider intervals of the form $(-\infty, a]$ as the next theorem shows.

Theorem. *The Borel σ-algebra $\mathcal{B}$ is generated by intervals of the form $(-\infty, a]$ where $a \in \mathbb{Q}$ is a* rational *number.*

Proof. Let $\mathcal{C}$ denote the collection of all open intervals. Since every open set in $\mathbb{R}$ is a countable union of open intervals, we must have $\sigma(\mathcal{C}) = \mathcal{B}$. Let $\mathcal{D}$ denote the collection of all intervals of the form $(-\infty, a]$, $a \in \mathbb{Q}$. Let $(a, b) \in \mathcal{C}$ for some $b > a$ with $b \in \mathbb{Q}$. Let

$$a_n = a + \frac{1}{n}$$

so that $a_n \downarrow a$ as $n \to \infty$, and let

$$b_n = b + \frac{1}{n}$$

so that $b_n \uparrow b$ as $n \to \infty$. Thus,

$$(a, b) = \bigcup_{n=1}^{\infty} (a_n, b_n] = \bigcup_{n=1}^{\infty} \{(-\infty, b_n] \cap (-\infty, a_n]^c\}$$

which implies that $(a, b) \in \sigma(\mathcal{D})$. That is, $\mathcal{C} \subseteq \sigma(\mathcal{D})$ so that $\sigma(\mathcal{C}) \subseteq \sigma(\mathcal{D})$. However, every element of $\mathcal{D}$ is a closed set which implies that

$$\sigma(\mathcal{D}) \subseteq \mathcal{B}.$$

This gives the chain of containments

$$\mathcal{B} = \sigma(\mathcal{C}) \subseteq \sigma(\mathcal{D}) \subseteq \mathcal{B}$$

and so $\sigma(\mathcal{D}) = \mathcal{B}$ proving the theorem. □

Remark. Exercises 2.9, 2.10, 2.11, 2.12 can be proved using Definition 2.3. For instance, since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, we can use use countable additivity (axiom 2) to conclude

$$P\{\Omega\} = P\{A \cup A^c\} = P\{A\} + P\{A^c\}.$$

We now use axiom 1, the fact that $P\{\Omega\} = 1$, to find

$$1 = P\{A\} + P\{A^c\} \quad \text{and so} \quad P\{A^c\} = 1 - P\{A\}.$$

Proposition. *If $A \subseteq B$, then $P\{A\} \leq P\{B\}$.*

Proof. We write

$$B = (A \cap B) \cup (A^c \cap B) = A \cup (A^c \cap B)$$

and note that

$$(A \cap B) \cap (A^c \cap B) = \emptyset.$$

Thus, by countable additivity (axiom 2) we conclude

$$P\{B\} = P\{A\} + P\{A^c \cap B\} \geq P\{A\}$$

using the fact that $0 \leq P\{A\} \leq 1$ for every event A. □

Statistics 851 (Winter 2008) January 14, 2008
Prof. Michael Kozdron

Lecture #4: Definition of Probability Measure

Reference. Chapter 2 pages 7–11

We begin with the axiomatic definition of a probability measure.

Definition. Let Ω be a sample space and let $\mathcal{A}$ be a σ-algebra of subsets of Ω. A set function $P : \mathcal{A} \to [0,1]$ is called a *probability measure* on $\mathcal{A}$ if

(1) $P\{\Omega\} = 1$, and

(2) if $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint, then

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\}.$$

Note that this definition is slightly different than the one given in Lecture #2. The axioms given in the definition are actually the minimal axioms one needs to assume about P. Every other fact about P can (and must) be proved from these two.

Theorem. *If $P : \mathcal{A} \to [0,1]$ is a probability measure, then $P\{\emptyset\} = 0$.*

Proof. Let $A_1 = \Omega, A_2 = \emptyset, A_3 = \emptyset, \ldots$ and note that this sequence is pairwise disjoint. Thus, we have

$$P\left\{\bigcup_{n=1}^{\infty} A_n\right\} = P\{\Omega \cup \emptyset \cup \emptyset \cup \cdots\} = P\{\Omega\} = 1$$

where the last equality follows from axiom 1. However, by axiom 2 we have

$$P\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \sum_{n=1}^{\infty} P\{A_n\} = P\{\Omega\} + P\{\emptyset\} + P\{\emptyset\} + P\{\emptyset\} + \cdots = P\{\Omega\} + \sum_{n=2}^{\infty} P\{\emptyset\}.$$

This implies that

$$1 = 1 + \sum_{n=2}^{\infty} P\{\emptyset\} \quad \text{or} \quad \sum_{n=2}^{\infty} P\{\emptyset\} = 0.$$

However, since $P\{A\} \geq 0$ for every $A \in \mathcal{A}$, we see that the only way for a countable sum of constants to equal 0 is if each of them equals zero. Thus, $P\{\emptyset\} = 0$ as required. □

In general, countable additivity implies finite additivity as the next theorem shows. Exercise 2.17 shows the converse is not true; this is discussed in more detail below.

Theorem. *If $A_1, A_2, \ldots, A_n$ is a finite collection of pairwise disjoint elements of $\mathcal{A}$, then*

$$P\left\{\bigcup_{i=1}^{n} A_i\right\} = \sum_{i=1}^{n} P\{A_i\}.$$

Proof. For $m > n$, let $A_m = \emptyset$. Therefore, we find

$$\begin{aligned} P\left\{\bigcup_{m=1}^{\infty} A_m\right\} &= P\{A_1 \cup A_2 \cup \cdots \cup A_n \cup \emptyset \cup \emptyset \cup \cdots\} \\ &= \sum_{m=1}^{\infty} P\{A_m\} \quad \text{by axiom 2} \\ &= P\{A_1\} + P\{A_2\} + \cdots + P\{A_n\} + P\{\emptyset\} + P\{\emptyset\} \cdots \\ &= \sum_{m=1}^{n} P\{A_m\} \quad \text{since } P\{\emptyset\} = 0. \end{aligned}$$

That is, we've shown that

$$P\left\{\bigcup_{m=1}^{n} A_m\right\} = \sum_{m=1}^{n} P\{A_m\}$$

as required. □

Remark. Exercise 2.9 asks you to show that $P\{A \cup B\} = P\{A\} + P\{B\}$ if $A \cap B = \emptyset$. This fact follows immediately from this theorem.

Remark. For Exercises 2.10 and 2.12, it might be easier to do Exercise 2.12 first and show this implies Exercise 2.10.

Remark. Venn diagrams are useful for intuition. However, they are not a substitute for a careful proof. For instance, to show that $(A \cup B)^c = A^c \cap B^c$ we show the two containments

(i) $(A \cup B)^c \subseteq A^c \cap B^c$, and

(ii) $A^c \cap B^c \subseteq (A \cup B)^c$.

The way to show (i) is to let $x \in (A \cup B)^c$ be arbitrary. Use facts about unions, complements, and intersections to deduce that x must then be in $A^c \cap B^c$.

Remark. Exercise 2.17 shows that finite additivity does not imply countable additivity. Use the algebra defined in Exercise 2.17, but give an example of a sample space Ω for which the algebra of finite sets or sets with finite complement is not a σ-algebra.

Conditional Probability and Independence

Reference. Chapter 3 pages 15–16

Definition. Events A and B are said to be *independent* if

$$P\{A \text{ and } B\} = P\{A\} \cdot P\{B\}.$$

A collection $(A_i)_{i \in I}$ is an *independent collection* if every finite subset J of I satisfies

$$P\left\{\bigcap_{i \in J} A_i\right\} = \prod_{i \in J} P\{A_i\}.$$

We often say that (A_i) are *mutually independent*.

Example. Let $\Omega = \{1, 2, 3, 4\}$ and let $\mathcal{A} = 2^{\Omega}$. Define the probability $P : \mathcal{A} \to [0, 1]$ by

$$P\{A\} = \frac{|A|}{4}, \quad A \in \mathcal{A}.$$

In particular,

$$P\{1\} = P\{2\} = P\{3\} = P\{4\} = \frac{1}{4}.$$

Let $A = \{1, 2\}$, $B = \{1, 3\}$, and $C = \{2, 3\}$.

- Since
$$P\{A \cap B\} = P\{1\} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P\{A\} \cdot P\{B\}$$
we conclude that A and B are independent.

- Since
$$P\{A \cap C\} = P\{2\} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P\{A\} \cdot P\{C\}$$
we conclude that A and C are independent.

- Since
$$P\{B \cap C\} = P\{3\} = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P\{B\} \cdot P\{C\}$$
we conclude that B and C are independent.

However,

$$P\{A \cap B \cap C\} = P\{\emptyset\} = 0 \neq P\{A\} \cdot P\{B\} \cdot P\{C\}$$

so that A, B, C are NOT independent. Thus, we see that the events A, B, C are *pairwise independent* but not *mutually independent*.

Notation. We often use *independent* as synonymous with *mutually independent*.

Definition. Let A and B be events with $P\{B\} > 0$. The *conditional probability* of A given B is defined by

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}.$$

Theorem. *Let $P : \mathcal{A} \to [0,1]$ be a probability and let A, $B \in \mathcal{A}$ be events.*

(a) If $P\{B\} > 0$, then A and B are independent if and only if $P\{A|B\} = P\{A\}$.

(b) If $P\{B\} > 0$, then the operation $A \mapsto P\{A|B\}$ from $\mathcal{A}$ to $[0,1]$ defines a new probability measure on $\mathcal{A}$ called the conditional probability measure given B.

Proof. To prove (a) we must show both containments. Assume first that A and B are independent. Then by definition,

$$P\{A \cap B\} = P\{A\} \cdot P\{B\}.$$

But also by definition we have

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}.$$

Thus, substituting the first expression into the second gives

$$P\{A|B\} = \frac{P\{A\} \cdot P\{B\}}{P\{B\}} = P\{A\}$$

as required. Conversely, suppose that $P\{A|B\} = P\{A\}$. By definition,

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

which implies that

$$P\{A\} = \frac{P\{A \cap B\}}{P\{B\}}$$

and so $P\{A \cap B\} = P\{A\} \cdot P\{B\}$. Thus, A and B are independent.

To show (b), define the set function $Q : \mathcal{A} \to [0,1]$ by setting $Q\{A\} = P\{A|B\}$. In order to show that Q is a probability measure, we must check both axioms. Since $\Omega \in \mathcal{A}$, we have

$$Q\{\Omega\} = P\{\Omega|B\} = \frac{P\{\Omega \cap B\}}{P\{B\}} = \frac{P\{B\}}{P\{B\}} = 1.$$

If $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint, then

$$Q\left\{\bigcup_{i=1}^{\infty} A_i\right\} = P\left\{\bigcup_{i=1}^{\infty} A_i \middle| B\right\} = \frac{P\left\{\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right\}}{P\{B\}} = \frac{P\left\{\bigcup_{i=1}^{\infty} (A_i \cap B)\right\}}{P\{B\}}.$$

However, since the (A_i) are pairwise disjoint, so too are the $(A_i \cap B)$. Thus, by axiom 2,

$$P\left\{\bigcup_{i=1}^{\infty}(A_i \cap B)\right\} = \sum_{i=1}^{\infty} P\{A_i \cap B\} = \sum_{i=1}^{\infty} P\{A_i|B\}P\{B\}$$

which implies that

$$Q\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i|B\} = \sum_{i=1}^{\infty} Q\{A_i\}$$

as required. □

As noted earlier, finite additivity of P does not, in general, imply countable additivity of P (axiom 2). However, the following theorem gives us some conditions under which they are equivalent.

Theorem. *Let $\mathcal{A}$ be a σ-algebra and suppose that $P : \mathcal{A} \to [0,1]$ is a set function satisfying*

(1) $P\{\Omega\} = 1$, and

(2) if $A_1, A_2, \ldots, A_n \in \mathcal{A}$ are pairwise disjoint, then

$$P\left\{\bigcup_{i=1}^{n} A_i\right\} = \sum_{i=1}^{n} P\{A_i\}.$$

Then, the following are equivalent.

(i) countable additivity: If $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint, then

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\}.$$

(ii) If $A_n \in \mathcal{A}$ with $A_n \downarrow \emptyset$, then $P\{A_n\} \downarrow 0$.

(iii) If $A_n \in \mathcal{A}$ with $A_n \downarrow A$, then $P\{A_n\} \downarrow P\{A\}$.

(iv) If $A_n \in \mathcal{A}$ with $A_n \uparrow \Omega$, then $P\{A_n\} \uparrow 1$.

(v) If $A_n \in \mathcal{A}$ with $A_n \uparrow A$, then $P\{A_n\} \uparrow P\{A\}$.

Notation. The notation $A_n \uparrow A$ means

$$A_n \subseteq A_{n+1} \quad \text{for all } n \text{ and } \quad \bigcup_{n=1}^{\infty} A_n = A$$

(i.e., grow out and stop at A), and the notation $A_n \downarrow A$ means

$$A_n \supseteq A_{n+1} \quad \text{for all } n \text{ and } \quad \bigcap_{n=1}^{\infty} A_n = A$$

(i.e., shrink in and stop at A).

Statistics 851 (Winter 2008) January 16, 2008
Prof. Michael Kozdron

Lecture #5: Conditional Probability

Reference. Chapter 3 pages 15–18

Definition. A collection of events (E_n) is called a *partition* (of Ω) if $E_n \in \mathcal{A}$ with $P\{E_n\} > 0$ for all n, the events (E_n) are pairwise disjoint, and

$$\bigcup_n E_n = \Omega.$$

Remark. Banach and Tarski showed that if one assumes the Axiom of Choice (as it is throughout conventional mathematics), then given any two bounded subsets A and B of $\mathbb{R}^3$ (each with non-empty interior) it is possible to partition A into n pieces and B into n pieces, i.e.,

$$A = \bigcup_{i=1}^{n} A_i \quad \text{and} \quad B = \bigcup_{i=1}^{n} B_i,$$

in such a way that A_i is Euclid-congruent to B_i for each i. That is, we can disassemble A and rebuild it as B!

Theorem (Partition Theorem)**.** *If (E_n) partition Ω and $A \in \mathcal{A}$, then*

$$P\{A\} = \sum_n P\{A|E_n\}P\{E_n\}.$$

Proof. Notice that

$$A = A \cap \Omega = A \cap \left(\bigcup_n E_n \right) = \bigcup_n (A \cap E_n)$$

since (E_n) partition Ω. Since the (E_n) are disjoint, so too are the $(A \cap E_n)$. Therefore, by axiom 2 of the definition of probability, we find

$$P\{A\} = P\left\{ \bigcup_n (A \cap E_n) \right\} = \sum_n P\{A \cap E_n\}.$$

By the definition of conditional probability, $P\{A \cap E_n\} = P\{A|E_n\}P\{E_n\}$ and so

$$P\{A\} = \sum_n P\{A|E_n\}P\{E_n\}$$

as required. □

"Easy" Bayes' Theorem

Since $A \cap B = B \cap A$ we see that $P\{A \cap B\} = P\{B \cap A\}$ and so by the definition of conditional probability

$$P\{A|B\}P\{B\} = P\{B|A\}P\{A\}.$$

Solving gives

$$P\{B|A\} = \frac{P\{A|B\}P\{B\}}{P\{A\}}.$$

"Medium" Bayes' Theorem

If $P\{B\} \in (0,1)$, then since (B, B^c) partition Ω, we can use the partition theorem to conclude

$$P\{A\} = P\{A|B\}P\{B\} + P\{A|B^c\}P\{B^c\}$$

and so

$$P\{B|A\} = \frac{P\{A|B\}P\{B\}}{P\{A|B\}P\{B\} + P\{A|B^c\}P\{B^c\}}.$$

"Hard" Bayes' Theorem

Theorem (Bayes' Theorem)**.** *If (E_n) partition Ω and $A \in \mathcal{A}$ with $P\{A\} > 0$, then*

$$P\{E_j|A\} = \frac{P\{A|E_j\}P\{E_j\}}{\sum_n P\{A|E_n\}P\{E_n\}}.$$

Proof. As in the "easy" Bayes' theorem, we have

$$P\{E_j|A\} = \frac{P\{A|E_j\}P\{E_j\}}{P\{A\}}.$$

By the partition equation, we have

$$P\{A\} = \sum_n P\{A|E_n\}P\{E_n\}$$

and so combining these two equations proves the theorem. □

The following exercise illustrates the use of Bayes' theorem.

Exercise. Approximately 1/125 of all births are fraternal twins and 1/300 of births are identical twins. Elvis Presley had a twin brother (who died at birth). What is the probability that Elvis was an identical twin? (You may approximate the probability of a boy or a girl as 1/2.)

Theorem. *If $A_1, A_2, \ldots, A_n \in \mathcal{A}$ and $P\{A_1 \cap A_2 \cap \cdots \cap A_{n-1}\} > 0$, then*

$$P\{A_1 \cap \cdots \cap A_n\} = P\{A_1\}P\{A_2|A_1\}P\{A_3|A_1 \cap A_2\} \cdots P\{A_n|A_1 \cap \cdots \cap A_{n-1}\}. \qquad (*)$$

Proof. This result follows by induction on n. For $n = 1$ it is a tautology and for $n = 2$ we have

$$P\{A_1 \cap A_2\} = P\{A_1\}P\{A_2|A_1\}$$

by the definition of conditional probability. For general k suppose that $(*)$ holds. We will show that it must hold for $k + 1$. Let $B = A_1 \cap A_2 \cap \cdots \cap A_k$ so that

$$P\{A_{k+1} \cap B\} = P\{B\}P\{A_{k+1}|B\}.$$

By the inductive hypothesis, we have assumed that

$$P\{B\} = P\{A_1\}P\{A_2|A_1\}P\{A_3|A_1 \cap A_2\} \cdots P\{A_k|A_1 \cap \cdots \cap A_{k-1}\}.$$

Substituting in for B and $P\{B\}$ gives the result. □

Remark. We have finished with Chapters 1, 2, 3. Read through these chapters, especially the proof of Theorem 2.3.

Probability on a Countable Space

Reference. Chapter 4 pages 21–25

Definition. A space Ω is *finite* if for some natural number $N < \infty$, the elements of Ω can be written

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}.$$

We write $|\Omega| = \#(\Omega) = N$ for the *cardinality* of Ω.

Fact. If Ω is finite, then the power set of Ω, written 2^Ω, is a finite σ-algebra with cardinality $2^{|\Omega|}$.

Definition. A space Ω is *countable* if its elements can be put into a one-to-one correspondence with $\mathbb{N}$, the set of natural numbers.

Example. The following sets are all countable:

- the natural numbers, $\mathbb{N} = \{1, 2, 3, \ldots\}$,
- the non-negative integers, $\mathbb{Z}^+ = \{0, 1, 2, 3, \ldots\}$,
- the integers, $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$, and
- the rational numbers, $\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$.

The cardinality of a countable set is defined as $\aleph_0$ and pronounced aleph-nought.

With countable sets "funny things" can happen such as the following example shows.

Example. Although the even numbers are a strict subset of the natural numbers, there are the "same" number of evens as naturals! That is, they both have cardinality $\aleph_0$ since the set of even numbers can be put into a one-to-one correspondence with the natural numbers:

$$\begin{array}{cccccc} 1 & 2 & 3 & 4 & \text{etc.} \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\ 2 & 4 & 6 & 8 & \end{array}$$

Definition. A space which is neither countable nor finite is called *uncountable*.

Example. The space $\mathbb{R}$ is uncountable, the interval $[0, 1]$ is uncountable, and the irrational numbers are uncountable.

Fact. Exercise 2.1 can be modified to show that the power set of *any* space (whether finite, countable, or uncountable) is a σ-algebra.

Our goal now is to define probability measures on countable spaces Ω. That is, we have $\Omega = \{\omega_1, \omega_2, \ldots\}$ and the σ-algebra 2^Ω.

To define P we need to define $P\{A\}$ for all $A \in \mathcal{A}$. It is enough to specify $P\{\omega\}$ for all $\omega \in \Omega$ instead. (We will prove this.) We must also ensure that

$$\sum_{\omega \in \Omega} P\{\omega\} = \sum_{i=1}^{\infty} P\{\omega_i\} = 1.$$

Definition. We often say that P is *supported* on $\Omega' \subseteq \Omega$ is $P\{\omega\} > 0$ for every $\omega \in \Omega'$, and we call these elements of Ω' the *atoms* of P.

Example. Let $\Omega = \{1, 2, 3, \ldots\}$ and let

$$P\{2\} = \frac{1}{3}, \quad P\{17\} = \frac{5}{9}, \quad P\{851\} = \frac{1}{9},$$

and $P\{\omega\} = 0$ for all other ω. The support of P is $\Omega' = \{2, 17, 851\}$ and the atoms of P are 2, 17, 851.

Example. The Poisson distribution with parameter $\lambda > 0$ is the probability defined on $\mathbb{Z}^+$ by

$$P\{n\} = \frac{e^{-\lambda}\lambda^n}{n!}, \quad n = 0, 1, 2, \ldots.$$

Note that $P\{n\} > 0$ for all $n \in Z^+$ so that the support of P is $\mathbb{Z}^+$. Furthermore,

$$\sum_{n=0}^{\infty} P\{n\} = \sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} e^{\lambda} = 1.$$

Example. The geometric distribution with parameter α, $0 \leq \alpha < 1$, is the probability defined on $\mathbb{Z}^+$ by

$$P\{n\} = (1-\alpha)\alpha^n, \quad n = 0, 1, 2, \ldots.$$

By convention, we take $0^0 = 1$. Note that if $\alpha = 0$, then $P\{0\} = 1$ and $P\{n\} = 0$ for all $n = 1, 2, \ldots$. Furthermore,

$$\sum_{n=0}^{\infty} P\{n\} = P\{0\} = 1.$$

If $\alpha \neq 0$, then $P\{n\} > 0$ for all $n \in \mathbb{Z}^+$. We also have

$$\sum_{n=0}^{\infty} P\{n\} = \sum_{n=0}^{\infty} (1-\alpha)\alpha^n = (1-\alpha) \sum_{n=0}^{\infty} \alpha^n = (1-\alpha) \cdot \frac{1}{1-\alpha} = 1.$$

Theorem. *(a) A probability P on a finite or countable set Ω is characterized by $P\{\omega\}$, its value on the atoms.*

(b) Let (p_ω), $\omega \in \Omega$, be a family of real numbers indexed by Ω (either finite or countable). There exists a unique probability P such that $P\{\omega\} = p_\omega$ if and only if $p_\omega \geq 0$ and

$$\sum_{\omega \in \Omega} p_\omega = 1.$$

Statistics 851 (Winter 2008) January 18, 2008
Prof. Michael Kozdron

Lecture #6: Probability on a Countable Space

Reference. Chapter 4 pages 21–24

We begin with the proof of the theorem given at the end of last lecture.

Theorem. *(a) A probability P on a finite or countable set Ω is characterized by $P\{\omega\}$, its value on the atoms.*

(b) Let (p_ω), $\omega \in \Omega$, be a family of real numbers indexed by Ω (either finite or countable). There exists a unique probability P such that $P\{\omega\} = p_\omega$ if and only if $p_\omega \geq 0$ and

$$\sum_{\omega\in\Omega} p_\omega = 1.$$

Proof. (a) Let $A \in \mathcal{A}$. We can then write

$$A = \bigcup_{\omega\in A} \{\omega\}$$

which is a disjoint union of at most countably many elements. Therefore, by axiom 2 we have

$$P\{A\} = P\left\{\bigcup_{\omega\in A} \{\omega\}\right\} = \sum_{\omega\in A} P\{\omega\}.$$

Hence, to compute the probability of any event $A \in \mathcal{A}$ it is sufficient to know $P\{\omega\}$ for every atom ω.

(b) Suppose that $P\{\omega\} = p_\omega$. Since P is a probability, we have by definition that $p_\omega \geq 0$ and

$$1 = P\{\Omega\} = P\left\{\bigcup_{\omega\in\Omega} \{\omega\}\right\} = \sum_{\omega\in\Omega} P\{\omega\} = \sum_{\omega\in\Omega} p_\omega.$$

Conversely, suppose that (p_ω), $\omega \in \Omega$, are given, and that $p_\omega \geq 0$ and

$$\sum_{\omega\in\Omega} p_\omega = 1.$$

Define the function $P\{\omega\} = p_\omega$ for every $\omega \in \Omega$ and define

$$P\{A\} = \sum_{\omega\in A} p_\omega$$

for every $A \in \mathcal{A}$. By convention, the "empty" sum equals zero; that is,

$$\sum_{\omega\in\emptyset} p_\omega = 0.$$

Now all we need to do is check that $P : \mathcal{A} \to [0,1]$ satisfies the definition of probability. Since

$$P\{\Omega\} = \sum_{\omega \in \Omega} p_\omega = 1$$

by assumption, and since countable additivity follows from

$$\sum_{i \in I} \sum_{\omega \in A_i} p_\omega = \sum_{\omega \in \bigcup_{i \in I} A_i} p_\omega$$

if (A_i) are pairwise disjoint, the proof is complete. □

Example. Since

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

we can define a probability measure on $\mathbb{N}$ by setting

$$P\{k\} = \frac{6}{\pi^2 k^2}$$

for $k = 1, 2, 3, \ldots$.

Example. A probability P on a finite set Ω is called *uniform* if $P\{\omega\} = p_\omega$ does not depend on ω. That is, if

$$P\{\omega\} = \frac{1}{|\Omega|} = \frac{1}{\#(\Omega)}.$$

For $A \in \mathcal{A}$, let

$$P\{A\} = \frac{|A|}{|\Omega|} = \frac{\#(A)}{\#(\Omega)}.$$

Example. Fix a natural number N and let $\Omega = \{0, 1, 2, \ldots, N\}$. Let p be a real number with $0 < p < 1$. The *binomial distribution with parameter* p is the probability P defined on (the power set of) Ω by

$$P\{k\} = \binom{N}{k} p^k (1-p)^{N-k}, \quad k = 0, 1, \ldots, N$$

where

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} = {}_N C_k.$$

Remark. Jacod and Protter choose to introduce the hypergeometric and binomial distributions in Chapter 4 via random variables. We will wait until Chapter 5 for this approach.

Random Variables on a Countable Space

Reference. Chapter 5 pages 27–32

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Recall that this means that Ω is a space of outcomes (also called the sample space), $\mathcal{A}$ is a σ-algebra of subsets of Ω (see Definition 2.1), and $P : \mathcal{A} \to [0, 1]$ is a probability measure on $\mathcal{A}$ (see Definition 2.3).

As we have already seen, care needs to be shown when dealing with "infinities." In particular, $\mathcal{A}$ must be closed under countable unions and P is countably additive.

Assume for the rest of this lecture that Ω is either finite or countable. As such, unless otherwise noted, we can consider $\mathcal{A} = 2^\Omega$ as our underlying σ-algebra.

Definition. A *random variable* is a function $X : \Omega \to \mathbb{R}$ given by $\omega \mapsto X(\omega)$.

Note. A random variable is not a variable in the high-school or calculus sense. It is a function.

- In calculus, we study the function $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto f(x)$
- In probability, we study the function $X : \Omega \to \mathbb{R}$ given by $\omega \to X(\omega)$.

Note. Since Ω is countable, every function $X : \Omega \to \mathbb{R}$ is a random variable. There will be problems with this definition when Ω is uncountable.

Note. Since Ω is countable, the *range* of X, denoted $X(\Omega) = T'$, is also countable. This notation might lead to a bit of confusion. A random variable defined on a countable space Ω is a real-valued function although its range is a countable subset of $\mathbb{R}$. The term that is often used is *codomain* to describe the role of the target space (in this case $\mathbb{R}$).

Definition. The *law* (or *distribution*) of X is the probability measure on $(T', 2^{T'})$ given by

$$P^X\{B\} = P\{\omega : X(\omega) \in B\}, \quad B \in 2^{T'}.$$

Note that X transforms the probability space $(\Omega, \mathcal{A}, P)$ into the probability space $(T', 2^{T'}, P^X)$.

Notation.

$$P^X\{B\} = P\{\omega : X(\omega) \in B\} = P\{X \in B\} = P\{X^{-1}(B)\}.$$

Notation. We write

$$P\{\omega\} = p_\omega \quad \text{and} \quad P^X\{j\} = p_j^X.$$

Since T' is countable, P^X, the law of X, can be characterized by p_j^X. That is,

$$P^X\{B\} = \sum_{j \in B} P^X\{j\} = \sum_{j \in B} p_j^X.$$

However, we also have

$$p_j^X = \sum_{\omega \in \Omega : X(\omega) = j} p_\omega.$$

That is,

$$P\{X = j\} = \sum_{\omega: X(\omega)=j} P\{\omega\}.$$

In undergraduate probability, we would say that the random variable X is *discrete* if T' is either finite or countable, and we would call $P\{X = j\}$, $j \in T'$, the *probability mass function* of X.

Definition. We say that X is a NAME random variable if the law of X has the NAME distribution.

Example. X is a Poisson random variable if P^X is the Poisson distribution. Recall that P^X has a Poisson distribution if

$$P^X\{j\} = \frac{e^{-\lambda}\lambda^j}{j!}, \quad j = 0, 1, 2, \ldots.$$

That is,

$$P\{X = j\} = P^X\{j\} = \frac{e^{-\lambda}\lambda^j}{j!}.$$

Lecture #7: Random Variables on a Countable Space

Reference. Chapter 5 pages 27–32

Suppose that Ω is a countable (or finite) space, and let $(\Omega, \mathcal{A}, P)$ be a probability space.

Let $X : \Omega \to T$ be a random variable which means that X is a T-valued function on Ω. We call T the *state space* (or *range space* or *co-domain*) and usually take $T = \mathbb{R}$ or $T = \mathbb{R}^n$. The *range* of X is defined to be

$$T' = X(\Omega) = \{t \in T : X(\omega) = t \text{ for some } \omega \in \Omega\}.$$

In particular, since Ω is countable, so too is T'.

As we saw last lecture, the random variable X transforms the probability space $(\Omega, \mathcal{A}, P)$ into a new probability space $(T', 2^{T'}, P^X)$ where

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

is the law of X and characterized by

$$p_j^X = P\{X = j\} = \sum_{\{\omega : X(\omega) = j\}} P\{\omega\}.$$

We often write $P\{\omega\} = p_\omega$.

One useful way to summarize a random variable is through its expected value.

Definition. Suppose that X is a random variable on a countable space Ω. The *expected value*, or *expectation*, of X is defined to be

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) P\{\omega\} = \sum_{\omega \in \Omega} X(w) p_\omega$$

provided the sum makes sense.

- If Ω is finite, then $\mathbb{E}(X)$ makes sense.
- If Ω is countable, then $\mathbb{E}(X)$ makes sense only if $\sum X(\omega) p_\omega$ is absolutely convergent.
- If Ω is countable and $X \geq 0$, then $\mathbb{E}(X)$ makes sense but $\mathbb{E}(X) = +\infty$ is allowed.

Remark. This definition is given in the "modern" language. We prefer to use it in anticipation of later developments.

In undergraduate probability classes, you were told: If X is a discrete random variable, then

$$\mathbb{E}(X) = \sum_j jP\{X = j\}.$$

We will now show that these definitions are equivalent. Begin by writing

$$\Omega = \bigcup_j \{\omega : X(\omega) = j\}$$

which represents the sample space Ω as a disjoint, countable union. This is a very, very useful decomposition. Notice that we are really partitioning Ω according to range of X. A less-useful decomposition is to partition Ω according to the domain of X, namely Ω itself. That is, if we enumerate Ω as $\Omega = \{\omega_1, \omega_2, \ldots\}$, then

$$\Omega = \bigcup_j \{\omega_j\}$$

is also a disjoint, countable union. We now find

$$\begin{aligned}
\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)p_\omega = \sum_{\omega\in\bigcup_j\{\omega:X(\omega)=j\}} X(\omega)p_\omega = \sum_j \sum_{\omega\in\{X(\omega)=j\}} X(\omega)p_\omega &= \sum_j \sum_{\omega\in\{X(\omega)=j\}} jp_\omega \\
&= \sum_j j \sum_{\omega\in\{X(\omega)=j\}} p_\omega \\
&= \sum_j jp_j^X \\
&= \sum_j jP\{X = j\}
\end{aligned}$$

which shows that our modern usage agrees with our undergraduate usage.

Definition. Write $\mathcal{L}^1$ to denote the set of all real-valued random variables on $(\Omega, \mathcal{A}, P)$ with finite expectation. That is,

$$\mathcal{L}^1 = \mathcal{L}^1(\Omega, \mathcal{A}, P) = \{X : (\Omega, \mathcal{A}, P) \to \mathbb{R} : \mathbb{E}(X) < \infty\}.$$

Proposition. *$\mathbb{E}$ is a linear operator on $\mathcal{L}^1$. That is, if X, $Y \in \mathcal{L}^1$ and α, $\beta \in \mathbb{R}$, then*

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y).$$

Proof. Using properties of summations, we find

$$\mathbb{E}(\alpha X + \beta Y) = \sum_{\omega\in\Omega}(\alpha X(\omega) + \beta Y(\omega))p_\omega = \alpha\sum_{\omega\in\Omega} X(\omega)p_\omega + \beta\sum_{\omega\in\Omega} Y(\omega)p_\omega = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$$

as required. □

Remark. It follows from this result that if X, $Y \in \mathcal{L}^1$ and α, $\beta \in \mathbb{R}$, then $\alpha X + \beta Y \in \mathcal{L}^1$.

Proposition. *If X, $Y \in \mathcal{L}^1$ with $X \leq Y$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.*

Proof. Using properties of summations, we find

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) p_\omega \leq \sum_{\omega \in \Omega} Y(\omega) p_\omega = \mathbb{E}(Y)$$

as required. □

Proposition. *If $A \in \mathcal{A}$ and $X(\omega) = 1_A(\omega)$, then $\mathbb{E}(X) = P\{A\}$.*

Proof. Recall that

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

Therefore,

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) p_\omega = \sum_{\omega \in \Omega} 1_A(\omega) p_\omega = \sum_{\omega \in A} 1_A(\omega) p_\omega + \sum_{\omega \in A^c} 1_A(\omega) p_\omega = \sum_{\omega \in A} p_\omega = P\{A\}$$

as required. □

Fact. Other facts about $\mathcal{L}^1$ include:

- $\mathcal{L}^1$ is a vector space,
- $\mathcal{L}^1$ is an inner product space with

$$(X, Y) = \langle X, Y \rangle = \mathbb{E}(XY),$$

- $\mathcal{L}^1$ contains all bounded random variables (that is, $X \leq c$ implies $\mathbb{E}(X) \leq c$), and
- if $X^2 \in \mathcal{L}^1$, then $X \in \mathcal{L}^1$.

Lecture #8: Expectation of Random Variables on a Countable Space

Reference. Chapter 5 pages 27–32

Recall that last lecture we defined the concept of expectation, or expected value, of a random variable on a countable space.

Throughout this lecture, we will assume that Ω is a countable space and we will let $(\Omega, \mathcal{A}, P)$ be a probability space.

Definition. Suppose that $X : \Omega \to \mathbb{R}$ is a random variable. The *expected value*, or *expectation*, of X is defined to be

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)P\{\omega\} = \sum_{\omega\in\Omega} X(w)p_\omega$$

provided the sum makes sense.

An important inequality involving expectation is given by the following theorem.

Theorem. *Suppose that $X : \Omega \to \mathbb{R}$ is a random variable and $h : \mathbb{R} \to [0, \infty)$ is a non-negative function. Then,*

$$P\{\omega : h(X(\omega)) \geq a\} \leq \frac{\mathbb{E}(h(X))}{a}$$

for any $a > 0$.

Proof. Let $Y = h(X)$ and note that Y is also a random variable. Let $a > 0$ and define the set A to be

$$A = Y^{-1}([a, \infty)) = \{\omega : h(X(\omega)) \geq a\} = \{h(X) \geq a\}.$$

Notice that $h(X) \geq a1_A$ where

$$1_A(\omega) = \begin{cases} 1, & \text{if } h(X(\omega)) \geq a, \\ 0, & \text{if } h(X(\omega)) < a. \end{cases}$$

Thus,

$$\mathbb{E}(h(X)) \geq \mathbb{E}(a1_A) = a\mathbb{E}(1_A) = aP\{A\},$$

or, in other words,

$$P\{h(X) \geq a\} \leq \frac{\mathbb{E}(h(X))}{a}$$

as required. □

Corollary (Markov's Inequality)**.** *If $a > 0$, then*

$$P\{|X| \geq a\} \leq \frac{\mathbb{E}(|X|)}{a}.$$

Proof. Take $h(x) = |x|$ in the previous theorem. □

Corollary (Chebychev's Inequality)**.** *If X^2 is a random variable with finite expectation, then*

(i) $P\{|X| \geq a\} \leq \dfrac{\mathbb{E}(X^2)}{a^2}$, *and*

(ii) $P\{|X - \mathbb{E}(X)| \geq a\} \leq \dfrac{\sigma_X^2}{a^2}$

for any $a > 0$ where $\sigma_X^2 = \mathbb{E}(X - \mathbb{E}(X))^2$ denotes the variance *of X.*

Proof. For (i), take $h(x) = x^2$ and note that

$$P\{|X| \geq a\} = P\{X^2 \geq a^2\}.$$

For (ii), take $Y = |X - \mathbb{E}(X)|$ and note that

$$P\{|X - \mathbb{E}(X)| \geq a\} = P\{Y \geq a\} = P\{Y^2 \geq a^2\}.$$

The results now follow from the previous theorem. □

Also recall that if X is a discrete random variable, then we can express the expected value of X as

$$\mathbb{E}(X) = \sum_j jP\{X = j\}.$$

This formula is often useful for calculations as the following examples show.

Example. Recall that X is a Poisson random variable with parameter $\lambda > 0$ if

$$P\{X = j\} = P^X\{j\} = \frac{e^{-\lambda}\lambda^j}{j!}.$$

Compute $\mathbb{E}(X)$.

Solution. We find

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

Exercise. If X is a Poisson random variable with parameter $\lambda > 0$, compute $\mathbb{E}(X(X-1))$. Use this result to determine $\mathbb{E}(X^2)$.

Example. The random variable X is Bernoulli with parameter α if

$$P\{X = 1\} = \alpha \quad \text{and} \quad P\{X = 0\} = 1 - \alpha$$

for some $0 < \alpha < 1$. Compute $\mathbb{E}(X)$.

Solution. We find

$$\mathbb{E}(X) = \sum_{j=0}^{1} jP\{X = j\} = 0 \cdot P\{X = 0\} + 1 \cdot P\{X = 1\} = \alpha.$$

Example. Fix a natural number N and let $\Omega = \{0, 1, 2, \ldots, N\}$. Let p be a real number with $0 < p < 1$. Recall that X is a Binomial random variable with parameter p if

$$P\{X = k\} = \binom{N}{k} p^k (1 - p)^{N-k}, \quad k = 0, 1, \ldots, N$$

where

$$\binom{N}{k} = \frac{N!}{k!(N - k)!}.$$

Compute $\mathbb{E}(X)$.

Solution. It is a fact that X can be represented as $X = Y_1 + \cdots + Y_N$ where $Y_1, Y_2, \ldots, Y_N$ are independent Bernoulli random variables with parameter p. (We will prove this fact later in the course.) Since $\mathbb{E}$ is a linear operator, we conclude

$$\mathbb{E}(X) = \mathbb{E}(Y_1 + \cdots + Y_N) = \mathbb{E}(Y_1) + \cdots + \mathbb{E}(Y_N) = p + \cdots + p = Np.$$

(We will also need to be careful about the precise probability spaces on which each random variable is defined in order to justify the calculation $\mathbb{E}(X) = Np$.) Note that, once justified, this calculation is much easier than attempting to compute

$$\mathbb{E}(X) = \sum_{k=0}^{N} k \cdot \binom{N}{k} p^k (1 - p)^{N-k}.$$

Example. Recall that X is a Geometric random variable if

$$P\{X = j\} = (1 - \alpha)\alpha^j, \quad j = 0, 1, 2, \ldots$$

for some $0 \leq \alpha < 1$. Compute $\mathbb{E}(X)$.

Solution. Note that if $\alpha = 0$, then $P\{X = 0\} = 1$ in which case $\mathbb{E}(X) = 0$. Therefore, assume that $0 < \alpha < 1$. We find

$$\mathbb{E}(X) = \sum_{j=0}^{\infty} j(1 - \alpha)\alpha^j = (1 - \alpha) \sum_{j=0}^{\infty} j\alpha^j = (1 - \alpha) \sum_{j=1}^{\infty} j\alpha^j = \alpha(1 - \alpha) \sum_{j=1}^{\infty} j\alpha^{j-1}.$$

Notice that

$$j\alpha^{j-1} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\alpha^j$$

and so

$$\sum_{j=1}^{\infty} j\alpha^{j-1} = \sum_{j=1}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\alpha}\alpha^j = \frac{\mathrm{d}}{\mathrm{d}\alpha}\sum_{j=1}^{\infty}\alpha^j.$$

The last equality relies on the assumption that the sum and derivative can be interchanged. (They can be, although we will not address this point here.) Also notice that

$$\sum_{j=1}^{\infty}\alpha^j = \sum_{j=0}^{\infty}\alpha^j - 1 = \frac{1}{1-\alpha} - 1 = \frac{\alpha}{1-\alpha}$$

since it is a geometric series. Thus, we conclude

$$\mathbb{E}(X) = \alpha(1-\alpha)\cdot\frac{\mathrm{d}}{\mathrm{d}\alpha}\frac{\alpha}{1-\alpha} = \alpha(1-\alpha)\cdot\frac{1}{(1-\alpha)^2} = \frac{\alpha}{1-\alpha}.$$

Exercise. If X is a Geometric random variable, compute $\mathbb{E}(X^2)$.

Remark. Note that Jacod and Protter also use

$$P\{X = j\} = \alpha(1-\alpha)^j, \quad j = 0, 1, 2, \ldots$$

for some $0 \leq \alpha < 1$ as the parametrization of a Geometric random variable. In this case, we would find

$$\mathbb{E}(X) = \frac{1-\alpha}{\alpha}.$$

Statistics 851 (Winter 2008) January 25, 2008
Prof. Michael Kozdron

Lecture #9: A Non-Measurable Set

Reference. Today's notes and §1.2 of [2]

We will begin our discussion of probability measures and random variables on uncountable spaces with the construction of a non-measurable set. What does this mean?

Consider the uncountable sample space $\Omega = [0,1]$. Our goal is to construct the uniform probability measure $P : \mathcal{A} \to [0,1]$ defined for all events A in some *appropriate* σ-algebra $\mathcal{A}$ of subsets of Ω.

As we discussed earlier, 2^Ω, the power set of Ω, is always a σ-algebra. The problem that we will encounter is that this σ-algebra will be too big! In particular, we will be able to construct sets $A \in 2^\Omega$ such that $P\{A\}$ cannot be defined in a consistent way.

We begin by recalling the definition of probability as given in Lecture #4.

Definition. Let Ω be a sample space and let $\mathcal{A}$ be a σ-algebra of subsets of Ω. A set function $P : \mathcal{A} \to [0,1]$ is called a *probability measure* on $\mathcal{A}$ if

(1) $P\{\Omega\} = 1$, and

(2) if $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint, then

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\}.$$

Suppose that P is (our candidate for) the uniform probability measure on $([0,1], 2^{[0,1]})$. This means that if $a < b$ then

$$P\{[a,b]\} = P\{(a,b)\} = P\{[a,b)\} = P\{(a,b]\} = b - a.$$

In other words, the probability of any interval is just its length. In particular,

$$P\{a\} = 0 \quad \text{for every } 0 \leq a \leq 1.$$

Furthermore, if $0 \leq a_1 < b_1 < a_2 < b_2 < \cdots < a_n < b_n \leq 1$, then

$$P\left\{\bigcup_{i=1}^{n} |a_i, b_i|\right\} = \sum_{i=1}^{\infty} P\{|a_i, b_i|\} = \sum_{i=1}^{\infty} (b_i - a_i)$$

where $|a,b|$ means that either end could be open or closed. Although this is the statement of countable additivity it should make sense to you intuitively. For instance, the probability that you fall in the interval $[0, 1/4]$ is $1/4$, the probability that you fall in the interval $[1/3, 1/2]$ is $1/6$, and the probability that you fall in either the interval $[0, 1/4]$ or $[1/3, 1/2]$ should be $1/4 + 1/6 = 5/12$. That is,

$$P\{[0,1/4] \cup [1/3,1/2]\} = P\{[0,1/4]\} + P\{[1/3,1/2]\} = \frac{1}{4} + \frac{1}{6} = \frac{5}{12}.$$

Remark. The reason that we don't allow uncountable additivity is the following. Begin by writing

$$[0,1] = \bigcup_{\omega\in[0,1]} \{\omega\}.$$

We therefore have

$$1 = P\{[0,1]\} = P\left\{\bigcup_{\omega\in[0,1]} \{\omega\}\right\}.$$

If uncountable additivity were allowed, we would have

$$P\left\{\bigcup_{\omega\in[0,1]} \{\omega\}\right\} = \sum_{\omega\in[0,1]} P\{\omega\} = \sum_{\omega\in[0,1]} 0.$$

This would force

$$1 = \sum_{\omega\in[0,1]} 0$$

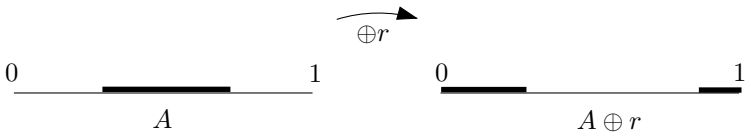
and we really don't want an uncountable sum of 0s to be 1!

If P is to be the uniform probability, then it should be unaffected by shifting. In particular, it should only depend on the length of the interval and not the endpoints themselves. For instance,

$$P\{[0,1/4]\} = P\{[3/4,1]\} = P\{[1/6,5/12]\} = \frac{1}{4}.$$

We can write this (allowing for "wrapping around" as)

$$P\{[0,1/4] \oplus r\} = P\{[r,1/4+r]\} = \frac{1}{4} \quad \text{for every } r.$$



Thus, it is reasonable to assume that

$$P\{A \oplus r\} = P\{A\}$$

where

$$A \oplus r = \{a + r : a \in A, a + r \le 1\} \cup \{a + r - 1 : a \in A, a + r > 1\}.$$

To prove that no uniform probability measure exists for every $A \in 2^{[0,1]}$ we will derive a contradiction. Suppose that there exists such as P.

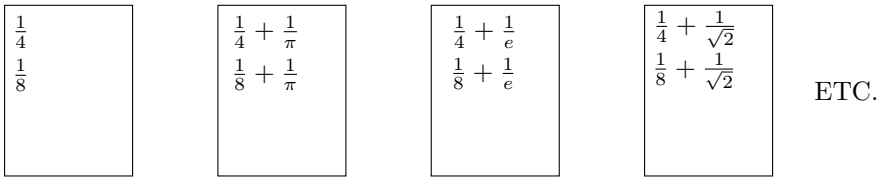
Define an *equivalence relation* $x \sim y$ if $y - x \in \mathbb{Q}$. For instance,

$$\frac{1}{2} \sim \frac{1}{4}, \quad \frac{1}{9} \sim \frac{1}{27}, \quad \frac{1}{9} \sim \frac{1}{4}, \quad \frac{1}{3} \not\sim \frac{1}{\pi}, \quad \frac{1}{e} \not\sim \frac{1}{\pi}.$$

This equivalence relationship partitions $[0,1]$ into a disjoint union of equivalence classes (with two elements of the same class differing by a rational, but elements of different classes differing by an irrational).

Let $\mathbb{Q}_1 = [0,1] \cap \mathbb{Q}$, and note that there are uncountably many equivalence classes. Formally, we can write this disjoint union as

$$[0,1] = \mathbb{Q}_1 \cup \left\{ \bigcup_{x \in [0,1] \setminus \mathbb{Q}_1} \{(\mathbb{Q}_1 + x) \cap [0,1]\} \right\}.$$



Let H be the subset of $[0,1]$ consisting of precisely one element from each equivalence class. (This step uses the Axiom of Choice.)

For definiteness, assume that $0 \notin H$. Therefore,

$$(0,1] = \bigcup_{r \in \mathbb{Q}_1, r \neq 1} \{H \oplus r\}$$

with $\{H \oplus r_i\} \cap \{H \oplus r_j\} = \emptyset$ for all $i \neq j$. This implies

$$P\{(0,1]\} = P\left\{ \bigcup_{r \in \mathbb{Q}_1, r \neq 1} \{H \oplus r\} \right\} = \sum_{r \in \mathbb{Q}_1, r \neq 1} P\{H \oplus r\} = \sum_{r \in \mathbb{Q}_1, r \neq 1} P\{H\}.$$

In other words,

$$1 = \sum_{r \in \mathbb{Q}_1, r \neq 1} P\{H\}.$$

This is a contradiction since a countable sum of a constant value must be either 0, $+\infty$, or $-\infty$. It can never be 1. Thus, $P\{H\}$ does not exist.

We can summarize our work with the following theorem.

Theorem. *Consider $\Omega = [0,1]$ with $\mathcal{A} = 2^\Omega$. There does not exist a definition of $P\{A\}$ for every $A \in \mathcal{A}$ satisfying*

- $P\{\emptyset\} = 0$, $P\{\Omega\} = 1$,
- $0 \leq P\{A\} \leq 1$ *for every* $A \in \mathcal{A}$,
- $P\{|a,b|\} = b - a$ *for all* $0 \leq a \leq b \leq 1$, *and*

- *if* $A_1, A_2, \ldots \in \mathcal{A}$ *are pairwise disjoint, then*

$$P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\}.$$

Remark. The fact that there exists a $A \in 2^{[0,1]}$ such that $P\{A\}$ does not exist means that the σ-algebra $2^{[0,1]}$ is simply too big! Instead, the "correct" σ-algebra to use is $\mathcal{B}_1$, the Borel sets of $[0,1]$. Later we will construct the uniform probability space $([0,1], \mathcal{B}_1, P)$.

Lecture #10: Construction of a Probability Measure

Reference. Chapter 6 pages 35–38

Suppose that Ω is a sample space. It Ω is finite or countable, then we can construct the probability space $(\Omega, 2^\Omega, P)$ without much trouble. However, it is not so easy if Ω is uncountable.

Problem. A "typical" probability on Ω will have

$$P\{\omega\} = p_\omega = 0$$

for *all* $\omega \in \Omega$. Hence, p_ω will NOT characterize P.

Problem. If Ω is uncountable, then, even though 2^Ω is a σ-algebra, it will often be too large. Thus, we will not be able to define $P\{A\}$ for *every* $A \in 2^\Omega$.

Solution. Our solution to both these problems is the following. We will consider an algebra $\mathcal{A}_0$ with $\sigma(\mathcal{A}_0) = \mathcal{A}$.

We will define a function $P : \mathcal{A}_0 \to [0,1]$ (although P will not necessarily be a probability) and extend $P : \mathcal{A} \to [0,1]$ in a unique way (so that P is now a probability).

Remark. When $\Omega = \mathbb{R}$, then the structure of $\mathbb{R}$ can simplify some things. We will see examples of this in Chaper 7.

Definition. A class $\mathcal{C}$ of subsets of Ω is *closed under finite intersections* if for every $n < \infty$ and for every $A_1, \ldots, A_n \in \mathcal{C}$,

$$\bigcap_{i=1}^{n} A_i \in \mathcal{C}.$$

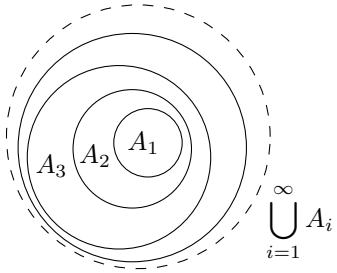
Example. If $\Omega = \{a, b, c, d\}$ and

$$\mathcal{C} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\},$$

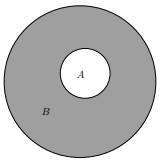
then $\mathcal{C}$ is not a σ-algebra (or even an algebra), but it is closed under finite intersections.

Definition. A class $\mathcal{C}$ of subsets of Ω is *closed under increasing limits* if for every collection $A_1, A_2, \ldots \in \mathcal{C}$ with $A_1 \subset A_2 \subset \cdots$, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{C}.$$



Definition. A class $\mathcal{C}$ of subsets of Ω is *closed under finite differences* if for every A, $B \in \mathcal{C}$ with $A \subset B$, then $B \setminus A \in \mathcal{C}$.



Theorem (Monotone Class Theorem)**.** *Let $\mathcal{C}$ be a class of subsets of Ω. Suppose that $\mathcal{C}$ contains Ω (that is, $\Omega \in \mathcal{C}$) and is closed under finite intersections. Let $\mathcal{D}$ be the smallest class containing $\mathcal{C}$ which is closed under increasing limits and finite differences. Then,*

$$\mathcal{D} = \sigma(\mathcal{C}).$$

Example. As in the previous example, suppose that $\Omega = \{a, b, c, d\}$ and let

$$\mathcal{C} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

so that $\mathcal{C}$ is closed under finite intersections. Furthermore, $\mathcal{C}$ is also closed under increasing limits and finite differences as a simple calculation shows. Therefore, if $\mathcal{D}$ denotes the smallest class containing $\mathcal{C}$ which is closed under increasing limits and finite differences, then clearly $\mathcal{D} = \mathcal{C}$ itself. However, $\mathcal{C}$ is not a σ-algebra; the conclusion of the Monotone Class Theorem suggests that $\mathcal{D} = \sigma(\mathcal{C})$ which would force $\mathcal{C}$ to be a σ-algebra. This is not a contradiction since the hypothesis that $\Omega \in \mathcal{C}$ is not met. Suppose that

$$\mathcal{C}' = \mathcal{C} \cup \Omega.$$

Then $\mathcal{C}'$ is still closed under finite intersections. However, it is no longer closed under finite differences. As a calculation shows, the smallest σ-algebra containing $\mathcal{C}'$ is now $\sigma(\mathcal{C}') = 2^{\Omega}$.

Proof. We begin by noting that the intersection of classes of sets closed under increasing limits and finite differences is again a class of that type. Hence, if we take the intersection of all such classes, then there will be a smallest class containing $\mathcal{C}$ which is closed under increasing limits and by finite differences. Denote this class by $\mathcal{D}$. Also note that a σ-algebra is necessarily closed under increasing limits and by finite differences. Thus, we conclude that $\mathcal{D} \subseteq \sigma(\mathcal{C})$. To complete the proof we will show the reverse containment, namely $\sigma(\mathcal{C}) \subseteq \mathcal{D}$.

For every set $B \subseteq \Omega$, let

$$\mathcal{D}_B = \{A \subseteq \Omega : A \in \mathcal{D} \text{ and } A \cap B \in \mathcal{D}\}.$$

Since $\mathcal{D}$ is closed under increasing limits and finite differences, a calculation shows that $\mathcal{D}_B$ must also closed under increasing limits and finite differences.

Since $\mathcal{C}$ is closed under finite intersections, $\mathcal{C} \subseteq \mathcal{D}_B$ for every $B \in \mathcal{C}$. That is, suppose that $B \in \mathcal{C}$ is fixed and let $C \in \mathcal{C}$ be arbitrary. Since $\mathcal{C}$ is closed under finite intersections, we must have $B \cap C \in \mathcal{C}$. Since $\mathcal{C} \subseteq \mathcal{D}$, we conclude that $B \cap C \in \mathcal{D}$ verifying that $C \in \mathcal{D}_B$

for every $B \in \mathcal{C}$. Note that by definition we have $\mathcal{D}_B \subseteq \mathcal{D}$ for every $B \in \mathcal{C}$ and so we have shown

$$\mathcal{C} \subseteq \mathcal{D}_B \subseteq \mathcal{D}$$

for every $B \in \mathcal{C}$. Since $\mathcal{D}_B$ is closed under increasing limits and finite differences, we conclude that $\mathcal{D}$, the smallest class containing $\mathcal{C}$ closed under increasing limits and finite differences, must be contained in $\mathcal{D}_B$ for every $B \in \mathcal{C}$. That is, $\mathcal{D} \subseteq \mathcal{D}_B$ for every $B \in \mathcal{C}$. Taken together, we are forced to conclude that $\mathcal{D} = \mathcal{D}_B$ for every $B \in \mathcal{C}$.

Now suppose that $A \in \mathcal{D}$ is arbitrary. We will show that $\mathcal{C} \subseteq \mathcal{D}_A$. If $B \in \mathcal{C}$ is arbitrary, then the previous paragraph implies that $\mathcal{D} = \mathcal{D}_B$. Thus, we conclude that $A \in \mathcal{D}_B$ which implies that $A \cap B \in \mathcal{D}$. It now follows that $B \in \mathcal{D}_A$. This shows that $\mathcal{C} \subseteq \mathcal{D}_A$ for every $A \in \mathcal{D}$ as required. Since $\mathcal{D}_A \subseteq \mathcal{D}$ for every $A \in \mathcal{D}$ by definition, we have shown

$$\mathcal{C} \subseteq \mathcal{D}_A \subseteq \mathcal{D}$$

for every $A \in \mathcal{D}$. The fact that $\mathcal{D}$ is the smallest class containing $\mathcal{C}$ which is closed under increasing limits and finite differences forces us, using the same argument as above, to conclude that $\mathcal{D} = \mathcal{D}_A$ for every $A \in \mathcal{D}$.

Since $\mathcal{D} = \mathcal{D}_A$ for all $A \in \mathcal{D}$, we conclude that $\mathcal{D}$ is closed under finite intersections. Furthermore, $\Omega \in \mathcal{D}$ and $\mathcal{D}$ is closed by finite differences which implies that $\mathcal{D}$ is closed under complementation. Since $\mathcal{D}$ is also closed by increasing limits, we conclude that $\mathcal{D}$ is a σ-algebra, and it is clearly the smallest σ-algebra containing $\mathcal{C}$. Thus, $\sigma(\mathcal{C}) \subseteq \mathcal{D}$ and the proof that $\mathcal{D} = \sigma(\mathcal{C})$ is complete. □

Corollary. *Let $\mathcal{A}$ be a σ-algebra, and let P, Q be two probabilities on $(\Omega, \mathcal{A})$. Suppose that P, Q agree on a class $\mathcal{C} \subseteq \mathcal{A}$ which is closed under finite intersections. If $\sigma(\mathcal{C}) = \mathcal{A}$, then $P = Q$.*

Proof. Since $\mathcal{A}$ is a σ-algebra we know that $\Omega \in \mathcal{A}$. Since $P\{\Omega\} = Q\{\Omega\} = 1$ we can assume without loss of generality that $\Omega \subseteq \mathcal{C}$. Define

$$\mathcal{D} = \{A \in \mathcal{A} : P\{A\} = Q\{A\}\}$$

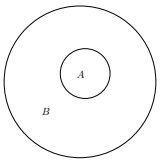
to be the class on which P and Q agree (and note that $\emptyset \in \mathcal{D}$ and $\Omega \in \mathcal{D}$ so that $\mathcal{D}$ is non-empty). By the definition of probability and Theorem 2.3, we see that $\mathcal{D}$ is closed by differences and increasing limits. By assumption, we also have $\mathcal{C} \subseteq \mathcal{D}$. Therefore, since $\sigma(\mathcal{C}) = \mathcal{A}$, we have $\mathcal{D} = \mathcal{A}$ by the Monotone Class Theorem. □

Lecture #11: Null Sets

Reference. Chapter 6 pages 35–38

Definition. Let P be a probability on $\mathcal{A}$. A *null set* (or a *negligible set*) for P is a subset $A \subseteq \Omega$ such that there exists a $B \in \mathcal{A}$ with $A \subseteq B$ and $P\{B\} = 0$.

Note. Suppose that $B \in \mathcal{A}$ with $P\{B\} = 0$. Let $A \subseteq B$ as shown below.



If $A \in \mathcal{A}$, then we can conclude that $P\{A\} = 0$. However, if $A \notin \mathcal{A}$, then $P\{A\}$ does not make sense.

In either case, A is a null set. Thus, it is natural to *define* $P\{A\} = 0$ for all null sets.

Theorem. *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space so that P is a probability on the σ-algebra $\mathcal{A}$. Let $\mathcal{N}$ denote the set of all null sets for P. Let*

$$\mathcal{A}' = \mathcal{A} \cup \mathcal{N} = \{A \cup N : A \in \mathcal{A}, N \in \mathcal{N}\}.$$

*Then $\mathcal{A}'$ is a σ-algebra, called the P-*completion *of $\mathcal{A}$, and is the smallest σ-algebra containing $\mathcal{A}$ and $\mathcal{N}$. Furthermore, P extends uniquely to a probability on $\mathcal{A}'$ (also called P) by setting*

$$P\{A \cup N\} = P\{A\}$$

for $A \in \mathcal{A}$, $N \in \mathcal{N}$.

Notation. We say that "Property B" holds *almost surely* if

$$P\{\omega : \text{Property B does not hold}\} = 0,$$

i.e., if $\{\omega : \text{Property B does not hold}\}$ is a null set.

Construction of Probabilities on $\mathbb{R}$

Reference. Chapter 7 pages 39–44

Recall that by a probability space $(\Omega, \mathcal{A}, P)$ we mean a triple where Ω is the sample space, $\mathcal{A}$ is a σ-algebra of subsets of Ω, and $P : \mathcal{A} \to [0, 1]$ is a probability measure.

A random variable is a special kind of function from Ω to $\mathbb{R}$. Schematically, we write

$$\begin{aligned} X : \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

If Ω is finite or countable, then ANY function $X : \Omega \to \mathbb{R}$ is a random variable.

However, if Ω is uncountable, then there are functions which are NOT random variables.

Example. Suppose that $\Omega = [0, 1]$ and $H \subset \Omega$ is the non-measurable set constructed in Lecture #9. The function $X : \Omega \to \mathbb{R}$ defined by

$$X(\omega) = 1_H\{\omega\}$$

is not a random variable. (This fact will be proved later.)

Also recall that if Ω is finite or countable, then we take $\mathcal{A} = 2^\Omega$ as our σ-algebra.

However, if Ω is uncountable, then the σ-algebra 2^Ω is too big. Thus, we must use a σ-algebra $\mathcal{B} \subsetneq 2^\Omega$.

If $\Omega \subseteq \mathbb{R}$ and is uncountable, say $\Omega = [0, 1]$ or $\Omega = \mathbb{R}$, then the "correct" σ-algebra to use is the Borel σ-algebra $\mathcal{B}$. To emphasize the Borel sets of Ω we write $\mathcal{B}(\Omega)$.

Recall that the Borel σ-algebra is generated by the open sets so that $\mathcal{B} = \sigma(\mathcal{O})$. For instance, $\mathcal{B}([0, 1]) = \sigma(\mathcal{O})$ where $\mathcal{O}$ denotes the set of open subsets of $[0, 1]$.

Fact. By Theorem 2.1, we know that $\mathcal{B}$ is generated by intervals of the form $(-\infty, a]$ for $a \in \mathbb{Q}$.

Suppose that $X : \Omega \to \mathbb{R}$ is a random variable. (This is not yet defined.) We have already seen that X induces a probability measure on $(\mathbb{R}, \mathcal{B})$ denoted by P^X where

$$P^X\{B\} = P\{\omega : X(\omega) \in B\} = P\{X \in B\} = P\{X^{-1}(B)\} \quad \text{for every } B \in \mathcal{B}$$

is called the *law of* X. Hence, we see that X transforms the probability space $(\Omega, \mathcal{A}, P)$ into the probability space $(\mathbb{R}, \mathcal{B}, P^X)$:

$$X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}, P^X).$$

Since the law of a random variable is always a probability measure on $(\mathbb{R}, \mathcal{B})$ (not yet proved) we have good reason to study probabilities on $\mathbb{R}$. This motivates Chapter 7.

Definition. Suppose that P is a probability measure on $(\mathbb{R}, \mathcal{B})$. The *distribution function* induced by P is the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = P\{(-\infty, x]\} \quad \text{for all } x \in \mathbb{R}.$$

Note. Since

$$(-\infty, x] = \bigcap_{n=1}^{\infty} \left(-\infty, x + \tfrac{1}{n}\right)$$

is the countable intersection of open sets, we see that $(-\infty, x] \in \mathcal{B}$ for every $x \in \mathbb{R}$. Thus, $P\{(-\infty, x]\}$ makes sense.

Question. We see that each different probability P gives rise to a distribution function F (i.e., $P \rightsquigarrow F$). Is the converse true? That is, does P uniquely characterize F and vice-versa?

The answer is yes—knowledge of F uniquely determines P as the following theorem shows.

Theorem. *The distribution function F induced by P characterizes P.*

Proof. (Sketch) Suppose that we know $F(x)$ for every $x \in \mathbb{R}$. We then know

$$P\{(-\infty, x]\}$$

for every $x \in \mathbb{R}$. In particular, we know

$$P\{(-\infty, a]\}$$

for every $a \in \mathbb{Q}$. Since

$$\{(-\infty, a] : a \in \mathbb{Q}\}$$

generates $\mathcal{B}(\mathbb{R})$, we know $P\{B\}$ for every $B \in \mathcal{B}$. □

Question. Can we tell which functions $F : \mathbb{R} \to [0, 1]$ are distribution functions?

Theorem. *A function $F : \mathbb{R} \to [0, 1]$ is the distribution function of a unique probability measure P on $(\mathbb{R}, \mathcal{B})$ if and only if*

(i) F is non-decreasing, i.e., $x < y$ implies $F(x) \leq F(y)$,

(ii) F is right-continuous, i.e., $F(y) \to F(x)$ as $y \to x+$,

$$\lim_{y \to x+} F(y) = \lim_{y \downarrow x} F(y) = F(x),$$

(iii) $F(x) \to 1$ as $x \to \infty$, i.e.,

$$\lim_{x \to \infty} F(x) = 1,$$

(iv) $F(x) \to 0$ as $x \to -\infty$, i.e.,

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

Example. Suppose that $\alpha \in \mathbb{R}$ and let

$$F(x) = \begin{cases} 1, & \text{if } x \geq \alpha, \\ 0, & \text{if } x < \alpha. \end{cases}$$

Since F satisfies all the conditions of the theorem, F must be a distribution function. The corresponding probability measure is

$$P\{B\} = \begin{cases} 1, & \text{if } \alpha \in B, \\ 0, & \text{if } \alpha \notin B, \end{cases}$$

for every $B \in \mathcal{B}$. We call P the *Dirac point mass at α*.

Lecture #12: Random Variables

Reference. Chapter 8 pages 47–50

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and let $X : \Omega \to \mathbb{R}$ given by $\omega \mapsto X(\omega)$ be a function.

Definition. We say that the function $X : \Omega \to \mathbb{R}$ is a *random variable* (or a *measurable function*) if $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$ where $\mathcal{B}$ denotes the Borel sets of $\mathbb{R}$.

Remark. We use the phrases random variable and measurable function synonymously. "Random variable" is preferred by probabilists while "measurable function" is preferred by analysts.

Note. The motivation for the definition of random variable is the following. We want to be able to compute

$$P\{\omega : X(\omega) \in B\} = P\{X \in B\} = P\{X^{-1}(B)\}$$

for every $B \in \mathcal{B}$. Therefore, we must have $X^{-1}(B) \in \mathcal{A}$ in order for $P\{X^{-1}(B)\}$ to make sense.

Remark. If Ω is either finite or countable, then every function $X : (\Omega, 2^{\Omega}) \to (\mathbb{R}, \mathcal{B})$ is measurable. If Ω is uncountable, then the trouble begins!

Example. Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and let $H \notin \mathcal{A}$ so that H is not an event. We saw an example of such a non-measurable H in Lecture #9. Define the function $X : \Omega \to \mathbb{R}$ by setting

$$X(\omega) = 1_H\{\omega\} = \begin{cases} 1, & \text{if } \omega \in H, \\ 0, & \text{if } \omega \notin H. \end{cases}$$

To prove that the function X is not a random variable, we must show that there exists a Borel set $B \in \mathcal{B}$ for which $X^{-1}(B) \notin \mathcal{A}$. Let $B = \{1\}$ which is clearly a closed set and therefore Borel. We find

$$X^{-1}(\{1\}) = \{\omega : X(\omega) = 1\} = \{\omega : \omega \in H\} = H.$$

However, by assumption, $H \notin \mathcal{A}$ so that $X^{-1}(\{1\}) \notin \mathcal{A}$. Thus, X is not a random variable.

Remark. Note that in order for the previous example to work we must have $\mathcal{A} \subsetneq 2^{\Omega}$. In the case of a discrete or countable Ω our standard choice of $\mathcal{A} = 2^{\Omega}$ makes such an example impossible. It is, however, possible to construct a probability space on a discrete space Ω in such a way that $\mathcal{A} \neq 2^{\Omega}$. For instance, let $\Omega = \{a, b, c, d\}$ and take $\mathcal{A} = \{\emptyset, \{a, b\}, \{c, d\}, \Omega\}$. Set $P\{a, b\} = P\{c, d\} = 1/2$. If $H = \{a\}$, then H is not an event and $X(\omega) = 1_H\{\omega\}$ is not a random variable. However, this example is both silly and contrived!

Definition. If $X : \Omega \to \mathbb{R}$ is a random variable, then the *law* (or *distribution*) of X is the probability measure on $(\mathbb{R}, \mathcal{B})$ given by

$$P^X\{B\} = P\{\omega : X(\omega) \in B\} = P\{X \in B\} = P\{X^{-1}(B)\}$$

for every $B \in \mathcal{B}$.

Although we have referred to the fact that P^X is a probability on $(\mathbb{R}, \mathcal{B})$, we have not yet proved it!

Theorem. *The law of a random variable $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ is a probability measure on $(\mathbb{R}, \mathcal{B})$.*

Proof. In order to prove this result, we must verify the conditions of Definition 2.3 are met. We begin by noting that $\mathcal{B}$ is, by definition, a σ-algebra on $\mathbb{R}$. Since $\emptyset \in \mathcal{B}$, we have

$$P^X\{\emptyset\} = P\{\omega : X(\omega) \in \emptyset\} = P\{\emptyset\} = 0$$

using the fact that $\emptyset \in \mathcal{A}$ and P is a probability on $(\Omega, \mathcal{A})$. Suppose that $B_1, B_2, \ldots \in \mathcal{B}$ is a sequence of pairwise, disjoint sets. Therefore,

$$P^X\left\{\bigcup_{i=1}^{\infty} B_i\right\} = P\left\{\omega : X(\omega) \in \bigcup_{i=1}^{\infty} B_i\right\} = P\left\{\omega : \omega \in \bigcup_{i=1}^{\infty} X^{-1}(B_i)\right\} = P\left\{\omega : \omega \in \bigcup_{i=1}^{\infty} A_i\right\}$$

where $A_i = X^{-1}(B_i)$. Since X is a random variable, we know that $A_i \in \mathcal{A}$ for each $i = 1, 2, \ldots$. Moreover, $A_1, A_2, \ldots$ are pairwise disjoint. Thus, using the fact that $A_1, A_2, \ldots \in \mathcal{A}$ are pairwise disjoint and P is a probability on $(\Omega, \mathcal{A})$ gives

$$P^X\left\{\bigcup_{i=1}^{\infty} B_i\right\} = P\left\{\bigcup_{i=1}^{\infty} A_i\right\} = \sum_{i=1}^{\infty} P\{A_i\} = \sum_{i=1}^{\infty} P\{X^{-1}(B_i)\} = \sum_{i=1}^{\infty} P^X\{B_i\}$$

as required. □

Since P^X is a probability measure on $(\mathbb{R}, \mathcal{B})$ we can discuss the corresponding distribution function.

Definition. If $X : \Omega \to \mathbb{R}$ is a random variable, then the *distribution function* of X, written $F_X : \mathbb{R} \to [0, 1]$, is defined by

$$F_X(x) = P^X\{(-\infty, x]\} = P\{\omega : X(\omega) \leq x\} = P\{X \leq x\}$$

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

Remark. By Theorem 7.1, distribution functions characterize probabilities on $\mathbb{R}$. That is, knowledge of one gives knowledge of the other:

$$X \text{ (random variable)} \leftrightsquigarrow P^X \text{ (law of } X\text{)} \leftrightsquigarrow F_X \text{ (distribution function of } X\text{)}$$

Remark. In Jacod and Protter, things are proved in a bit more generality. A function $X : (E, \mathcal{E}) \to (F, \mathcal{F})$ is called *measurable* if $X^{-1}(B) \in \mathcal{E}$ for every $B \in \mathcal{F}$. A space E and a σ-algebra $\mathcal{E}$ of subsets of E together are often called a *measurable space*. Technically, it is only when we introduce probability measures onto $(E, \mathcal{E})$ do measurable functions become random variables.

Summary of Important Objects

At this point, we have introduced all of the objects that are most important to modern probability. It is worth memorizing each of the definitions.

- Ω, the sample space of outcomes $\omega \in \Omega$,
- $\mathcal{A}$, a σ-algebra of subsets of Ω (Definition 2.1),
- P, a set function from $\mathcal{A}$ to $[0, 1]$ (Definition 2.3),
- $(\Omega, \mathcal{A})$ is called a measurable space and $(\Omega, \mathcal{A}, P)$ is called a probability space,
- the real numbers $\mathbb{R}$ together with the Borel sets $\mathcal{B}$ is a measurable space (Theorem 2.1),
- $X : \Omega \to R$ is a random variable (Definition 8.1),
- P^X is the law of X (page 50),
- F_X is the distribution function of X (Definition 7.1),
- $(\mathbb{R}, \mathcal{B}, P^X)$ is a probability space (Theorem 8.5).

It is important to know when a function of a random variable is again a random variable. The following theorem gives the answer.

Theorem. *Let $X : \Omega \to \mathbb{R}$ be a random variable and let $f : \mathbb{R} \to \mathbb{R}$ be measurable. If the function $f \circ X : \Omega \to \mathbb{R}$ is defined by*

$$f \circ X(\omega) = f(X(\omega)),$$

then $f \circ X$ is a random variable.

Proof. To show that $f \circ X$ is a random variable, we must show that

$$(f \circ X)^{-1}(B) \in \mathcal{A}$$

for every $B \in \mathcal{B}$. Since $f : \mathbb{R} \to \mathbb{R}$ is measurable, we know that

$$f^{-1}(B) \in \mathcal{B}$$

for every $B \in \mathcal{B}$. Since $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is a random variable, we know that

$$X^{-1}(B) \in \mathcal{A}$$

for every $B \in \mathcal{B}$. Since $f^{-1}(B) \in \mathcal{B}$, it must be the case that $X^{-1}(f^{-1}(B)) \in \mathcal{A}$. Since

$$(f \circ X)^{-1}(B) = X^{-1}(f^{-1}(B))$$

the proof is complete. □

The next theorem extends the example we considered at the beginning of this lecture.

Theorem. *Suppose that* $(\Omega, \mathcal{A}, P)$ *is a probability space and* $A \subseteq \Omega$*. The function* $1_A : \Omega \to \mathbb{R}$ *given by*

$$1_A\{\omega\} = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A, \end{cases}$$

is a random variable if and only if $A \in \mathcal{A}$*.*

Proof. For $B \subseteq \mathbb{R}$, we find

$$(1_A)^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0 \notin B,\ 1 \notin B, \\ A, & \text{if } 0 \notin B,\ 1 \in B, \\ A^c, & \text{if } 0 \in B,\ 1 \notin B, \\ \Omega, & \text{if } 0 \in B,\ 1 \in B. \end{cases}$$

Thus, $(1_A)^{-1}(B) \in \mathcal{A}$ if and only if $B \in \mathcal{B}$. □

Lecture #13: Random Variables

Reference. Chapter 8 pages 47–50

Theorem. *Suppose that $X : \Omega \to \mathbb{R}$ is a function. X is a random variable if and only if $X^{-1}(O) \in \mathcal{A}$ for every open set $O \in \mathcal{O}$.*

Proof. We being by recalling that the Borel sets are the σ-algebra generated by the open sets. That is, $\mathcal{B} = \sigma(\mathcal{O})$ where $\mathcal{O} = \{\text{open sets}\}$. If X is a random variable, then since an open set is necessarily a Borel set we have the forward implication $X^{-1}(O) \in \mathcal{A}$ for every open set $O \in \mathcal{O}$. To show that converse, suppose that $X^{-1}(O) \in \mathcal{A}$ for every open set $O \in \mathcal{O}$. We must now show that $X^{-1}(B) \in \mathcal{A}$ for every Borel set $B \in \mathcal{B}$. We leave it as an exercise to verify that that the following set relations hold:

$$X^{-1}\left(\bigcup_n B_n\right) = \bigcup_n X^{-1}(B_n), \quad X^{-1}\left(\bigcap_n B_n\right) = \bigcap_n X^{-1}(B_n), \quad \text{and} \quad X^{-1}(B^c) = \left[X^{-1}(B)\right]^c.$$

Our strategy for the proof is the following. Let $\mathcal{F} = \{B \in \mathcal{B} : X^{-1}(B) \in \mathcal{A}\}$. We will show that $\mathcal{F} = \mathcal{B}$. On the one hand, $\mathcal{F} \subseteq \mathcal{B}$ by definition. On the other hand, since X^{-1} commutes with unions, intersections, and complements, we see that $\mathcal{F}$ is a σ-algebra. By assumption $\mathcal{O} \subseteq \mathcal{F}$ which implies that $\sigma(\mathcal{O}) \subseteq \mathcal{F}$ since $\mathcal{F}$ is a σ-algebra. Since $\sigma(\mathcal{O}) = \mathcal{B}$ we conclude $\mathcal{B} \subseteq \mathcal{F}$. Thus, we have shown

$$\mathcal{B} \subseteq \mathcal{F} \subseteq \mathcal{B}$$

which implies that $\mathcal{B} = \mathcal{F}$. In other words,

$$X^{-1}(\mathcal{B}) \subseteq \sigma(X^{-1}(\mathcal{O})) \subseteq \mathcal{A}$$

meaning that $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$. □

Theorem. *Suppose that $X : \Omega \to \mathbb{R}$ and $X_n : \Omega \to \mathbb{R}$, $n = 1, 2, \ldots$ are functions.*

(a) X is a random variable if and only if $\{X \leq a\} = X^{-1}((-\infty, a]) \in \mathcal{A}$ for every $a \in \mathbb{R}$ if and only if $\{X < a\} \in \mathcal{A}$ for every $a \in \mathbb{R}$.

(b) If X_n, $n = 1, 2, \ldots$, are each random variables, then

$$\sup_n X_n, \quad \inf_n X_n, \quad \limsup_{n\to\infty} X_n, \quad \liminf_{n\to\infty} X_n$$

are also random variables.

(c) If X_n, $n = 1, 2, \ldots$, are each random variables and $X_n \to X$ pointwise, then X is a random variable.

Proof. (a) By Theorem 2.1, we know that

$$\sigma(\{(-\infty, a] : a \in \mathbb{R}\}) = \mathcal{B}$$

and so the result follows from the previous theorem. Recall that we write

$$(-\infty, a) = \bigcup_{n=1}^{\infty} \left(-\infty, a - \tfrac{1}{n}\right].$$

(b) Since X_n is a random variable, we see that $\{X_n \leq a\} = \{X_n(\omega) \in (-\infty, a]\} \in \mathcal{A}$ for each n. By definition,

$$\left\{\sup_n X_n \leq a\right\} = \bigcap_n \{X_n \leq a\} \in \mathcal{A}$$

and

$$\left\{\inf_n X_n < a\right\} = \bigcup_n \{X_n < a\} \in \mathcal{A}$$

so that both of these are random variables by (a). Since

$$\limsup_{n\to\infty} X_n = \inf_n \sup_{m \geq n} X_m,$$

if we define

$$Y_n = \sup_{m \geq n} X_m,$$

then Y_n is a random variable so that

$$\inf_n Y_n$$

is also a random variable. The proof that

$$\liminf_{n\to\infty} X_n = \sup_n \inf_{m \geq n} X_m$$

is a random variable is similar.

(c) If $X_n \to X$ pointwise, then

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

is a random variable by (b). □

Density Functions

Reference. Chapter 7 pages 42–44

Suppose that $f : \mathbb{R} \to [0, \infty)$ is non-negative and Riemann-integrable with

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1.$$

Define the function $F : \mathbb{R} \to [0, 1]$ by setting

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y.$$

It then follows that F is non-decreasing, F is right-continuous, $F(x) \to 1$ as $x \to \infty$, and $F(x) \to 0$ as $x \to -\infty$. In other words, F is a distribution function. (Check these facts!) We call f the *density (function) associated to* F.

Remark. It is not true that every distribution admits a density. For instance, there is no density associated with the Dirac point mass at α. However, many "famous" distributions admit densities.

Example. The uniform distribution on (a, b) has density

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

Example. The exponential distribution with parameter $\beta > 0$ has density

$$f(x) = \begin{cases} \beta e^{-\beta x}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Example. The normal distribution with parameters $\sigma > 0$, $\mu \in \mathbb{R}$, has density

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

Note. See pages 43–44 for other examples of densities.

Remark. Suppose $X : \Omega \to \mathbb{R}$ is a random variable. As we have already seen, P^X, the law of X, is a probability measure on $(\mathbb{R}, \mathcal{B})$. Hence, we say that X has density function f_X if f_X is the density associated with F_X, the distribution function of X.

Example. Consider the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = [0, 1]$, $\mathcal{A} = \mathcal{B}([0, 1])$, and P is the uniform probability on $[0, 1]$. Let $X : \Omega \to \mathbb{R}$ be defined by

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in \mathbb{Q} \cap [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Note that X is the indicator function of the rational numbers in $[0,1]$. As we have already seen, X is a random variable if and only if $\mathbb{Q} \cap [0.1] \in \mathcal{A}$. Is it? We write $\mathbb{Q}_1 = \mathbb{Q} \cap [0,1]$ so that

$$\mathbb{Q}_1 = \bigcap_{\omega \in \mathbb{Q}_1} \{\omega\}$$

expresses $\mathbb{Q}_1$ as a countable union of disjoint closed sets. Thus, $\mathbb{Q}_1$ is a Borel set so that $\mathbb{Q}_1 \in \mathcal{A}$ is measurable and $X = 1_{\mathbb{Q}_1}$ is a random variable. Let the function $Y : \Omega \to \mathbb{R}$ be defined by $Y(\omega) = \omega X(\omega)$. Is Y a random variable?

Here are some other questions to keep us thinking.

- What "type" of random variable is X? What about Y?
- What is P^X? What is P^Y?
- Does P^X have a distribution function F_X? What about P^Y?
- Do either F_X or F_Y admit a density?
- What is $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(Y)$, or $\mathbb{E}(Y^2)$?

Statistics 851 (Winter 2008) February 6, 2008
Prof. Michael Kozdron

Lecture #14: Some General Function Theory

Suppose that $f : \mathbf{X} \to \mathbf{Y}$ is a function. We are implicitly assuming that f is defined for all $x \in \mathbf{X}$. We call $\mathbf{X}$ the *domain* of f and call $\mathbf{Y}$ the *codomain* of f.

The *range* of f is the set

$$f(\mathbf{X}) = \{y \in \mathbf{Y} : f(x) = y \text{ for some } x \in \mathbf{X}\}.$$

Note that $f(\mathbf{X}) \subseteq \mathbf{Y}$. If $f(\mathbf{X}) = \mathbf{Y}$, then we say that f is *onto* $\mathbf{Y}$.

Let $B \subseteq \mathbf{Y}$. We define $f^{-1}(B)$ by

$$f^{-1}(B) = \{x \in \mathbf{X} : f(x) = y \text{ for some } y \in B\} = \{f \in B\} = \{x : f(x) \in B\}.$$

We call $\mathbf{X}$ a *topological space* if there is a notion of open subsets of $\mathbf{X}$. The Borel σ-algebra on $\mathbf{X}$, written $\mathcal{B}(\mathbf{X})$, is the σ-algebra generated by the open sets of $\mathbf{X}$.

Let $\mathbf{X}$ and $\mathbf{Y}$ be topological spaces. A function $f : \mathbf{X} \to \mathbf{Y}$ is called *continuous* if for every open set $V \subseteq \mathbf{Y}$, the set $U = f^{-1}(V) \subseteq \mathbf{X}$ is open.

A function $f : \mathbf{X} \to \mathbf{Y}$ is called measurable if for every measurable set $B \in \mathcal{B}(\mathbf{Y})$, the set $f^{-1}(B) \in \mathcal{B}(\mathbf{X})$ is a measurable set.

The proof of the first theorem proved in Lecture #13 can be generalized without any difficulty to yield the following result.

Theorem. *Suppose that $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ and $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$ are topological measure spaces. The function $f : \mathbf{X} \to \mathbf{Y}$ is measurable if and only if $f^{-1}(O) \in \mathcal{B}(\mathbf{X})$ for every open set $O \in \mathcal{B}(\mathbf{Y})$.*

The next theorem tells us that continuous functions are necessarily measurable functions.

Theorem. *Suppose that $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$ and $(\mathbf{Y}, \mathcal{B}(\mathbf{Y}))$ are topological measure spaces. If $f : \mathbf{X} \to \mathbf{Y}$ is continuous, then f is measurable.*

Proof. By definition, $f : \mathbf{X} \to \mathbf{Y}$ is continuous if and only if $f^{-1}(O) \subseteq \mathbf{X}$ is an open set for every open set $O \subseteq \mathbf{Y}$. Since an open set is necessarily a Borel set, we conclude that $f^{-1}(O) \in \mathcal{B}(\mathbf{X})$ for every open set $O \in \mathcal{B}(\mathbf{Y})$. However, it now follows immediately from the previous theorem that $f : \mathbf{X} \to \mathbf{Y}$ is measurable. □

The following theorem is a generalization of a theorem proved in Lecture #12 and shows that the composition of measurable functions is measurable.

Theorem. *Suppose that $(\mathbf{W}, \mathcal{F})$, $(\mathbf{X}, \mathcal{G})$, and $(\mathbf{Y}, \mathcal{H})$ are measurable spaces, and let $f : (\mathbf{W}, \mathcal{F}) \to (\mathbf{X}, \mathcal{G})$ and $g : (\mathbf{X}, \mathcal{G}) \to (\mathbf{Y}, \mathcal{H})$ be measurable. Then the function $h = g \circ f$ is a measurable function from $(\mathbf{W}, \mathcal{F})$ to $(\mathbf{Y}, \mathcal{H})$.*

Proof. Suppose that $H \in \mathcal{H}$. Since g is measurable, we have $g^{-1}(H) \in \mathcal{G}$. Since f is measurable, we have $f^{-1}(g^{-1}(H)) \in \mathcal{F}$. Since

$$h^{-1}(H) = (g \circ f)^{-1}(H) = f^{-1}(g^{-1}(H)) \in \mathcal{F}$$

the proof is complete. □

Recall that from Theorem 2.1 the Borel sets of $\mathbb{R}$ are generated by intervals of the form $(-\infty, a]$ for $a \in \mathbb{Q}$. Exactly the same proof shows that the Borel sets of $\mathbb{R}^n$, written $\mathcal{B}^n = \mathcal{B}(\mathbb{R}^n)$ are generated by quadrants of the form

$$\prod_{i=1}^{n}(-\infty, a_i], \quad a_1, \ldots, a_n \in \mathbb{Q}.$$

Theorem. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and suppose that $X_i : \Omega \to \mathbb{R}$ is a random variable for each $i = 1, \ldots, n$. Suppose further that $f : (\mathbb{R}^n, \mathcal{B}^n) \to (\mathbb{R}, \mathcal{B})$ is measurable. Then the function $f(X_1, \ldots, X_n) : \Omega \to \mathbb{R}$ is measurable.*

Proof. Define the (vector-valued) function $X : \Omega \to \mathbb{R}^n$ by setting $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))$. Since

$$X^{-1}\left(\prod_{i=1}^{n}(-\infty, a_i]\right) = \bigcap_{i=1}^{n} X_i^{-1}((-\infty, a_i]) = \bigcap_{i=1}^{n}\{X_i \leq a_i\} \in \mathcal{A}$$

we conclude that $X : \Omega \to \mathbb{R}^n$ is measurable. We often call X a *random vector*. Since $X : \Omega \to \mathbb{R}^n$ is measurable and $f : \mathbb{R}^n \to \mathbb{R}$ is measurable, we immediately conclude from the previous theorem $f \circ X = f(X_1, \ldots, X_n) : \Omega \to \mathbb{R}$ is measurable. □

Corollary. *If X and Y are random variables, then so too are $X + Y$, XY, X/Y for $Y \neq 0$, $X \wedge Y$, and $X \vee Y$.*

Expectation

Reference. Chapter 9 pages 51–60

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and let $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ be a random variable. Recall that this means that $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$. Our goal in this lecture is the following.

Goal. To define $\mathbb{E}(X)$ for every random variable X.

Definition. A random variable $X : \Omega \to \mathbb{R}$ is called *simple* if it can be written as

$$X = \sum_{i=1}^{n} a_i 1_{A_i}$$

where $a_i \in \mathbb{R}$, $A_i \in \mathcal{A}$ for $i = 1, 2, \ldots, n$.

Note that if a random variable is simple, then it takes on a finite number of values. As the next proposition shows, the converse is also true.

Proposition. *If $X : \Omega \to \mathbb{R}$ is a random variable that takes on finitely many values, then X is simple.*

Proof. Suppose X is a random variable and takes on finitely many values labelled $a_1, \ldots, a_n$. Let

$$A_i = \{X = a_i\} = \{\omega : X(\omega) = a_i\}, \quad i = 1, \ldots, n.$$

Since X is a random variable, $A_i = X^{-1}(\{a_i\})$, and $\{a_i\}$ is a Borel set, we conclude that $A_i \in \mathcal{A}$ for each $i = 1, \ldots, n$. Furthermore, we can represent X as

$$X = \sum_{i=1}^{n} a_i 1_{A_i}$$

which proves that X is simple. □

Thus, saying that X is a simple random variable means that the range of X is finite.

Remark. Recall that the function $X = 1_A$ is measurable if and only if $A \in \mathcal{A}$. Also recall that the composition and sum of measurable functions is again measurable. Therefore, we see that

$$X = \sum_{i=1}^{n} a_i 1_{A_i}$$

is measurable if and only if $A_i \in \mathcal{A}$ for each $i = 1, \ldots, n$.

Definition. If X is a simple random variable, then the *expectation* (or *expected value*) of X is defined to be

$$\mathbb{E}(X) = \sum_{i=1}^{n} a_i P\{A_i\}.$$

Remark. By definition, a simple random variable is necessarily discrete. However, the converse is not true. For instance, if Y is a Poisson(λ) random variable, then

$$P\{Y = k\} > 0 \quad \text{for every } k = 0, 1, 2, \ldots$$

so that Y is a discrete random variable which is not simple.

Example. Consider the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = [0, 1]$, $\mathcal{A} = \mathcal{B}(\Omega)$ are the Borel sets in $[0, 1]$, and P is the uniform probability measure on $[0, 1]$. Let $\mathbb{Q}_1 = [0, 1] \cap \mathbb{Q}$ and consider the random variable $X = 1_{\mathbb{Q}_1}$ which is given by

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in [0, 1] \cap \mathbb{Q}, \\ 0, & \text{if } \omega \notin [0, 1] \cap \mathbb{Q}. \end{cases}$$

Since the only two possible values of X are 0 and 1, we conclude that X is simple. Therefore, if we let $a_1 = 1$ and $a_2 = 0$ so that $A_1 = \mathbb{Q}_1$ and $A_2 = [0, 1] \setminus \mathbb{Q}_1$, then

$$\mathbb{E}(X) = a_1 P\{A_1\} + a_2 P\{A_2\} = P\{\mathbb{Q}_1\} = 0$$

since $P\{\mathbb{Q}_1\} = 0$.

A slight extension of this example is given by the following exercise.

Exercise. Suppose that $(\Omega, \mathcal{A}, P)$ is the same probability space as in the previous example. Let $A \subset \Omega$ be countable and define the function $X : \Omega \to \mathbb{R}$ by setting $X(\omega) = 1_A(\omega)$. Prove that $A \in \mathcal{A}$ with $P\{A\} = 0$ so that X is a random variable with $\mathbb{E}(X) = 0$.

Example. Suppose that $(\Omega, \mathcal{A}, P)$ is the same probability space and X is the same random variable as in the previous example. Define the function $Y : \Omega \to \mathbb{R}$ by setting $Y(\omega) = 0$ for all $\omega \in \Omega$. Therefore, Y is a simple random variable with $\mathbb{E}(Y) = 0$. Notice that both $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ equal 0. However, $X \neq Y$ since, for instance, $X(1/2) = 1$ but $Y(1/2) = 0$. The set where X and Y differ is

$$\{X \neq Y\} = \{\omega : X(\omega) \neq Y(\omega)\} = \{\omega : X(\omega) = 1, Y(\omega) = 0\} = \{\omega : \omega \in [0,1] \cap \mathbb{Q}\}.$$

However, $P\{\omega : \omega \in [0,1] \cap \mathbb{Q}\} = 0$ so that

$$P\{X \neq Y\} = 0.$$

In other words, $\{X \neq Y\}$ is a null set so that $X = Y$ almost surely. If we now define the function $Z : \Omega \to \mathbb{R}$ by setting

$$Z(\omega) = \begin{cases} 1, & \text{if } \omega \in [0, 1/2], \\ -1, & \text{if } \omega \notin [0, 1/2], \end{cases}$$

then Z is a simple random variable with $\mathbb{E}(Z) = 0$. However, $P\{Z \neq Y\} = 1$ since $\{Z \neq Y\} = \Omega$ and $P\{Z \neq X\} = 1$ since $P\{Z = X\} = P\{[0, 1/2] \cap \mathbb{Q}\} \leq P\{[0,1] \cap \mathbb{Q}\} = 0$. Therefore, neither $\{X \neq Z\}$ nor $\{Y \neq Z\}$ is a null set.

Lecture #15: Expectation of a Simple Random Variable

Reference. Chapter 9 pages 51–60

Recall that a simple random variable is one that takes on finitely many values.

Definition. A random variable $X : \Omega \to \mathbb{R}$ is called *simple* if it can be written as

$$X = \sum_{i=1}^{n} a_i 1_{A_i}$$

where $a_i \in \mathbb{R}$, $A_i \in \mathcal{A}$ for $i = 1, 2, \ldots, n$.

Example. Consider the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = [0, 1]$, $\mathcal{A} = \mathcal{B}(\Omega)$, and P is the uniform probability on Ω. Suppose that the random variable $X : \Omega \to \mathbb{R}$ is defined by

$$X(\omega) = \sum_{i=1}^{4} a_i 1_{A_i}(\omega)$$

where $a_1 = 4$, $a_2 = 2$, $a_3 = 1$, $a_4 = -1$, and

$$A_1 = [0, \tfrac{1}{2}), \quad A_2 = [\tfrac{1}{4}, \tfrac{3}{4}), \quad A_3 = (\tfrac{1}{2}, \tfrac{7}{8}], \quad A_4 = (\tfrac{7}{8}, 1].$$

Show that there exist finitely many real constants $c_1, \ldots, c_n$ and *disjoint* sets $C_1, \ldots, C_n \in \mathcal{A}$ such that

$$X = \sum_{i=1}^{n} c_i 1_{C_i}.$$

Solution. We find

$$X(\omega) = \begin{cases} 4, & \text{if } 0 \leq \omega < 1/4, \\ 6, & \text{if } 1/4 \leq \omega < 1/2, \\ 2, & \text{if } \omega = 1/2, \\ 3, & \text{if } 1/2 < \omega < 3/4, \\ 1, & \text{if } 3/4 \leq \omega < 7/8, \\ 0, & \text{if } \omega = 7/8, \\ -1, & \text{if } 7/8 < \omega \leq 1, \end{cases}$$

so that

$$X = \sum_{i=1}^{7} c_i 1_{C_i}$$

where $c_1 = 4$, $c_2 = 6$, $c_3 = 2$, $c_4 = 3$, $c_5 = 1$, $c_6 = 0$, $c_7 = -1$ and

$$C_1 = [0, \tfrac{1}{4}), \quad C_2 = [\tfrac{1}{4}, \tfrac{1}{2}), \quad C_3 = \{\tfrac{1}{2}\}, \quad C_4 = (\tfrac{1}{2}, \tfrac{3}{4}), \quad C_5 = [\tfrac{3}{4}, \tfrac{7}{8}), \quad C_6 = \{\tfrac{7}{8}\}, \quad C_7 = (\tfrac{7}{8}, 1].$$

Proposition. *If X and Y are simple random variables, then*

$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y)$$

for every α, $\beta \in \mathbb{R}$.

Proof. Suppose that X and Y are simple random variables with

$$X = \sum_{i=1}^{n} a_i 1_{A_i} \quad \text{and} \quad Y = \sum_{j=1}^{m} b_j 1_{B_j}$$

where $A_1, \ldots, A_n \in \mathcal{A}$ and $B_1, \ldots, B_m \in \mathcal{A}$ each partition Ω. Since

$$\alpha X = \alpha \sum_{i=1}^{n} a_i 1_{A_i} = \sum_{i=1}^{n} (\alpha a_i) 1_{A_i}$$

we conclude by definition that

$$\mathbb{E}(\alpha X) = \sum_{i=1}^{n} (\alpha a_i) P\{A_i\} = \alpha \sum_{i=1}^{n} a_i P\{A_i\} = \alpha \mathbb{E}(X).$$

The proof of the theorem will be completed by showing $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. Notice that

$$\{A_i \cap B_j : 1 \leq i \leq n, 1 \leq j \leq m\}$$

consists of pairwise disjoint events whose union is Ω and

$$X + Y = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) 1_{A_i \cap B_j}.$$

Therefore, by definition,

$$\begin{aligned}
\mathbb{E}(X + Y) &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) P\{A_i \cap B_j\} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i P\{A_i \cap B_j\} + \sum_{i=1}^{n} \sum_{j=1}^{m} b_j P\{A_i \cap B_j\} \\
&= \sum_{i=1}^{n} a_i P\{A_i\} + \sum_{j=1}^{m} b_j P\{B_j\}
\end{aligned}$$

and the proof is complete. □

Fact. If X and Y are simple random variables with $X \leq Y$, then

$$\mathbb{E}(X) \leq \mathbb{E}(Y).$$

Exercise. Prove the previous fact.

Goal. To construct $\mathbb{E}(X)$ for any random variable.

We have already defined $\mathbb{E}(X)$ for simple random variables.

Suppose that X is a *positive random variable*. That is, $X(\omega) \geq 0$ for all $\omega \in \Omega$. (We will need to allow $X(\omega) \in [0, +\infty]$ for some consistency.)

Definition. If X is a positive random variable, define the *expectation* of X to be

$$\mathbb{E}(X) = \sup\{\mathbb{E}(Y) : Y \text{ is simple and } 0 \leq Y \leq X\}.$$

That is, we approximate positive random variables by simple random variables. Of course, this leads to the question of whether or not this is possible.

Fact. For every random variable $X \geq 0$, there exists a sequence (X_n) of positive, simple random variables with $X_n \uparrow X$ (that is, X_n increases to X).

An example of such a sequence is given by

$$X_n(\omega) = \begin{cases} \frac{k}{2^n}, & \text{if } \frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n} \text{ and } 0 \leq k \leq n2^n - 1, \\ n, & \text{if } X(\omega) \geq n. \end{cases}$$

(Draw a picture.)

Fact. If $X \geq 0$ and (X_n) is a sequence of simple random variables with $X_n \uparrow X$, then $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$.

Now suppose that X is any random variable. Write

$$X^+ = \max\{X, 0\} \quad \text{and} \quad X^- = -\min\{X, 0\}$$

for the *positive part* and the *negative part* of X, respectively.

Note that $X^+ \geq 0$ and $X^- \geq 0$ so that the positive part and negative part of X are both positive random variables and

$$X = X^+ - X^- \quad \text{and} \quad |X| = X^+ + X^-.$$

Definition. A random variable X is called *integrable* (or has *finite expectation*) if both $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ are finite. In this case we define $\mathbb{E}(X)$ to be

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Definition. If one of $\mathbb{E}(X^+)$ or $\mathbb{E}(X^-)$ is infinite, then we can still define $\mathbb{E}(X)$ by setting

$$\mathbb{E}(X) = \begin{cases} +\infty, & \text{if } \mathbb{E}(X^+) = +\infty \text{ and } \mathbb{E}(X^-) < \infty, \\ -\infty, & \text{if } \mathbb{E}(X^+) < \infty \text{ and } \mathbb{E}(X^-) = +\infty. \end{cases}$$

However, X is not integrable in this case.

Definition. If both $\mathbb{E}(X^+) = +\infty$ and $\mathbb{E}(X^-) = +\infty$, then $\mathbb{E}(X)$ does not exist.

Summary of Constructing $\mathbb{E}(X)$ for General Random Variables

The outline of how to construct $\mathbb{E}(X)$ is sometimes called the "standard machine" and follows these steps.

(1) Step 1: define $\mathbb{E}(X)$ for simple random variables,

(2) Step 2: define $\mathbb{E}(X)$ for positive random variables,

(3) Step 3: define $\mathbb{E}(X)$ for general random variables.

Of course, we still have to prove that all of the facts given above are true.

Lecture #16: Integration and Expectation

Reference. Chapter 9 pages 51–60

Definition. If X is a positive random variable, then we define

$$\mathbb{E}(X) = \sup\{\mathbb{E}(Y) : Y \text{ is simple and } 0 \leq Y \leq X\}.$$

If X is any random variable, then we can write $X = X^+ - X^-$ where both X^+ and X^- are positive random variables. Provided that both $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ are finite, we define $\mathbb{E}(X)$ to be

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$$

and we say that X is *integrable* (or has *finite expectation*).

Remark. If $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}, P^X)$ is a random variable, then we sometimes write

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

That is, the expectation of a random variable is the *Lebesgue integral* of X. We will say more about this later.

Definition. Let $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ be the set of real-valued random variables on $(\Omega, \mathcal{A}, P)$ with finite expectation. That is,

$$\mathcal{L}^1(\Omega, \mathcal{A}, P) = \{X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B}) : \ X \text{ is a random variable with } \mathbb{E}(X) < \infty\}.$$

We will often write $\mathcal{L}^1$ for $\mathcal{L}^1(\Omega, \mathcal{A}, P)$ and suppress the dependence on the underlying probability space.

Theorem. *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space, and let $X_1, X_2, \ldots, X$, and Y all be real-valued random variables on $(\Omega, \mathcal{A}, P)$.*

(a) $\mathcal{L}^1$ is a vector space and expectation is a linear map on $\mathcal{L}^1$. Furthermore, expectation is positive. That is, if X, $Y \in \mathcal{L}^1$ with $0 \leq X \leq Y$, then $0 \leq \mathbb{E}(X) \leq \mathbb{E}(Y)$.

(b) $X \in \mathcal{L}^1$ if and only if $|X| \in \mathcal{L}^1$, in which case we have

$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|).$$

(c) If $X = Y$ almost surely (i.e., if $P\{\omega : X(\omega) = Y(\omega)\} = 1$), then $\mathbb{E}(X) = \mathbb{E}(Y)$.

(d) (Monotone Convergence Theorem) *If the random variables $X_n \geq 0$ for all n and $X_n \uparrow X$ (i.e., $X_n \to X$ and $X_n \leq X_{n+1}$), then*

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}\left(\lim_{n\to\infty} X_n\right) = \mathbb{E}(X).$$

(We allow $\mathbb{E}(X) = +\infty$ if necessary.)

(e) (Fatou's Lemma) *If the random variables X_n all satisfy $X_n \geq Y$ almost surely for some $Y \in \mathcal{L}^1$ and for all n, then*

$$\mathbb{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathbb{E}(X_n). \qquad (*)$$

In particular, $(*)$ holds if $X_n \geq 0$ for all n.

(f) (Lebesgue's Dominated Convergence Theorem) *If the random variables $X_n \to X$, and if for some $Y \in \mathcal{L}^1$ we have $|X_n| \leq Y$ almost surely for all n, then $X_n \in \mathcal{L}^1$, $X \in \mathcal{L}^1$, and*

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

Remark. This theorem contains **ALL** of the central results of Lebesgue integration theory.

Theorem. *Let X_n be a sequence of random variables.*

(a) If $X_n \geq 0$ for all n, then

$$\mathbb{E}\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty} \mathbb{E}(X_n) \qquad (*)$$

with both sides simultaneously being either finite or infinite.

(b) If

$$\sum_{n=1}^{\infty} \mathbb{E}(X_n) < \infty,$$

then

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

converges almost surely to some random variable $Y \in \mathcal{L}^1$. In other words,

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

is integrable with

$$\mathbb{E}\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty} \mathbb{E}(X_n).$$

Thus, $(*)$ holds with both sides being finite.

Notation. Since $X = Y$ almost surely implies $\mathbb{E}(X) = \mathbb{E}(Y)$, we can consider "almost sure equality" as an equivalence relation; that is,

$$X \sim Y \text{ if } X = Y \text{ almost surely.}$$

Let L^1 denote $\mathcal{L}^1$ modulo almost sure equality.

Notation. Let

$$\mathcal{L}^p = \{\text{random variables } X : |X|^p \in \mathcal{L}^1\}, \quad 1 \le p < \infty,$$

and let L^p denote $\mathcal{L}^p$ modulo almost sure equality.

Theorem. *(a)* (Cauchy-Schwartz Inequality) *If X, $Y \in \mathcal{L}^2$, then $XY \in \mathcal{L}^1$ and*

$$|\mathbb{E}(XY)| \le \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

(b) $L^2 \subset L^1$ and $X \in L^2$ implies

$$[\mathbb{E}(X)]^2 \le \mathbb{E}(X^2).$$

(c) L^2 is a linear space. That is, if X, $Y \in L^2$ and α, $\beta \in \mathbb{R}$, then $\alpha X + \beta Y \in L^2$. In fact, as we will see in Chapter 22, L^2 is a Hilbert space.

Remark. Although L^2 is a Hilbert space, $\mathcal{L}^2$ is not. This is the reason that we consider random variables up to equality almost surely.

Theorem. *Let $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ be a random variable.*

(a) (Markov's Inequality) *If $X \in \mathcal{L}^1$, then*

$$P\{|X| \ge a\} \le \frac{\mathbb{E}(|X|)}{a}$$

for every $a > 0$.

(b) (Chebychev's Inequality) *If $X \in \mathcal{L}^2$, then*

$$P\{|X| \ge a\} \le \frac{\mathbb{E}(X^2)}{a^2}$$

for every $a > 0$.

Theorem (Expectation Rule)**.** *Let $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ be a random variable, and let $h : (\mathbb{R}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B})$ be measurable.*

(a) The random variable $h(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ if and only if $h \in \mathcal{L}^1(\mathbb{R}, \mathcal{B}, P^X)$.

(b) If $h \ge 0$, or if either condition in (a) holds, then

$$\begin{aligned}\mathbb{E}(h(X)) &= \int_\Omega h(X(\omega))\, P(\mathrm{d}\omega) \quad \textit{Lebesgue integral (definition)} \\ &= \int_{\mathbb{R}} h(x)\, P^X(\mathrm{d}x) \quad \textit{Riemann-Stieltjes integral (theorem).}\end{aligned}$$

(c) If X has a density f, and either $\mathbb{E}(|h(X)|) < \infty$ or $h \ge 0$, then

$$\mathbb{E}(h(X)) = \int_{\mathbb{R}} h(x) f(x)\, \mathrm{d}x \quad \textit{Riemann integral (theorem).}$$

Remark. Parts (b) and (c) explain how to perform a "change-of-variables."

Statistics 851 (Winter 2008)
Interchanging Limits and Integration

The following example shows that you cannot carelessly interchange limits!

Example. Suppose that $f_n(x) \to f(x)$. When does

$$\int f(x)\,\mathrm{d}x = \int \left(\lim_{n\to\infty} f_n(x)\right)\,\mathrm{d}x = \lim_{n\to\infty} \int f_n(x)\,\mathrm{d}x?$$

Let $f_n(x) = nx(1-x^2)^n$, $0 \le x \le 1$, for $n = 1, 2, \ldots$ so that

- $f_n(0) = 0$ for all n,
- $f_n(1) = 1$ for all n,
- $f_n(x) \to 0$ for all $0 < x < 1$.

That is,

$$\lim_{n\to\infty} f_n(x) = 0 \quad \text{for all } 0 \le x \le 1$$

and so

$$\int_0^1 \left(\lim_{n\to\infty} f_n(x)\right)\,\mathrm{d}x = \int_0^1 0\,\mathrm{d}x = 0$$

However,

$$\int_0^1 f_n(x)\,\mathrm{d}x = \int_0^1 nx(1-x^2)^n\,\mathrm{d}x = \frac{n}{2n+2}$$

so that

$$\lim_{n\to\infty} \int_0^1 f_n(x)\,\mathrm{d}x = \lim_{n\to\infty} \frac{n}{2n+2} = \frac{1}{2}.$$

Statistics 851 (Winter 2008) February 13, 2008
Prof. Michael Kozdron

Lecture #17: Comparison of Lebesgue and Riemann Integrals

Reference. Today's notes!

The purpose of today's lecture is to compare the Lebesgue integral with the Riemann integral.

In order to illustrate the differences between the two, we will not use probability notation, but instead use calculus notation.

Consider the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = [0,1]$, $\mathcal{A}$ denotes the Borel sets of $[0,1]$, and P denotes the uniform probability on $[0,1]$. Let $X : \Omega \to \mathbb{R}$ be a random variable and consider the expectation of X given by

$$\int_\Omega X(\omega)\, P(\mathrm{d}\omega). \qquad (*)$$

The uniform probability measure on $[0,1]$ is sometimes called *Lebesgue measure* and is denoted by μ. Instead of using $\omega \in [0,1]$ we will use $x \in [0,1]$, and instead of writing $X : [0,1] \to \mathbb{R}$ for a random variable we will write $h : [0,1] \to \mathbb{R}$ for a measurable function. Therefore, in our new language, $(*)$ is equivalent to

$$\int_{[0,1]} h(x)\, \mu(\mathrm{d}x) = \int_0^1 h(x)\, \mu(\mathrm{d}x)$$

which we call the *Lebesgue integral* of h.

By using this notation, we immediately see the analogy with

$$\int_0^1 h(x)\, \mathrm{d}x,$$

the *Riemann integral* of h.

If we translate back to the language of probability, the Riemann integral of the random variable X would be written as

$$\int_\Omega X(\omega)\, \mathrm{d}\omega.$$

Example. Suppose that $h(x) = 1_{\mathbb{Q}_1}(x)$ is the indicator function of the rational numbers in $[0,1]$. That is,

$$h(x) = \begin{cases} 1, & \text{if } x \in [0,1] \cap \mathbb{Q}, \\ 0, & \text{if } x \notin [0,1] \cap \mathbb{Q}. \end{cases}$$

Note that h is a *simple function*; that is, we can write

$$h(x) = \sum_{i=1}^{2} a_i 1_{A_i}(x)$$

where $a_1 = 1$, $a_2 = 0$, $A_1 = [0,1] \cap \mathbb{Q}$, and $A_2 = [0,1] \cap \mathbb{Q}^c$. By definition, the Lebesgue integral of h is

$$\int_0^1 h(x)\,\mu(\mathrm{d}x) = \sum_{i=1}^{2} a_i \mu(A_i) = 1 \cdot\ \mu(A_1) + 0 \cdot \mu(A_2) = \mu(A_1) = 0$$

since A_1 is a countable set.

Suppose that we now try to compute the Riemann integral

$$\int_0^1 h(x)\,\mathrm{d}x.$$

If h has an antiderivative, then we can use the Fundamental Theorem of Calculus. The trouble is that h is discontinuous everywhere and so h cannot possibly have an antiderivative. Therefore, in order to compute the Riemann integral we must resort to using the definition.

Definition. Suppose that $\mathcal{P} = \{0 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = 1\}$ is a partition of $[0,1]$, and let $\Delta x_i = x_{i+1} - x_i$ for $i = 0, 1, \ldots, N-1$. The function $h : [0,1] \to \mathbb{R}$ is *Riemann integrable* if

$$\lim_{N\to\infty} \sum_{i=0}^{N-1} h(x_i^*)\Delta x_i$$

exists for all choices of partitions $\mathcal{P}$ and $x_i^* \in [x_i, x_{i+1}]$. If so, we write

$$\int_0^1 h(x)\,\mathrm{d}x = \lim_{N\to\infty} \sum_{i=0}^{N-1} h(x_i^*)\Delta x_i$$

for the common value of this limit.

Example (continued). We will make our first attempt to compute

$$\int_0^1 h(x)\,\mathrm{d}x.$$

Let

$$\mathcal{P} = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\right\}$$

so that

$$\Delta x_i = \frac{1}{N}.$$

For $i = 0, 1, \ldots, N-1$, let

$$x_i^* = \frac{i}{N} + \frac{1}{2N}$$

and note that x_i^* is always rational. Therefore, by the definition of h we have

$$h(x_i^*) = 1$$

and so

$$\sum_{i=0}^{N-1} h(x_i^*)\Delta x_i = \sum_{i=0}^{N-1} 1 \cdot \frac{1}{N} = 1$$

which implies that

$$\lim_{N\to\infty} \sum_{i=0}^{N-1} h(x_i^*)\Delta x_i = 1.$$

Thus, we might be tempted to say

$$\int_0^1 h(x)\,\mathrm{d}x = 1. \qquad (**)$$

However, this is true only if the limit exists for *all* choices of partitions and x_i^*.

We now make our second attempt to compute the integral. For $i = 0, 1, \ldots, N-1$, let

$$x_i^* = \frac{i}{N} + \frac{1}{\pi N}$$

and note that x_i^* is always irrational. Therefore, by the definition of h we have

$$h(x_i^*) = 0$$

and so

$$\sum_{i=0}^{N-1} h(x_i^*)\Delta x_i = \sum_{i=0}^{N-1} 0 \cdot \frac{1}{N} = 0$$

which implies that

$$\lim_{N\to\infty} \sum_{i=0}^{N-1} h(x_i^*)\Delta x_i = 0.$$

This would force

$$\int_0^1 h(x)\,\mathrm{d}x = 0$$

which contradicts $(**)$ above.

Thus, we are forced to conclude that the Riemann integral

$$\int_0^1 h(x)\,\mathrm{d}x$$

does not exist.

Remark. The lesson here is that there exist simple functions whose Riemann integrals do not exist, but whose Lebesgue integrals do exist. The lesson for probabilists is that we define the expectation of a random variable as a Lebesgue integral so that expectations will be defined for a wider class of random variables.

Statistics 851 (Winter 2008) February 15, 2008
Prof. Michael Kozdron

Lecture #18: An Example of a Random Variable with a Non-Borel Range

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable.

Recall that the notation $X : \Omega \to \mathbb{R}$ means that the domain of X is Ω and the co-domain of X is $\mathbb{R}$. The function X is necessarily defined for every $\omega \in \Omega$. However, there is *no* requirement that X be onto $\mathbb{R}$. We write $X(\Omega)$ for the range of $\mathbb{R}$ so that $X(\Omega) \subseteq \mathbb{R}$.

Also recall that the condition for the function $X : \Omega \to \mathbb{R}$ to be a random variable is that the inverse image of a Borel set be an event. That is, we must have $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$.

The question might now be asked why we need to consider "inverse images" of Borel sets. Why is not enough to consider "images" of events in $\mathcal{A}$ under X?

Notice that we necessarily must have $X^{-1}(\mathbb{R}) = \Omega$ for *every* real-valued function $X : \Omega \to \mathbb{R}$. In particular, this is true whether or not X is measurable.

However, it need not be the case that $X(\Omega) = \mathbb{R}$. Again, it is the case that the range of X is all of $\mathbb{R}$ if and only if X is onto $\mathbb{R}$.

Example. For instance, suppose that $\Omega = [0, 1]$ and consider the two functions $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ given by

$$X(\omega) = \omega$$

and

$$Y(\omega) = \begin{cases} 0, & \text{if } \omega = 0, \\ -\frac{1}{4\omega}, & \text{if } 0 < \omega < \frac{1}{4}, \\ \frac{1}{2-4\omega}, & \text{if } \frac{1}{4} \leq \omega < \frac{1}{2}, \\ 4\omega - 3, & \text{if } \frac{1}{2} \leq \omega \leq 1. \end{cases}$$

It is not hard to show that both X and Y are random variables with respect to the uniform probability measure on $(\Omega, \mathcal{B}(\Omega))$. However, $X(\Omega) = [0, 1] \subsetneq \mathbb{R}$ while $Y(\Omega) = \mathbb{R}$.

As we have already noted, it is necessarily the case that the range of X satisfies $X(\Omega) \subseteq \mathbb{R}$. Hence, it is natural to ask whether or not $X(\Omega) \in \mathcal{B}$.

However, as the following example shows, it need not be the case that $X(\Omega) \in \mathcal{B}$ *even if* X is a random variable.

Example. Suppose that $H \subset [0, 1]$ is a non-measurable set and let $q \in H$ be a given point. Recall from Lecture #9 that the construction of H and the identification of q requires the Axiom of Choice.

Let $\Omega = H$ and let $\mathcal{A} = 2^{\Omega}$ be the power set of Ω. For $A \in \mathcal{A}$, let

$$P\{A\} = \begin{cases} 1, & \text{if } q \in A, \\ 0, & \text{if } q \notin A, \end{cases}$$

so that P is, in fact, a probability measure on $\mathcal{A}$. Define the function $X : \Omega \to \mathbb{R}$ by setting $X(\omega) = \omega$. If $B \in \mathcal{B}$ is a Borel set, then $X^{-1}(B) \subset \Omega$ so that we necessarily have $X^{-1}(B) \in \mathcal{A}$. In other words, X is a random variable. However, it is clearly the case that $X(\Omega) = H$ so that $X(\Omega) \notin \mathcal{B}$.

We end by proving a related result; namely, the fact that if $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is a random variable, then $X^{-1}(\mathcal{B})$ is a σ-algebra.

Proposition. *If $\mathcal{F} = X^{-1}(\mathcal{B}) = \{X^{-1}(B) : \text{for some } B \in \mathcal{B}\}$, then $\mathcal{F}$ is a σ-algebra.*

Proof. In order to prove that $\mathcal{F}$ is a σ-algebra, we need to verify that the three conditions in the definition are met.

(i) Since $\mathcal{B}$ is a σ-algebra, we know that $\emptyset \in \mathcal{B}$. However, it then follows that $\emptyset = X^{-1}(\emptyset) \in \mathcal{F}$.

(ii) Suppose that $A \in \mathcal{F}$ so that $A = X^{-1}(B)$ for some $B \in \mathcal{B}$. However, $A^c = [X^{-1}(B)]^c = X^{-1}(B^c)$. Since $\mathcal{B}$ is a σ-algebra, we know that $B^c \in \mathcal{B}$ and so $A^c \in \mathcal{F}$. In particular, combining this with (i) implies that $\Omega = \emptyset^c \in \mathcal{F}$.

(iii) Suppose that $A_1, A_2, \ldots \in \mathcal{F}$ so that $A_i = X^{-1}(B_i)$, $i = 1, 2, \ldots$, for some $B_1, B_2, \ldots \in \mathcal{B}$. However,

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} X^{-1}(B_i) = X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right).$$

Since $\mathcal{B}$ is a σ-algebra, we know that

$$\bigcup_{i=1}^{\infty} B_i \in \mathcal{B}$$

and so

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Thus, $\mathcal{F}$ is, in fact, a σ-algebra. □

The σ-algebra given in the previous proposition is important enough to have its own name.

Definition. Suppose that $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is a random variable. The *σ-algebra generated by X* is defined to be $\sigma(X) = X^{-1}(\mathcal{B})$.

Remark. Notice that the previous proposition never actually used the fact that X is a random variable. This result is true for *any* function $X : \Omega \to (\mathbb{R}, \mathcal{B})$.

Remark. In Exercise 9.1 you are asked to show that if $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ is a random variable, then it is also measurable as a function from $(\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$. In this case, we have $\sigma(X) \subseteq \mathcal{A}$.

However, if $X : \Omega \to \mathbb{R}$ is any function, then, although X is necessarily measurable as a function from $(\Omega, \sigma(X)) \to (\mathbb{R}, \mathcal{B})$, it is no longer true that $\sigma(X) \subseteq \mathcal{A}$.

Lecture #19: Construction of Expectation

Reference. Chapter 9, pages 51–60

Goal. To define

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

for all random variables $X : \Omega \to \mathbb{R}$.

Our strategy will be as follows. We will

(1) define $\mathbb{E}(X)$ for simple random variables,

(2) define $\mathbb{E}(X)$ for positive random variables,

(3) define $\mathbb{E}(X)$ for general random variables.

This strategy is sometimes called the "standard machine" and is the outline that we will follow to prove most results about expectation of random variables.

Step 1: Simple Random Variables

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Suppose that $X : \Omega \to \mathbb{R}$ is a simple random variable so that

$$X(\omega) = \sum_{j=1}^{m} a_j 1_{A_j}(\omega)$$

where $a_1, \ldots, a_m \in \mathbb{R}$ and $A_1, \ldots, A_m \in \mathcal{A}$. We define the expectation of X to be

$$\mathbb{E}(X) = \sum_{j=1}^{m} a_j P\{A_j\}.$$

Step 2: Positive Random Variables

Suppose that X is a positive random variable. That is, $X(\omega) \geq 0$ for all $\omega \in \Omega$. (We will need to allow $X(\omega) \in [0, +\infty]$ for some consistency.) We are also assuming at this step that X is not a simple random variable.

Definition. If X is a positive random variable, define the *expectation* of X to be

$$\mathbb{E}(X) = \sup\{\mathbb{E}(Y) : Y \text{ is simple and } 0 \leq Y \leq X\}. \qquad (*)$$

Proposition 1. *For every random variable $X \geq 0$, there exists a sequence (X_n) of positive, simple random variables with $X_n \uparrow X$ (that is, X_n increases to X).*

Proof. Let $X \geq 0$ be given and define the sequence

$$X_n(\omega) = \begin{cases} \frac{k}{2^n}, & \text{if } \frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n} \text{ and } 0 \leq k \leq n2^n - 1, \\ n, & \text{if } X(\omega) \geq n. \end{cases}$$

Then it follows that $X_n \leq X_{n+1}$ for every $n = 1, 2, 3, \ldots$ and $X_n \to X$ which completes the proof. □

Proposition 2. *If $X \geq 0$ and (X_n) is a sequence of simple random variables with $X_n \uparrow X$, then $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$. That is, if $X_n \leq X_{n+1}$ and*

$$\lim_{n\to\infty} X_n = X,$$

then

$$\lim_{n\to\infty} \int X_n \, \mathrm{d}P \uparrow \int \left(\lim_{n\to\infty} X_n \right) \mathrm{d}P = \int X \, \mathrm{d}P.$$

Proof. Suppose that $X \geq 0$ is a random variable and let (X_n) be a sequence of simple random variables with $X_n \geq 0$ and $X_n \uparrow X$. Observe that since the X_n are increasing, we have $\mathbb{E}(X_n) \leq \mathbb{E}(X_{n+1})$. Therefore, $\mathbb{E}(X_n)$ increases to some limit $a \in [0, \infty]$; that is,

$$\mathbb{E}(X_n) \uparrow a$$

for some $0 \leq a \leq \infty$. (If $\mathbb{E}(X_n)$ is an unbounded sequence, then $a = \infty$. However, if $\mathbb{E}(X_n)$ is a bounded sequence, then $a < \infty$ follows from the fact that increasing, bounded sequences have unique limits.) Therefore, it follows from $(*)$, the definition of $\mathbb{E}(X)$, that $a \leq \mathbb{E}(X)$.

We will now show $a \geq \mathbb{E}(X)$. As a result of $(*)$, we only need to show that if Y is a simple random variable with $0 \leq Y \leq X$, then $\mathbb{E}(Y) \leq a$. To this end, let Y be simple and write

$$Y = \sum_{k=1}^{m} a_k 1\{Y = a_k\}.$$

That is, take $A_k = \{\omega : Y(\omega) = a_k\}$. Let $0 < \epsilon \leq 1$ and define

$$Y_{n,\epsilon} = (1 - \epsilon) Y 1_{\{(1-\epsilon)Y \leq X_n\}}.$$

Note that $Y_{n,\epsilon} = (1 - \epsilon)a_k$ on the set

$$\{(1 - \epsilon)Y \leq X_n\} \cap A_k = A_{k,n,\epsilon}$$

and that $Y_{n,\epsilon} = 0$ on the set $\{(1 - \epsilon)Y > X_n\}$. Clearly $Y_{n,\epsilon} \leq X_n$ and so

$$\mathbb{E}(Y_{n,\epsilon}) = (1 - \epsilon) \sum_{k=1}^{m} a_k P\{A_{k,n,\epsilon}\} \leq \mathbb{E}(X_n).$$

Therefore, since

$$Y \leq X = \lim_{n\to\infty} X_n$$

we see that

$$(1-\epsilon)Y < \lim_{n\to\infty} X_n$$

as soon as $Y > 0$. (This requires that X is not identically 0.) Hence,

$$A_{k,n,\epsilon} \to A_k$$

as $n \to \infty$. By Theorem 2.4 we conclude that

$$P\{A_{k,n,\epsilon}\} \to P\{A_k\}$$

so that

$$\mathbb{E}(Y_{n,\epsilon}) = (1-\epsilon)\sum_{k=1}^{m} a_k P\{A_{k,n,\epsilon}\} \to (1-\epsilon)\sum_{k=1}^{m} a_k P\{A_k\} = (1-\epsilon)\mathbb{E}(Y) \leq a.$$

That is,

$$E(Y_{n,\epsilon}) \leq \mathbb{E}(X_n)$$

and

$$E(Y_{n,\epsilon}) \to (1-\epsilon)\mathbb{E}(Y) \quad \text{and} \quad \mathbb{E}(X_n) \to a$$

so that

$$(1-\epsilon)\mathbb{E}(Y) \leq a$$

(which follows since everything is increasing). Let $\epsilon > 0$ so that

$$\mathbb{E}(Y) \leq a.$$

By $(*)$,

$$\sup\{\mathbb{E}(Y)\} \leq a$$

which implies that

$$\mathbb{E}(X) \leq a.$$

Combined with our earlier result that $a \leq \mathbb{E}(X)$ we conclude $\mathbb{E}(X) = a$ and the proof is complete. □

Step 3: General Random Variables

Now suppose that X is any random variable. Write

$$X^+ = \max\{X, 0\} \quad \text{and} \quad X^- = -\min\{X, 0\}$$

for the positive part and the negative part of X, respectively. Note that $X^+ \geq 0$ and $X^- \geq 0$ so that the positive part and negative part of X are both positive random variables and $X = X^+ - X^-$.

Definition. A random variable X is called *integrable* (or has *finite expectation*) if both $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ are finite. In this case we define $\mathbb{E}(X)$ to be

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-).$$

Remark. Having proved this result, we see that the standard machine is really not that hard to implement. In fact, it is usually enough to prove a result for simple random variables and then extend that result to positive random variables using the result of Step 2. Step 3 for general random variables usually follows by definition.

Lecture #20: Independence

Reference. Chapter 10, pages 65–67

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Recall that events $A \in \mathcal{A}$ and $B \in \mathcal{A}$ are *independent* if

$$P\{A \cap B\} = P\{A\} \cdot P\{B\}.$$

We can extend this idea to collections of events as follows.

Definition. Two sub-σ-algebras $\mathcal{A}_1 \subseteq \mathcal{A}$ and $\mathcal{A}_2 \subseteq \mathcal{A}$ are *independent* if

$$P\{A_1 \cap A_2\} = P\{A_1\} \cdot P\{A_2\}$$

for every $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.

Furthermore, a collection $(\mathcal{A}_i)$, $i \in I$, of sub-σ-algebras of $\mathcal{A}$ is *independent* if for every finite subset J of I, we have

$$P\left\{\bigcap_{i\in J} A_i\right\} = \prod_{i\in J} P\{A_i\}$$

for all $A_i \in \mathcal{A}_i$.

We can now discuss independent random variables.

Recall. Suppose that $(E, \mathcal{E})$ is a measurable space. The function $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ is called measurable if $X^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{E}$. Usually we take $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B})$ and call X a random variable.

Now, as shown in Lecture #18,

$$X^{-1}(\mathcal{E}) = \{A \in \mathcal{A} : X(A) = B \text{ for some } B \in \mathcal{E}\}$$

is a σ-algebra with $X^{-1}(\mathcal{E}) \subseteq \mathcal{A}$.

This sub-σ-algebra is so important that it gets its own name and notation.

Notation. We write $\sigma(X)$ to denote the *σ-algebra generated by X*. In other words, $\sigma(X) = X^{-1}(\mathcal{E})$.

Definition. Suppose that $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are measurable spaces, and let $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ and $Y : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$ be random variables. We say that X and Y are *indpendent* if $\sigma(X)$ and $\sigma(Y)$ are independent σ-algebras.

Notice that both $\sigma(X)$ and $\sigma(Y)$ are sub-σ-algebras of $\mathcal{A}$ even though X and Y might have different target spaces.

Theorem. *The random variables X and Y are independent if and only if*

$$P\{X \in A, Y \in B\} = P\{X \in A\} \cdot P\{Y \in B\}$$

for all $A \in \mathcal{E}$ and $B \in \mathcal{F}$.

Proof. This follows from the definition of $\sigma(X)$ since $X^{-1}(\mathcal{E})$ is exactly the σ-algebra generated by events of the form $\{X \in A\}$ for $A \in \mathcal{E}$. That is,

$$\sigma(X) = \sigma(X \in A, A \in \mathcal{E}) \quad \text{and} \quad \sigma(Y) = \sigma(Y \in B, B \in \mathcal{F})$$

and so the theorem follows. □

Corollary. *The random variables X and Y are independent if and only if*

$$P^{X,Y}\{A, B\} = P^X\{A\} \cdot P^Y\{B\}$$

where $P^{X,Y}$ denotes the joint law of X and Y.

Corollary. *The random variables X and Y are independent if and only if*

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y)$$

where $F_{X,Y}$ denotes the joint distribution function of X and Y.

Corollary. *If the random variables X and Y have densities f_X and f_Y, respectively, and* joint density *$f_{X,Y}$, then X and Y are independent if and only if*

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

Remark. We still need to define the terms *joint law*, *joint distribution function*, and *joint density function*. This will require the notion of product spaces.

Theorem. *If X and Y are independent random variables, and if f and g are measurable functions, then $f \circ X$ and $g \circ Y$ are independent random variables.*

Proof. Given f and g we have

$$(f \circ X)^{-1}(\mathcal{E}) = X^{-1}(f^{-1}(\mathcal{E})) \subseteq X^{-1}(\mathcal{E})$$

and

$$(g \circ Y)^{-1}(\mathcal{F}) = Y^{-1}(g^{-1}(\mathcal{F})) \subseteq Y^{-1}(\mathcal{F}).$$

If X and Y are independent, then $X^{-1}(\mathcal{E})$ and $Y^{-1}(\mathcal{F})$ are independent so that $X^{-1}(f^{-1}(\mathcal{E}))$ and $Y^{-1}(g^{-1}(\mathcal{F}))$ are independent. That is,

$$\sigma(f \circ X) \quad \text{and} \quad \sigma(g \circ Y)$$

are independent and the proof is complete. □

Theorem. *If X and Y are independent random variables, then*

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X)) \cdot \mathbb{E}(g(Y)) \qquad (*)$$

for every pair of bounded, measurable functions f, g, or positive, measurable functions f, g.

Remark. If X and Y are jointly continuous random variables, then this result is easily proved as in STAT 351:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(y)f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)g(y)f_X(x)f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} f(x)f_X(x)\,\mathrm{d}x \cdot \int_{-\infty}^{\infty} g(y)f_Y(y)\,\mathrm{d}y.\end{aligned}$$

Proof. In order to prove this result, we follow that standard machine. Suppose that X and Y are random variables and that f and g are measurable functions. The first step is to show that the result holds when f and g are simple functions. Suppose that $f(x) = 1_A(x)$ and $g(y) = 1_B(y)$. Then,

$$\begin{aligned}\mathbb{E}(f(X)g(Y)) = \mathbb{E}(1_A(X)1_B(Y)) &= \mathbb{E}(1\{X \in A, Y \in B\}) \\ &= P\{X \in A, Y \in B\} \\ &= P\{X \in A\} \cdot P\{Y \in B\} \\ &= \mathbb{E}(1_A(X)) \cdot \mathbb{E}(1_B(Y)) \\ &= \mathbb{E}(f(X)) \cdot \mathbb{E}(g(Y)).\end{aligned}$$

If

$$f(x) = \sum_{i=1}^{n} a_i 1_{A_i}(x) \quad \text{and} \quad g(y) = \sum_{j=1}^{m} b_j 1_{B_j}(y),$$

then the result $(*)$ holds by linearity.

The second step is to show that the result holds for positive functions. Thus, let f, g be positive functions and assume that f_n and g_n are sequences of positive, simple functions with $f_n \uparrow f$ and $g_n \uparrow g$. Note that

$$f_n(x) \cdot g_n(y) \uparrow f(x) \cdot g(y).$$

Therefore,

$$\begin{aligned}\mathbb{E}(f(X) \cdot g(Y)) &= \mathbb{E}\left(\lim_{n\to\infty} f_n(X) \cdot g_n(Y)\right) \\ &= \lim_{n\to\infty} \mathbb{E}(f_n(X) \cdot g_n(Y)) \quad \text{by the Monotone Convergence Theorem} \\ &= \lim_{n\to\infty} [\mathbb{E}(f_n(X)) \cdot \mathbb{E}(g_n(Y))] \quad \text{by Step 1} \\ &= \lim_{n\to\infty} \mathbb{E}(f_n(X)) \cdot \lim_{n\to\infty} \mathbb{E}(g_n(Y)) \\ &= \mathbb{E}\left(\lim_{n\to\infty} f_n(X)\right) \cdot \mathbb{E}\left(\lim_{n\to\infty} g_n(Y)\right) \quad \text{by the Monotone Convergence Theorem} \\ &= \mathbb{E}(f(X)) \cdot \mathbb{E}(g(Y)).\end{aligned}$$

The third step is to conclude that if f, g are both bounded, then the result follows by writing

$$f = f^+ - f^- \quad \text{and} \quad g = g^+ - g^-$$

and using linearity. $\square$

Lecture #21: Product Spaces

Reference. Chapter 10, pages 67–71

We begin with the following result which gives a characterization of independent σ-algebras in terms of the probabilities of the events.

Theorem. *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space. The σ-algebra $\mathcal{A}$ is independent of itself if and only if $P\{A\} \in \{0,1\}$ for all $A \in \mathcal{A}$.*

Proof. Suppose that $\mathcal{A}$ is independent of itself. This implies that $P\{A \cap B\} = P\{A\} \cdot P\{B\}$ for every A, $B \in \mathcal{A}$. In particular, let $A \in \mathcal{A}$ so that $A^c \in \mathcal{A}$. Therefore,

$$0 = P\{\emptyset\} = P\{A \cap A^c\} = P\{A\} \cdot P\{A^c\}$$

which can only be satisfied if either $P\{A\} = 0$ or $P\{A^c\} = 0$. That is, we must have $P\{A\} \in \{0,1\}$.

Conversely, suppose that $P\{A\} \in \{0,1\}$ for all $A \in \mathcal{A}$. If A, $B \in \mathcal{A}$, then there are three cases that need to be considered.

(i) Suppose that $P\{A\} = P\{B\} = 0$. Therefore, since

$$0 \leq P\{A \cap B\} \leq P\{A\} = 0$$

we conclude that

$$P\{A \cap B\} = P\{A\} \cdot P\{B\} \ \ (= 0).$$

(ii) Suppose that $P\{A\} = 0$ and $P\{B\} = 1$. Therefore, since

$$0 \leq P\{A \cap B\} \leq P\{A\} = 0$$

we again conclude that

$$P\{A \cap B\} = P\{A\} \cdot P\{B\} \ \ (= 0).$$

(Similarly, we reach the same conclusion if $P\{A\} = 1$ and $P\{B\} = 0$.)

(iii) Suppose that $P\{A\} = P\{B\} = 1$. Therefore,

$$0 \leq P\{(A \cap B)^c\} = P\{A^c \cup B^c\} \leq P\{A^c\} + P\{B^c\} = 0$$

which implies that $P\{A \cap B\} = 1$ and so

$$P\{A \cap B\} = P\{A\} \cdot P\{B\} \ \ (= 1).$$

Thus we have demonstrated both implications, the theorem is proved. □

An equivalent formulation of the theorem is given by the contrapositive.

Theorem. *The σ-algebra $\mathcal{A}$ is not independent of itself if and only if there exists an $A \in \mathcal{A}$ with $P\{A\} \in (0,1)$.*

Remark. Therefore, suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and suppose that $\mathcal{A}_1$ and $\mathcal{A}_2$ are two sub-σ-algebras of $\mathcal{A}$. It is possible to discuss the independence (or non-independence) of $\mathcal{A}_1$ and $\mathcal{A}_2$ even if $\mathcal{A}_1 = \mathcal{A}_2$.

As indicated last class, it is necessary to talk about joint distributions. In order to do so, however, we must define product spaces.

Let $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ be a random variable and let $Y : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$ be a random variable.

Definition. If E and F are sets, then we write

$$E \times F = \{(x, y) : x \in E, y \in F\}$$

for their *Cartesian product*.

Consider the function $Z : \Omega \to E \times F$ given by

$$\omega \mapsto Z(\omega) = (X(\omega), Y(\omega)).$$

It should seem "obvious" that Z is a random variable. However, in order to prove that $Z : \Omega \to E \times F$ is, in fact, measurable, we need to have a σ-algebra on $E \times F$.

If $\mathcal{E}$ and $\mathcal{F}$ are σ-algebras, then

$$\mathcal{E} \times \mathcal{F} = \{A \subset E \times F : A = \Lambda \times \Gamma, \Lambda \in \mathcal{E}, \Gamma \in \mathcal{F}\}$$

is well-defined although it need not be a σ-algebra. Let

$$\sigma(\mathcal{E}, \mathcal{F}) = \mathcal{E} \otimes \mathcal{F}$$

denote the smallest σ-algebra containing both $\mathcal{E}$ and $\mathcal{F}$.

Hence, $(E \times F, \mathcal{E} \otimes \mathcal{F})$ is also a measurable space, and this leads us to the following definition.

Definition. A function $f : (E \times F, \mathcal{E} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is *measurable* if $f^{-1}(B) \in \mathcal{E} \otimes \mathcal{F}$ for every $B \in \mathcal{B}$.

Suppose further that P_1 is a probability on $(E, \mathcal{E})$ and that P_2 is a probability on $(F, \mathcal{F})$. Is it possible to define a probability P on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ in a natural way?

Answer. YES!

Intuition. Let $C \in \mathcal{E} \times \mathcal{F}$ so that $C = A \times B$. We should define the probability $P\{C\} = P_1\{A\} \cdot P_2\{B\}$. That is, we should define the probability of the product as the product of the probabilities. The problem, however, is that $\mathcal{E} \times \mathcal{F}$ need not equal $\mathcal{E} \otimes \mathcal{F}$. That is, there are events $C \in \mathcal{E} \otimes \mathcal{F}$ which are NOT of the form $C = A \times B$ for some $A \in \mathcal{E}$, $B \in \mathcal{F}$. How should we define P for these sets? The answer will be given by the Fubini-Tonelli Theorem.

Statistics 851 (Winter 2008)
Exercises on Independence

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space with $\Omega = \{a, b, c, d, e\}$ and $\mathcal{A} = 2^{\Omega}$. Let X and Y be the real-valued random variables defined by

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in \{a, b\}, \\ 0, & \text{if } \omega \notin \{a, b\}, \end{cases} \qquad Y(\omega) = \begin{cases} 2, & \text{if } \omega \in \{a, c\}, \\ 0, & \text{if } \omega \notin \{a, c\}. \end{cases}$$

(a) Give explicitly (by listing all the elements) the σ-algebras $\sigma(X)$ and $\sigma(Y)$ generated by X and Y, respectively.

(b) Find the σ-algebra $\sigma(X, Y)$ generated (jointly) by X and Y.

(c) If $Z = X + Y$, does $\sigma(Z) = \sigma(X, Y)$?

For real numbers α, $\beta \geq 0$ with $\alpha + \beta \leq 1/2$, let P be the probability measure on $\mathcal{A}$ determined by the relations

$$P(\{a\}) = P(\{b\}) = \alpha, \quad P(\{c\}) = P(\{d\}) = \beta, \quad P(\{e\}) = 1 - 2(\alpha + \beta).$$

(d) Find all α, β for which $\sigma(X)$ and $\sigma(Y)$ are independent. Simplify!

(e) Find all α, β for which X and $Z = X + Y$ are independent.

Statistics 851 (Winter 2008)
Exercises on Independence (Solutions)

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space with $\Omega = \{a, b, c, d, e\}$ and $\mathcal{A} = 2^{\Omega}$. Let X and Y be the real-valued random variables defined by

$$X(\omega) = \begin{cases} 1, & \text{if } \omega \in \{a, b\}, \\ 0, & \text{if } \omega \notin \{a, b\}, \end{cases} \qquad Y(\omega) = \begin{cases} 2, & \text{if } \omega \in \{a, c\}, \\ 0, & \text{if } \omega \notin \{a, c\}. \end{cases}$$

(a) Give explicitly (by listing all the elements) the σ-algebras $\sigma(X)$ and $\sigma(Y)$ generated by X and Y, respectively.

Suppose that $B \in \mathcal{B}$ so that

$$X^{-1}(B) = \begin{cases} \Omega, & \text{if } 1 \in B, 0 \in B, \\ \{a, b\}, & \text{if } 1 \in B, 0 \notin B, \\ \{c, d, e\}, & \text{if } 1 \notin B, 0 \in B, \\ \emptyset, & \text{if } 1 \notin B, 0 \notin B. \end{cases}$$

Thus, the σ-algebra generated by X is $\sigma(X) = \{\emptyset, \Omega, \{a, b\}, \{c, d, e\}\}$. Similarly,

$$Y^{-1}(B) = \begin{cases} \Omega, & \text{if } 2 \in B, 0 \in B, \\ \{a, c\}, & \text{if } 2 \in B, 0 \notin B, \\ \{b, d, e\}, & \text{if } 2 \notin B, 0 \in B, \\ \emptyset, & \text{if } 2 \notin B, 0 \notin B. \end{cases}$$

so that $\sigma(Y) = \{\emptyset, \Omega, \{a, c\}, \{b, d, e\}\}$.

(b) Find the σ-algebra $\sigma(X, Y)$ generated (jointly) by X and Y.

By definition,

$$\begin{aligned} \sigma(X, Y) &= \sigma(\sigma(X), \sigma(Y)) = \sigma(\{a\}, \{b\}, \{c\}, \{d, e\}) \\ &= \{\emptyset, \Omega, \{a\}, \{b\}, \{c\}, \{d, e\}, \{a, b\}, \{a, c\}, \ \{a, d, e\}, \{b, c\}, \{b, d, e\}\{c, d, e\}, \{a, b, c\}, \\ &\quad \{a, b, d, e\}, \{a, c, d, e\}, \{b, c, d, e\}\}. \end{aligned}$$

(c) If $Z = X + Y$, does $\sigma(Z) = \sigma(X, Y)$?

If $Z = X + Y$ then

$$Z(\omega) = \begin{cases} 3, & \text{if } \omega = a \\ 2, & \text{if } \omega = c \\ 1, & \text{if } \omega = b \\ 0, & \text{if } \omega \in \{d, e\} \end{cases}$$

so that

$$\sigma(Z) = \sigma(\{a\}, \{b\}, \{c\}, \{d, e\}) = \sigma(X, Y).$$

For real numbers α, $\beta \geq 0$ with $\alpha + \beta \leq 1/2$, let P be the probability measure on $\mathcal{A}$ determined by the relations

$$P(\{a\}) = P(\{b\}) = \alpha, \quad P(\{c\}) = P(\{d\}) = \beta, \quad P(\{e\}) = 1 - 2(\alpha + \beta).$$

(d) Find all α, β for which $\sigma(X)$ and $\sigma(Y)$ are independent. Simplify!

Recall that $\sigma(X)$ and $\sigma(Y)$ are independent iff $P(A \cap B) = P(A)P(B)$ for every $A \in \sigma(X)$ and $B \in \sigma(Y)$. Notice that if either A or B are either $\emptyset$ or Ω, then $P(A \cap B) = P(A)P(B)$. Thus, in order to show that $\sigma(X)$ and $\sigma(Y)$ are independent, we need to simultaneously satisfy the four equalities.

$$P(\{a,b\} \cap \{a,c\}) = P(\{a,b\})P(\{a,c\}), \quad P(\{c,d,e\} \cap \{a,c\}) = P(\{c,d,e\})P(\{a,c\}),$$

$$P(\{a,b\} \cap \{b,d,e\}) = P(\{a,b\})P(\{b,d,e\}), \quad P(\{c,d,e\} \cap \{b,d,e\}) = P(\{c,d,e\})P(\{b,d,e\}).$$

By the definition of P, we have

$$P(\{a,b\}) = 2\alpha, \quad P(\{c,d,e\}) = 1 - 2\alpha, \quad P(\{a,c\}) = \alpha + \beta, \quad P(\{b,d,e\}) = 1 - \alpha - \beta$$

as well as

$$P(\{a,b\} \cap \{a,c\}) = P(\{a\}) = \alpha, \quad P(\{c,d,e\} \cap \{a,c\}) = P(\{c\}) = \beta,$$

$$P(\{a,b\} \cap \{b,d,e\}) = P(\{b\}) = \alpha, \quad P(\{c,d,e\} \cap \{b,d,e\}) = P(\{d,e\}) = 1 - 2\alpha - \beta.$$

Hence, our system of equations for α and β becomes

$$\begin{aligned}
\alpha &= 2\alpha(\alpha + \beta) \\
\beta &= (1 - 2\alpha)(\alpha + \beta) \\
\alpha &= 2\alpha(1 - \alpha - \beta) \\
1 - 2\alpha - \beta &= (1 - 2\alpha)(1 - \alpha - \beta)
\end{aligned}$$

which has solution set

$$\left\{(\alpha, \beta) : 0 \leq \alpha \leq \frac{1}{2},\ 0 \leq \beta \leq \frac{1}{2},\ \alpha + \beta = \frac{1}{2}\right\}.$$

(e) Find all α, β for which X and $Z = X + Y$ are independent.

In order for X and Z to be independent, we must have

$$P(X = i, Z = j) = P(X = i)P(Z = j)$$

for $i = 0, 1$, $j = 0, 1, 2, 3$. Notice, however, that

$$(X, Z) = \begin{cases} (1,3), & \text{if } \omega \in \{a\}, \\ (1,1), & \text{if } \omega \in \{b\}, \\ (0,2), & \text{if } \omega \in \{c\}, \\ (0,0), & \text{if } \omega \in \{d,e\}. \end{cases}$$

These imply that

$$P(\{a\}) = P(\{a,b\})P(\{a\}), \;\; P(\{b\}) = P(\{a,b\})P(\{b\}),$$

$$P(\{c\}) = P(\{c,d,e\})P(\{c\}), \;\; P(\{d,e\}) = P(\{c,d,e\}P(\{d,e\})$$

which are simultaneously satisfied by either ($\alpha = 1/2$, $\beta = 0$) or ($\alpha = 0$, $\beta = 0$).

Lecture #22: Product Spaces

Reference. Chapter 10, pages 67–71

Recall. We introduced the notion of the product space of two measurable spaces. Intuitively, the product space can be viewed as a model of independence. That is, suppose that $(\Omega_1, \mathcal{A}_1, P_1)$ and $(\Omega_2, \mathcal{A}_2, P_2)$ are two probability spaces. Consider the space

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1,\, \omega_2 \in \Omega_2\}.$$

Although the product $\mathcal{A}_1 \times \mathcal{A}_2$ might seem like the natural choice for a σ-algebra on $\Omega_1 \times \Omega_2$, it is not necessarily a σ-algebra. Hence, in order to turn $\Omega_1 \times \Omega_2$ into a measurable space, we consider

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma(\mathcal{A}_1 \times \mathcal{A}_2)$$

which is the smallest σ-algebra containing $\mathcal{A}_1 \times \mathcal{A}_2$.

The question now is, "Can we define a probability measure on $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ in a natural way?"

The answer is YES!

If $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$, then $C = A_1 \times A_2 \in \mathcal{A}_1 \times \mathcal{A}_2$ and so we can *define* the probability of C to be

$$P\{C\} = P\{A_1 \times A_2\} = P_1\{A_1\} \cdot P_2\{A_2\}.$$

This is the natural definition of P and is motivated by the definition of independence.

The trouble, however, is that $\mathcal{A}_1 \times \mathcal{A}_2$ need not equal $\mathcal{A}_1 \otimes \mathcal{A}_2$ and so we need to figure out how to assign probability to events in $\mathcal{A}_1 \otimes \mathcal{A}_2 \setminus \mathcal{A}_1 \times \mathcal{A}_2$.

Example. Suppose that $\Omega_1 = \Omega_2 = \{a, b\} = \Omega$ and

$$\mathcal{A}_1 = \mathcal{A}_2 = 2^\Omega = \{\emptyset, \{a\}, \{b\}, \Omega\}.$$

Define the probabilities P_1 and P_2 by setting

$$P_1\{a\} = P_2\{a\} = p \quad \text{and} \quad P_1\{b\} = P_2\{b\} = 1 - p$$

for some $0 < p < 1$.

Consider the product $\Omega_1 \times \Omega_2$ which is given by

$$\Omega_1 \times \Omega_2 = \{(a, a), (a, b), (b, a), (b, b)\}$$

and consider the set $C \subset \Omega_1 \times \Omega_2$ given by

$$C = \{(a, a), (a, b), (b, b)\}.$$

Note that C is NOT of the form $C = A_1 \times A_2$. Therefore, we cannot assign a probability to C by simply defining

$$P\{C\} = P_1\{A_1\} \cdot P_2\{A_2\}$$

and so we need to be able to extend $P_1 \times P_2$ to be a probability on all of $\mathcal{A}_1 \otimes \mathcal{A}_2$.

Finally, note that in this example, which models two independent tosses of a coin, we would like to assign

$$P\{C\} = p^2 + p(1-p) + (1-p)^2 = 1 - p + p^2.$$

Exercise. Explicitly list all of the elements of $\mathcal{A}_1 \times \mathcal{A}_2$ and $\mathcal{A}_1 \otimes \mathcal{A}_2$. Give the probability $P_1 \times P_2\{C\}$ for all $C \in \mathcal{A}_1 \times \mathcal{A}_2$. Define $P\{C\}$ (in a consistent way) for all $C \in \mathcal{A}_1 \otimes \mathcal{A}_2 \setminus \mathcal{A}_1 \times \mathcal{A}_2$.

Random Vectors

We begin with the following result given in Exercise 10.1 which shows that (X_1, X_2) is a random vector if and only if X_1 and X_2 are both random variables.

Theorem. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and suppose that $X = (X_1, X_2) : \Omega \to E \times F$. The function $X : (\Omega, \mathcal{A}) \to (E \times F, \mathcal{E} \otimes \mathcal{F})$ is measurable if and only if $X_1 : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ and $X_2 : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$ are measurable.*

Proof. Suppose that both $X_1 : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ and $X_2 : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$ are measurable. Let $A \in \mathcal{E}$, $B \in \mathcal{F}$, so that $X_1^{-1}(A) \in \mathcal{A}$ and $X_2^{-1}(B) \in \mathcal{A}$ by the measurability of X_1, X_2, respectively. Therefore,

$$X^{-1}(A \times B) = X_1^{-1}(A) \cap X_2^{-1}(B) \in \mathcal{A}$$

so that X is measurable by the definition of the product σ-algebra $\mathcal{E} \otimes \mathcal{F}$ and Theorem 8.1 (since $\mathcal{E} \otimes \mathcal{F}$ is generated by $\mathcal{E} \times \mathcal{F} = \{A \times B : A \in \mathcal{E},\ B \in \mathcal{F}\}$).

Conversely, suppose that $X = (X_1, X_2) : \Omega \to E \times F$ is measurable so that $X^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{E} \otimes \mathcal{F}$. In particular, $X^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{E} \times \mathcal{F}$. If $C \in \mathcal{E} \times \mathcal{F}$, then it is necessarily of the form $A \times B$ for some $A \in \mathcal{E}$ and some $B \in \mathcal{F}$, and so

$$X^{-1}(C) = X^{-1}(A \times B) = X_1^{-1}(A) \cap X_2^{-1}(B) \in \mathcal{A}. \qquad (*)$$

In order to show that $X_1 : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$ is measurable, we must show $X_1^{-1}(A) \in \mathcal{A}$ for all $A \in \mathcal{E}$. To do this, we simply make a judicious choice of B in $(*)$. Choosing $B = F$ gives

$$\mathcal{A} \ni X_1^{-1}(A) \cap X_2^{-1}(F) = X_1^{-1}(A) \cap \Omega = X_1^{-1}(A).$$

Similarly, to show $X_2 : (\Omega, \mathcal{A}) \to (F, \mathcal{F})$ is measurable we choose $A = E$ in $(*)$. This gives

$$\mathcal{A} \ni X_1^{-1}(E) \cap X_2^{-1}(B) = \Omega \cap X_2^{-1}(B) = X_2^{-1}(B).$$

Taken together, the proof is complete. □

Sections of a measurable function

Suppose that $f : (E \times F, \mathcal{E} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ given by

$$(x, y) \mapsto f(x, y) = z$$

is measurable.

Fix $x \in E$ and let $f_x : F \to \mathbb{R}$ be given by

$$f_x(y) = f(x, y).$$

In other words, we start with a measurable function of two variables, say $f(x, y)$, and view it as a function of y only for fixed x. Sometimes we will also write

$$f_x(\cdot) = f(x, \cdot) \quad \text{or} \quad y \mapsto f(x, y).$$

Similarly, fix $y \in F$ and let $f_y : E \to \mathbb{R}$ be given by

$$f_y(x) = f(x, y).$$

That is, $f_y(\cdot) = f(\cdot, y)$ or $x \mapsto f(x, y)$.

We call f_x and f_y the *sections of* f.

Remark. There is a similarity between the sections of a measurable function and the marginal densities of a random vector. Suppose that (X, Y) is a random vector with joint density function $f(x, y)$. The sections of f are the functions $f_x(y)$ and $f_y(x)$ obtained simply by viewing $f(x, y)$ as a function of a single variable. The marginal densities, however, are obtained by "integrating out" the other variable. That is,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}y$$

is the marginal density function of X and

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, \mathrm{d}x$$

is the marginal density function of Y.

Theorem. *If $f : (E \times F, \mathcal{E} \otimes \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ is measurable, then the section $f_x : F \to \mathbb{R}$ is $\mathcal{F}$-measurable and the section $f_y : E \to \mathbb{R}$ is $\mathcal{E}$-measurable.*

Remark. The converse is false in general.

Proof. We will apply the standard machine and show that f_x is $\mathcal{F}$-measurable. (The proof that f_y is $\mathcal{E}$-measurable is identical.)

Let $f(x, y) = 1_C(x, y)$ for $C \in \mathcal{E} \otimes \mathcal{F}$ and let

$$\mathcal{H} = \{C \in \mathcal{E} \otimes \mathcal{F} : y \mapsto 1_C(x, y) \text{ is } \mathcal{F}\text{-measurable for each fixed } x \in E\}.$$

Note that $\mathcal{H}$ is a σ-algebra with $\mathcal{H} \subseteq \mathcal{E} \otimes \mathcal{F}$. Furthermore, $\mathcal{H} \supseteq \mathcal{E} \times \mathcal{F}$ which implies that

$$\mathcal{H} \supseteq \sigma(\mathcal{E} \times \mathcal{F}).$$

That is,

$$\mathcal{H} = \mathcal{E} \otimes \mathcal{F}.$$

Therefore, the result holds if f is an indicator function, and therefore holds if

$$f(x, y) = \sum_{i=1}^{n} c_i 1_{C_i}(x, y)$$

where $c_1, \ldots, c_n \in \mathbb{R}$ and $C_1, \ldots, C_n \in \mathcal{E} \otimes \mathcal{F}$ by linearity (Theorem 8.4).

Now suppose that $f \geq 0$ is a positive function and let (f_n) be a sequence of positive, simple functions with $f_n \uparrow f$. If $x \in E$ is fixed and

$$g_n(y) = f_n(x, y),$$

then g_n is $\mathcal{F}$-measurable by the first part. We now let

$$f_x(y) = \lim_{n\to\infty} g_n(y) = \lim_{n\to\infty} f_n(x, y)$$

which is $\mathcal{F}$-measurable since the limit of measurable functions is measurable (Theorem 8.4).

Finally, for general functions f, we consider $f = f^+ - f^-$ which expresses f as the difference of measurable functions and is therefore itself measurable. (This is also part of Theorem 8.4). □

Theorem (Fubini-Tonelli)**.** *(a) Let $R\{A \times B\} = P\{A\} \cdot Q\{B\}$ for $A \in \mathcal{E}$, $B \in \mathcal{F}$. Then R extends uniquely to a probability on $(E \times F, \mathcal{E} \otimes \mathcal{F})$ which we denote by $P \otimes Q$.*

(b) Suppose that $f : E \times F \to \mathbb{R}$. If f is either (i) $\mathcal{E} \otimes \mathcal{F}$-measurable and positive, or (ii) $\mathcal{E} \otimes \mathcal{F}$-measurable and integrable with respect to $P \otimes Q$, then

- *the function $x \mapsto \int f(x, y)Q(\mathrm{d}y)$ is $\mathcal{E}$-measurable,*
- *the function $y \mapsto \int f(x, y)P(\mathrm{d}x)$ is $\mathcal{F}$-measurable, and*
- $$\begin{aligned}\int f \,\mathrm{d}(P \otimes Q) &= \int f(x, y)P \otimes Q(\mathrm{d}x, \mathrm{d}y) \\ &= \int \int f(x, y)Q(\mathrm{d}y)P(\mathrm{d}x) \\ &= \int \int f(x, y)P(\mathrm{d}x)Q(\mathrm{d}y).\end{aligned}$$

Remark. Compare (a) with Theorem 6.4. We define R in the "natural" way and extend it to null sets. Part (b) says that we can evaluate a double integral in either order provided that f is (i) measurable and positive, or (ii) measurable and integrable.

Remark. We will omit the proof. Part (a) uses the Monotone Class Theorem while part (b) uses the standard machine.

Lecture #23: The Fubini-Tonelli Theorem

Reference. Chapter 10, pages 67–71

We begin by recalling the Fubini-Tonelli Theorem which gives us conditions under which we can interchange the order of Lebesgue integration:

$$\int\int f(x,y)Q(\mathrm{d}y)P(\mathrm{d}x) = \int\int f(x,y)P(\mathrm{d}x)Q(\mathrm{d}y). \qquad (*)$$

Formally, suppose that $(E,\mathcal{E},P)$ and $(F,\mathcal{F},Q)$ are measure spaces and consider the product space $(E\times F,\mathcal{E}\otimes\mathcal{F},P\otimes Q)$. Let $f:E\times F\to\mathbb{R}$ be a function.

The theorem states that equality holds in $(*)$ if f is either (i) $\mathcal{E}\otimes\mathcal{F}$-measurable and positive, or (ii) $\mathcal{E}\otimes\mathcal{F}$-measurable and integrable with respect to $P\otimes Q$.

We also recall the following theorem which tells us that joint measurability of $f(x,y)$ implies measurability of the sections $f_x(y)$ and $f_y(x)$.

Theorem. *If $f:(E\times F,\mathcal{E}\otimes\mathcal{F})\to(\mathbb{R},\mathcal{B})$ is measurable, then for fixed $x\in E$ the section $f_x:F\to\mathbb{R}$ given by $f_x(y)=f(x,y)$ is $\mathcal{F}$-measurable, and for fixed $y\in F$ the section $f_y:E\to\mathbb{R}$ given by $f_x(y)=f(x,y)$ is $\mathcal{E}$-measurable.*

Recall. If f is continuous, then f is measurable. (In fact, if f is continuous almost surely, then f is measurable.)

Example. Suppose that $E=[0,2]$, $\mathcal{E}=\mathcal{B}(E)$, P is the uniform probability on E, and suppose further than $F=[0,1]$, $\mathcal{F}=\mathcal{B}(F)$, Q is the uniform probability on F.

Consider the function

$$f(x,y)=\begin{cases}\frac{xy(x^2-y^2)}{(x^2+y^2)^3}, & \text{if } (x,y)\neq(0,0),\\ 0, & \text{if } (x,y)=(0,0).\end{cases}$$

It is clear that f is $\mathcal{E}\otimes\mathcal{F}$-measurable since the only possible point of discontinuity is at $(0,0)$.

Fix $x\in[0,2]$ and consider the section $f_x(y)$. For instance, if $x=1$, then

$$f_{x=1}(y)=\frac{y(1-y^2)}{(1+y^2)^3} \quad \text{for } y\in[0,1].$$

Clearly $f_{x=1}(y)$ is continuous (and hence measurable) for $y\in[0,1]=F$. Or, for instance, suppose that $x=0$. Then $f_{x=0}(y)=0$ for $y\in[0,1]$. In general, for fixed $x\in(0,2]$,

$$f_x(y)=\frac{xy(x^2-y^2)}{(x^2+y^2)^3}, \quad y\in[0,1],$$

is continuous for $y \in F$ and hence is $\mathcal{F}$-measurable.

Fix $y \in [0,1]$ and consider the section $f_y(x)$. For instance, if $y = 1/2$, then

$$f_{y=1/2}(x) = \frac{x(x^2 - 1/4)}{2(x^2 + 1/4)^3} \quad \text{for } x \in [0,2]$$

is continuous in x and so $f_{y=1/2}(x)$ is $\mathcal{E}$-measurable. Or, for instance, suppose that $y = 0$. Then $f_{y=0}(x) = 0$ for $x \in [0,2]$. In general, for fixed $y \in (0,1]$,

$$f_y(x) = \frac{xy(x^2 - y^2)}{(x^2 + y^2)^3}, \quad x \in [0,2],$$

is continuous for $x \in E$ and hence is $\mathcal{E}$-measurable.

We now calculate

$$\int_F f(x,y)Q(\mathrm{d}y) = \int_0^1 f(x,y)Q(\mathrm{d}y) = \int_0^1 f(x,y)\,\mathrm{d}y = \frac{x}{2(x^2+1)^2}$$

which follows since Q is the uniform probability on $[0,1]$. Hence,

$$\begin{aligned}\int_E \int_F f(x,y)Q(\mathrm{d}y)P(\mathrm{d}x) = \int_E \frac{x}{2(x^2+1)^2}P(\mathrm{d}x) = \int_0^2 \frac{x}{2(x^2+1)^2}\cdot\frac{1}{2}\,\mathrm{d}x = -\frac{1}{8}\left(\frac{1}{x^2+1}\right)\Bigg|_0^2 \\ = \frac{1}{10}\end{aligned}$$

which follows since P is the uniform probability on $[0,2]$.

On the other hand,

$$\int_E f(x,y)P(\mathrm{d}x) = \int_0^2 f(x,y)\cdot\frac{1}{2}\,\mathrm{d}x = -\frac{y}{2(4+y^2)^2}$$

and so

$$\begin{aligned}\int_F \int_E f(x,y)P(\mathrm{d}x)Q(\mathrm{d}y) = \int_F -\frac{y}{2(4+y^2)^2}Q(\mathrm{d}y) = -\int_0^1 \frac{y}{2(4+y^2)^2}\,\mathrm{d}y = \frac{1}{2}\left(\frac{1}{4+y^2}\right)\Bigg|_0^1 \\ = -\frac{1}{40}.\end{aligned}$$

The reason that

$$\int\int f(x,y)Q(\mathrm{d}y)P(\mathrm{d}x) \neq \int\int f(x,y)P(\mathrm{d}x)Q(\mathrm{d}y)$$

has to do with f itself. Notice that

$$f(2t,t) = \frac{6}{125t^2} \quad \text{and} \quad f(t,2t) = -\frac{6}{125t^2}.$$

In order for Fubini-Tonelli to work, it must be the case that f is either

(i) $\mathcal{E} \otimes \mathcal{F}$-measurable and positive, or

(ii) $\mathcal{E} \otimes \mathcal{F}$-measurable and $P \otimes Q$-integrable.

Since we have shown that the order of integration *does* matter for $f(x, y)$, we can conclude that f violates both (i) and (ii). In particular, f is neither positive nor $P \otimes Q$-integrable. As we noted above f *is* $\mathcal{E} \otimes \mathcal{F}$-measurable and so this example also shows that joint measurability alone is not sufficient for integrability.

Tail Events

Reference. Chapter 10, pages 71–72

Let $(\Omega, \mathcal{A}, P)$ be a probability space and let A_n be a sequence of events in $\mathcal{A}$. Define

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m = \lim_{n\to\infty} \bigcup_{m=n}^{\infty} A_m$$

which can be interpreted probabilistically as

$$\{A_n \text{ occur infinitely often}\} = \{A_n \text{ i.o.}\}.$$

Theorem (Borel-Cantelli)**.** *Let A_n be a sequence of events in $(\Omega, \mathcal{A}, P)$.*

(a) If

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty,$$

then $P\{A_n \text{ i.o.}\} = 0$.

(b) If $P\{A_n \text{ i.o.}\} = 0$ and A_n are mutually independent, then

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty.$$

Remark. An equivalent formulation of (b) is: "If A_n are mutually independent and if

$$\sum_{n=1}^{\infty} P\{A_n\} = \infty,$$

then $P\{A_n \text{ i.o.}\} = 1$."

Suppose that X_n are all random variables defined on $(\Omega, \mathcal{A}, P)$. Define the σ-algebras $\mathcal{B}_n = \sigma(X_n)$ and

$$\mathcal{C}_n = \sigma(X_n, X_{n+1}, X_{n+2}, \ldots) = \sigma\left(\bigcup_{p\geq n} \mathcal{B}_p\right).$$

Let

$$\mathcal{C}_\infty = \bigcap_{n=1}^{\infty} \mathcal{C}_n.$$

Definition. By Exercise 2.2, we have that $\mathcal{C}_\infty$ is a σ-algebra which we call the *tail σ-algebra (associated to the sequence $X_1, X_2, \ldots$).*

Theorem (Kolmogorov's Zero-One Law)**.** *If X_n are independent random variables and $C \in \mathcal{C}_\infty$, then either $P\{C\} = 0$ or $P\{C\} = 1$.*

Lecture #24: The Borel-Cantelli Lemma

Reference. Chapter 10, pages 71–72

Let $(\Omega, \mathcal{A}, P)$ be a probability space and let A_n be a sequence of events in $\mathcal{A}$. Define

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m = \lim_{n\to\infty} \bigcup_{m=n}^{\infty} A_m$$

which can be interpreted probabilistically as

$$\{A_n \text{ occur infinitely often}\} = \{A_n \text{ i.o.}\}.$$

Theorem (Borel-Cantelli)**.** *Let A_n be a sequence of events in $(\Omega, \mathcal{A}, P)$.*

(a) If

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty,$$

then $P\{A_n \text{ i.o.}\} = 0$.

(b) If $P\{A_n \text{ i.o.}\} = 0$ and A_n are mutually independent, then

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty.$$

Proof. (a) Suppose that $a_n = P\{A_n\} = \mathbb{E}(1_{A_n})$. By Theorem 9.2,

$$\sum_{n=1}^{\infty} P\{A_n\} = \sum_{n=1}^{\infty} a_n < \infty$$

implies

$$\sum_{n=1}^{\infty} 1_{A_n} < \infty \text{ a.s.}$$

However,

$$\sum_{n=1}^{\infty} 1_{A_n}(\omega) = \infty$$

if and only if

$$\omega \in \bigcap_{n=1}^{\infty} \bigcup_{m\geq n} A_m = \limsup_{n\to\infty} A_n.$$

Together these give

$$\sum_{n=1}^{\infty} P\{A_n\} < \infty \quad \Rightarrow \quad P\{A_n \text{ i.o.}\} = 0.$$

(b) Suppose that A_n are mutually independent. Then

$$\begin{aligned}
P\{\limsup_{n\to\infty} A_n\} &= P\left\{\bigcap_{n=1}^{\infty}\bigcup_{m\geq n} A_m\right\} \\
&= \lim_{n\to\infty}\lim_{k\to\infty} P\left\{\bigcup_{m=n}^{k} A_m\right\} \quad \text{by Theorem 2.4} \\
&= \lim_{n\to\infty}\lim_{k\to\infty}\left[1 - P\left\{\bigcap_{m=n}^{k} A_m^c\right\}\right] \quad \text{by DeMorgan's law} \\
&= 1 - \lim_{n\to\infty}\lim_{k\to\infty}\prod_{m=n}^{k}\left[1 - P\left\{A_m\right\}\right] \quad \text{by independence} \\
&= 1 - \lim_{n\to\infty}\lim_{k\to\infty}\prod_{m=n}^{k}(1 - a_m).
\end{aligned}$$

By hypothesis,

$$P\{\limsup_{n\to\infty} A_n\} = 0$$

so that

$$\lim_{n\to\infty}\lim_{k\to\infty}\prod_{m=n}^{k}(1 - a_m) = 1.$$

Taking logarithms gives

$$\lim_{n\to\infty}\lim_{k\to\infty}\sum_{m=n}^{k}\log(1 - a_m) = 0$$

which implies that

$$\lim_{n\to\infty}\sum_{m=n}^{\infty}\log(1 - a_m) = 0$$

and so we conclude that

$$\sum_{m=1}^{\infty}\log(1 - a_m)$$

converges. However,

$$|\log(1 - x)| \geq x \quad \text{for } 0 < x < 1$$

and so

$$\sum_{m=1}^{\infty}|\log(1 - a_m)| \geq \sum_{m=1}^{\infty} a_m$$

which implies that

$$\sum_{m=1}^{\infty} a_m = \sum_{m=1}^{\infty} P\{A_m\} < \infty$$

as required. □

Remark. An equivalent formulation of (b) is: "If A_n are mutually independent and if

$$\sum_{n=1}^{\infty} P\{A_n\} = \infty,$$

then $P\{A_n \text{ i.o.}\} = 1$."

Example. Suppose that $\Omega = [0,1]$ and P is the uniform probability measure on $[0,1]$. For $n = 1, 2, \ldots$, let

$$A_n = \left[0, \tfrac{1}{n}\right]$$

and note that

$$\bigcup_{m \geq n} A_m = \bigcup_{m \geq n} \left[0, \tfrac{1}{m}\right] = \left[0, \tfrac{1}{n}\right].$$

Therefore,

$$\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m = \bigcap_{n=1}^{\infty} \left[0, \tfrac{1}{n}\right] = \{0\}$$

which implies that

$$P\{A_n \text{ i.o.}\} = P\{0\} = 0.$$

We will now show that the A_n are not independent. Note that

$$P\{A_n\} = P\left\{\left[0, \tfrac{1}{n}\right]\right\} = \frac{1}{n}.$$

Consider A_2 and A_3. We find

$$P\{A_2 \cap A_3\} = P\left\{\left[0, \tfrac{1}{2}\right] \cap \left[0, \tfrac{1}{3}\right]\right\} = P\left\{\left[0, \tfrac{1}{3}\right]\right\} = \frac{1}{3}$$

although

$$P\{A_2\} \cdot P\{A_3\} = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

In general, if $k < j$, then $A_k \cap A_j = A_k$. It is curious to note, however, that A_1 is independent of A_j, $j > 1$, since

$$P\{A_1 \cap A_j\} = P\{A_j\} = \frac{1}{j}$$

and

$$P\{A_1\} \cdot P\{A_j\} = 1 \cdot \frac{1}{j} = \frac{1}{j}.$$

Nonetheless, $\{A_n, n = 1, 2, \ldots\}$ is not an independent sequence of events. Finally, we note that

$$\sum_{n=1}^{\infty} P\{A_n\} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

and so this gives a "counter-example" to part (b) of the Borel-Cantelli Lemma. In other words,

$$P\{A_n \text{ i.o.}\} = 0 \text{ and } A_n \text{ NOT independent does NOT imply } \sum_{n=1}^{\infty} P\{A_n\} < \infty.$$

Example. Let Ω and P be as in the previous example and for $n = 1, 2, \ldots$, let

$$A_n = \left[0, \tfrac{1}{n^2}\right].$$

As before, $P\{A_n \text{ i.o.}\} = 0$ and A_n are not independent. However, in this example we have

$$\sum_{n=1}^{\infty} P\{A_n\} = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Thus, we conclude that $P\{A_n \text{ i.o.}\} = 0$ gives no information about

$$\sum_{n=1}^{\infty} P\{A_n\}$$

in general.

Theorem (Kolmogorov's Zero-One Law)**.** *Suppose that $\{X_n, n = 1, 2, 3, \ldots\}$ are independent random variables and $\mathcal{C}_\infty$ is the tail σ-algebra generated by $X_1, X_2, \ldots$. If $C \in \mathcal{C}_\infty$, then either $P\{C\} = 0$ or $P\{C\} = 1$.*

Proof. Let $\mathcal{B}_n = \sigma(X_n)$ and $\mathcal{C}_n = \sigma(X_n, X_{n+1}, X_{n+2}, \ldots)$ so that $\mathcal{C}_\infty = \cap_n \mathcal{C}_n$. Define $\mathcal{D}_n = \sigma(X_1, \ldots, X_{n-1})$ so that $\mathcal{C}_n$ and $\mathcal{D}_n$ are independent σ-algebras (since the X_n are independent random variables). Thus, if $A \in \mathcal{C}_n$ and $B \in \mathcal{D}_n$, then

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

If $A \in \mathcal{C}_\infty$, then $(*)$ holds for all $B \in \cup_n \mathcal{D}_n$.

Hence, by the Monotone Class Theorem, $(*)$ holds for all $A \in \mathcal{C}_\infty$ and $B \in \sigma\left(\cup_n \mathcal{D}_n\right)$. But

$$\mathcal{C}_\infty \subsetneq \sigma\left(\bigcup_{n=1}^{\infty} \mathcal{D}_n\right)$$

so that $(*)$ holds for $A = B \in \mathcal{C}_\infty$. Therefore,

$$P\{A\} = P\{A \cap A\} = P\{A\} \cdot P\{A\}$$

which implies that $P\{A\} = P\{A\}^2$. This can be satisfied only if $P\{A\} = 0$ or $P\{A\} = 1$. $\square$

Lecture #25: Midterm Review

In preparation for the midterm next class, we will consider the following exercises.

Exercise. Suppose that X and Y are random variables with $X = Y$ almost surely. Prove $\mathbb{E}(X) = \mathbb{E}(Y)$.

Exercise. Suppose that $\Omega = \{1, 2, 3, \ldots\}$. Let $\mathcal{A}$ denote the collection of subsets $A \subseteq \Omega$ such that the limit

$$\lim_{n\to\infty} \frac{\#(A \cap \{1, \ldots, n\})}{n}$$

exists. For $A \in \mathcal{A}$, define

$$P(A) = \lim_{n\to\infty} \frac{\#(A \cap \{1, \ldots, n\})}{n}.$$

Prove that $(\Omega, \mathcal{A}, P)$ is not a probability space.

Lecture #26: Convergence Almost Surely and Convergence in Probability

Reference. Chapter 17, pages 141–147

Suppose that X_n, $n = 1, 2, 3, \ldots$, and X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$.

We say that the sequence X_n *converges pointwise* to X if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \text{for every } \omega \in \Omega.$$

In other words, X_n converges pointwise to X if

$$\{\omega : X_n(\omega) \to X(\omega)\} = \Omega.$$

It turns out that convergence pointwise is useless for probability because there are often "impossible events" for which $X_n(\omega) \not\to X(\omega)$.

A much more useful notion is that of almost sure convergence. We say that X_n *converges almost surely* to X if

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \text{for almost every } \omega \in \Omega.$$

In other words, X_n converges almost surely to X if

$$P\{\omega : X_n(\omega) \to X(\omega)\} = 1.$$

Note that almost sure convergence allows for there to exist some ω for which $X_n(\omega) \not\to X(\omega)$. However, these ω belong to a null set and so they are "impossible."

Example. Flip a fair coin repeatedly. Let $X_n = 1$ if the nth flip is heads and 0 otherwise so that

$$P\{X_n = 1\} = P\{X_n = 0\} = \frac{1}{2}.$$

We would like to say that the long run average of heads approaches 1/2. That is,

$$\left\{\omega : \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = \frac{1}{2}\right\} = \Omega.$$

But, if we let $\omega_0 = (T, T, T, T, \ldots)$, then $X_j(\omega_0) = 0$ for all j so that

$$\frac{X_1(\omega_0) + \cdots + X_n(\omega_0)}{n} = \frac{0 + \cdots + 0}{n} = 0$$

which implies

$$\lim_{n\to\infty} \frac{X_1(\omega_0) + \cdots + X_n(\omega_0)}{n} = 0 \neq \frac{1}{2}.$$

This shows that

$$\left\{\omega : \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = \frac{1}{2}\right\} \neq \Omega.$$

However, we will prove as a consequence of the Strong Law of Large Numbers that

$$P\left\{\omega : \lim_{n\to\infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = \frac{1}{2}\right\} = 1.$$

In other words, the long run average of heads approaches 1/2 **almost surely**.

Example. Suppose that $X_1, X_2, \ldots$ are independent random variables with

$$P\{X_n = 1\} = \frac{1}{n} \quad \text{and} \quad P\{X_n = 0\} = 1 - \frac{1}{n}.$$

If we imagine X_n as the indicator of the event that heads is flipped on the nth toss of a coin, then we see that as n increases, the probability of heads goes to 0. Thus, we are tempted to say that

$$\lim_{n\to\infty} X_n = 0 \quad \text{or, in other words,} \quad X_n \to 0 \text{ a.s.}$$

However, this is not true. To see this, let A_n be the event that the nth flip is heads so that $P\{A_n\} = 1/n$ and let $X_n = 1_{A_n}$. Since the A_n are independent with

$$\sum_{n=1}^{\infty} P\{A_n\} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

we conclude from part (b) of the Borel-Cantelli Lemma that

$$P\{A_n \text{ i.o.}\} = 1.$$

In other words, "we will see heads appear infinitely often" and so

$$P\left\{\omega : \limsup_{n\to\infty} X_n(\omega) = 1\right\} = 1. \qquad (*)$$

Since A_n^c is the event that the nth flip is tails, and since the A_n^c are also independent and satisfy

$$\sum_{n=1}^{\infty} P\{A_n^c\} = \sum_{n=1}^{\infty} \frac{n-1}{n} = \infty,$$

we conclude from part (b) of the Borel-Cantelli Lemma that

$$P\{A_n^c \text{ i.o.}\} = 1.$$

In other words, "we will see tails appear infinitely often" and so

$$P\left\{\omega : \liminf_{n\to\infty} X_n(\omega) = 0\right\} = 1. \qquad (**)$$

In particular, $(*)$ and $(**)$ show that

$$P\left\{\omega : \lim_{n\to\infty} X_n(\omega) \text{ does not exist}\right\} = 1$$

which implies that X_n does not converge almost surely to 0 (or to anything else).

This example seems troublesome since X_n should really go to zero! Fortunately, there is another notion of convergence that can help us out here.

Definition. Suppose that X_n, $n = 1, 2, \ldots$, and X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. We say that X_n *converges in probability* to X if

$$\lim_{n\to\infty} P\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\} = 0 \quad \text{for every } \epsilon > 0.$$

Remark. One of the standard ways to show convergence in probability is to use Markov's inequality which states that

$$P\{|Y| > a\} \leq \frac{\mathbb{E}|Y|}{a}$$

for every $a > 0$ (assuming $Y \in L^1$).

Example (continued). We see that

$$\mathbb{E}|X_n| = \mathbb{E}(X_n) = 1 \cdot P\{X_n = 1\} + 0 \cdot P\{X_n = 0\} = \frac{1}{n}.$$

Therefore, if $\epsilon > 0$, then

$$P\{|X_n - 0| > \epsilon\} \leq \frac{\mathbb{E}|X_n|}{\epsilon} = \frac{1}{n\epsilon}$$

and so

$$\lim_{n\to\infty} P\{|X_n - 0| > \epsilon\} \leq \lim_{n\to\infty} \frac{1}{n\epsilon} = 0.$$

Thus, $X_n \to 0$ in probability.

Example. Consider the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = [0, 1]$ with the σ-algebra $\mathcal{A}$ equal to the Borel sets in $[0, 1]$ and P is the uniform probability measure on $[0, 1]$. For $n = 1, 2, \ldots$, define the random variable

$$X_n(\omega) = n^{1/2} 1_{(0, \frac{1}{n}]}(\omega).$$

Show that X_n converges in probability.

Solution. Note that if X_n converges in probability, then it must be to 0. We find

$$\mathbb{E}|X_n| = \mathbb{E}(X_n) = n^{1/2} P\left\{\left(0, \tfrac{1}{n}\right]\right\} = n^{1/2} \cdot \frac{1}{n} = \frac{1}{\sqrt{n}}.$$

Therefore, if $\epsilon > 0$, then

$$P\{|X_n - 0| > \epsilon\} \leq \frac{\mathbb{E}|X_n|}{\epsilon} = \frac{1}{\epsilon\sqrt{n}}$$

and so

$$\lim_{n\to\infty} P\{|X_n - 0| > \epsilon\} \leq \lim_{n\to\infty} \frac{1}{\epsilon\sqrt{n}} = 0.$$

Thus, $X_n \to 0$ in probability.

Statistics 851 (Winter 2008) March 17, 2008
Prof. Michael Kozdron

Lecture #27: Convergence of Random Variables

Reference. Chapter 17, pages 141–147

Suppose that X_n, $n = 1, 2, 3, \ldots$, and X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$.

Definition. We say that X_n *converges almost surely* to X if

$$P\{\omega : X_n(\omega) \to X(\omega)\} = 1.$$

In other words, $X_n \to X$ a.s. if

$$N = \{\omega : X_n(\omega) \not\to X(\omega)\}$$

is a null set. Note that

$$\{\omega : X_n(\omega) \to X(\omega)\} = \left\{\omega : \bigcap_{\epsilon>0} \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} |X_m(\omega) - X(\omega)| \leq \epsilon\right\}$$

which says that $X_n \to X$ almost surely iff "for every $\epsilon > 0$ there exists an n such that $|X_m(\omega) - X(\omega)| \leq \epsilon$ for all $m \geq n$." Taking complements gives

$$N = \{\omega : X_n(\omega) \not\to X(\omega)\} = \left\{\omega : \bigcup_{\epsilon>0} \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} |X_m(\omega) - X(\omega)| > \epsilon\right\}.$$

Definition. We say that X_n *converges in probability* to X if

$$\lim_{n\to\infty} P\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\} = 0 \quad \text{for every } \epsilon > 0.$$

Recall. We saw an example last lecture of a sequence of random variables X_n which converged to 0 in probability but not almost surely. The following theorem shows that convergence almost surely always implies convergence in probability.

Theorem. *If $X_n \to X$ almost surely, then $X_n \to X$ in probability.*

Proof. Suppose that $X_n \to X$ almost surely so that

$$P\left\{\omega : \bigcup_{\epsilon>0} \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} |X_m(\omega) - X(\omega)| > \epsilon\right\} = 0$$

which is equivalent to saying that $X_n \to X$ almost surely if for every $\epsilon > 0$,

$$P\left\{\omega : \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} |X_m(\omega) - X(\omega)| > \epsilon\right\} = 0.$$

Let $A_m = \{|X_m - X| > \epsilon\}$ and let

$$B_n = \bigcup_{m \geq n} A_m$$

so that $X_n \to X$ almost surely iff for every $\epsilon > 0$,

$$P\left\{\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} |X_m - X| > \epsilon\right\} = P\left\{\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\right\} = P\left\{\bigcap_{n=1}^{\infty} B_n\right\} = 0.$$

Note that B_n is a decreasing sequence ($B_n \supseteq B_{n+1}$) so that by Theorem 2.3,

$$P\left\{\bigcap_{n=1}^{\infty} B_n\right\} = \lim_{n\to\infty} P\{B_n\} = \lim_{n\to\infty} P\left\{\bigcup_{m\geq n} A_m\right\} = 0.$$

In other words, we see that $X_n \to X$ almost surely iff for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|X_m - X| > \epsilon \text{ for some } m \geq n\} = 0.$$

Thus, it is now clear that almost sure convergence is stronger than convergence in probability. That is,

$$\lim_{n\to\infty} P\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\} \leq \lim_{n\to\infty} P\{|X_m - X| > \epsilon \text{ for some } m \geq n\}$$

which shows that if $X_n \to X$ almost surely, then $X_n \to X$ in probability. □

Another useful convergence concept is that of convergence in L^p.

Definition. Let $1 \leq p < \infty$. We say that X_n *converges in* L^p to X if $|X_n| \in L^p$, $|X| \in L^p$, and

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0.$$

Theorem. *If $X_n \to X$ in L^p for any $p \in [1, \infty)$, then $X_n \to X$ in probability.*

Proof. Assume that $|X_n| \in L^p$, $|X| \in L^p$, and $\mathbb{E}(|X_n - X|^p) \to 0$. For any $\epsilon > 0$, notice that

$$|X_n - X|^p \geq \epsilon^p 1_{\{|X_n - X| > \epsilon\}}.$$

Taking expectations (i.e., Markov's inequality) gives

$$\mathbb{E}(|X_n - X|^p) \geq \epsilon^p P\{|X_n - X| > \epsilon\}.$$

Therefore, for every $\epsilon > 0$.

$$\lim_{n\to\infty} P\{|X_n - X| > \epsilon\} \leq \frac{1}{\epsilon^p} \lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$$

so that $X_n \to X$ in probability. □

Example. Consider the probability space $(\Omega, \mathcal{A}, P)$ where $\Omega = [0,1]$ with the σ-algebra $\mathcal{A}$ equal to the Borel sets in $[0,1]$ and P is the uniform probability measure on $[0,1]$. For $n = 1, 2, \ldots$, define the random variable

$$X_n(\omega) = n^{1/2} 1_{(0,\frac{1}{n}]}(\omega).$$

Last class we showed that $X_n \to 0$ in probability. We now show that X_n does not converge in L^p for $p \geq 2$.

Solution. For fixed n, we see that

$$\mathbb{E}(|X_n|^p) = \mathbb{E}(X_n^p) = n^{p/2} \cdot \frac{1}{n} = n^{p/2-1} < \infty,$$

and so we conclude that $|X_n| \in L^p$ for all $p \geq 1$. However,

$$\lim_{n\to\infty} \mathbb{E}(|X_n - 0|^p) = 0 \quad \text{if and only if} \quad \frac{p}{2} - 1 < 0$$

which implies that $X_n \not\to 0$ in L^p for $p \geq 2$.

Remark. You will see an example on the homework which shows that $X_n \to 0$ in probability does not imply $X_n \to 0$ in L^p for any $p \geq 1$.

Definition. We say that X_n *converges completely* to X if for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P\{|X_n - X| > \epsilon\} < \infty.$$

Theorem. *If $X_n \to X$ completely, then $X_n \to X$ almost surely.*

Proof. Suppose that $X_n \to X$ completely so that

$$\sum_{n=1}^{\infty} P\{|X_n - X| > \epsilon\} < \infty.$$

If we let $A_n = \{|X_n - X| > \epsilon\}$, then part (a) of the Borel-Cantelli lemma implies that

$$P\{A_n \text{ i.o.}\} = 0$$

which means that $X_n \to X$ almost surely. □

Summary of Modes of Convergence

Suppose that X_n, $n = 1, 2, \ldots$, X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. The facts about convergence are:

$$\begin{array}{lcl} X_n \to X \text{ completely} & \Rightarrow & X_n \to X \text{ almost surely} \\ & & \Downarrow \\ & & X_n \to X \text{ in probability} \\ & & \Uparrow \\ & & X_n \to X \text{ in } L^p \end{array}$$

with no other implications holding in general.

Remark. We have proved all of the implications. Furthermore, we have shown that all of the implications are strict except for complete and almost sure convergence. Since Exercise 10.13 on page 73 shows that if $X_1, X_2, \ldots$ are independent, then $X_n \to X$ completely is equivalent to $X_n \to X$ almost surely, we see that it will be a little involved to construct a sequence of random variables which converge almost surely but not completely.

Lecture #28: Convergence of Random Variables

Reference. Chapter 17, pages 141–147

Recall. Suppose that X_n, $n = 1, 2, \ldots$, and X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. The facts about convergence are:

$$\begin{array}{rcl} X_n \to X \text{ completely} & \Rightarrow & X_n \to X \text{ almost surely} \\ & & \Downarrow \\ & & X_n \to X \text{ in probability} \\ & & \Uparrow \\ & & X_n \to X \text{ in } L^p \end{array}$$

with no other implications holding in general.

We have already proved the three implications and have given examples to show that two of the converse implications do not hold. All that remains is to provide an example to show that convergence almost surely does not imply complete convergence. The following example accomplishes this and is a slight modification of an example considered in Lecture #24.

Example. Suppose that $\Omega = [0, 1]$ and P is the uniform probability measure on $[0, 1]$. For $n = 1, 2, \ldots$, let

$$A_n = \left[0, \tfrac{1}{n}\right]$$

and note that

$$\bigcup_{m \geq n} A_m = \bigcup_{m \geq n} \left[0, \tfrac{1}{m}\right] = \left[0, \tfrac{1}{n}\right].$$

Therefore,

$$P\{A_n \text{ i.o.}\} = P\left\{\bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m\right\} = P\left\{\bigcap_{n=1}^{\infty} \left[0, \tfrac{1}{n}\right]\right\} = P\{0\} = 0.$$

We now define the random variable X_n by setting

$$X_n(\omega) = n \cdot 1_{A_n}(\omega).$$

Suppose that $\epsilon > 0$ is arbitrary (and also suppose that $\epsilon < 1$ to make things clearer). It then follows that

$$|X_n(\omega)| > \epsilon \quad \text{iff} \quad X_n(\omega) = n \quad \text{iff} \quad \omega \in A_n.$$

That is,

$$\{|X_n| > \epsilon \text{ i.o.}\} = \{A_n \text{ i.o.}\}$$

and so

$$P\{|X_n| > \epsilon \text{ i.o.}\} = 0$$

which implies that $X_n \to 0$ almost surely. However, we see that

$$\sum_{n=1}^{\infty} P\{|X_n| > \epsilon\} = \sum_{n=1}^{\infty} P\{A_n\} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

so that X_n does not converge completely to 0.

Remark. If you are uncomfortable with assuming $0 < \epsilon < 1$ (although you should not be), just note that if $\epsilon > 0$ is fixed, say $\epsilon = K$, then we only need to consider X_n, $n > K$, since $\{A_n,\, n \geq 1,\, \text{i.o.}\} = \{A_n,\, n > K,\, \text{i.o.}\}$. That is, a finite number of the A_n do not affect whether or not A_n happens infinitely often.

Example. If we consider the random variable $X_n = n1_{A_n}$ as in the previous example, then for each $n = 1, 2, \ldots$, we have $|X_n| \in L^p$ for every $1 \leq p < \infty$ since

$$\mathbb{E}|X_n|^p = n^p P\{A_n\} = n^{p-1} < \infty.$$

However, if $p = 1$, then

$$\lim_{n\to\infty} \mathbb{E}|X_n - 0| = \lim_{n\to\infty} 1 = 1$$

which shows that X_n does not converge to 0 in L^1. If $p > 1$, then

$$\lim_{n\to\infty} \mathbb{E}|X_n - 0|^p = \lim_{n\to\infty} n^{p-1} = \infty$$

which shows that X_n does not converge to 0 in L^p, $1 < p < \infty$. In particular, this example shows that convergence almost surely does not necessarily imply convergence in L^p for any $1 \leq p < \infty$.

Theorem. *Suppose f is continuous.*

(a) If $X_n \to X$ almost surely, then $f(X_n) \to f(X)$ almost surely.

(b) If $X_n \to X$ in probability, then $f(X_n) \to f(X)$ in probability.

Proof. To prove (a), suppose that N is the null set on which $X_n(\omega) \not\to X(\omega)$. For $\omega \notin N$, we have

$$\lim_{n\to\infty} f(X_n(\omega)) = f\left(\lim_{n\to\infty} X_n(\omega)\right) = f(X(\omega))$$

where the first equality follows from the fact that f is continuous. Since this is true for any $\omega \notin N$ and $P\{N\} = 0$, it follows that $f(X_n) \to f(X)$ almost surely.

For the proof of (b), see page 147. □

Converses to the Implications

We have already shown the "facts about convergence" by demonstrating that three implications hold with no others holding in general. We now prove a number of "almost"-converses; that is, converse implications which hold under additional hypotheses.

Theorem. *If $X_n \to X$ almost surely and the X_n are independent, then $X_n \to X$ completely.*

Proof. Assume that $X_n \to X$ almost surely and the X_n are independent. Let $\epsilon > 0$ and let $A_n = \{|X_n| > \epsilon\}$ so that the A_n are independent and

$$P\{A_n \text{ i.o.}\} = 0.$$

Part (b) of the Borel-Cantelli lemma now applies and we conclude

$$\sum_{n=1}^{\infty} P\{A_n\} = \sum_{n=1}^{\infty} P\{|X_n| > \epsilon\} < \infty.$$

In other words, $X_n \to X$ completely. □

Theorem. *If $X_n \to X$ in probability, then there exists some subsequence n_k such that $X_{n_k} \to X$ almost surely.*

Proof. Let $n_0 = 0$ and for every integer $k \geq 1$ let

$$n_k = \inf\left\{n > n_{k-1} : P\left\{\omega : |X_n(\omega) - X(\omega)| > \frac{1}{k}\right\} \leq \frac{1}{2^k}\right\}.$$

For every $\epsilon > 0$, we have $\frac{1}{k} < \epsilon$ eventually (i.e., for $k > k_0$ where $k_0 = \lceil 1/\epsilon \rceil$). Hence, for $m > k_0$,

$$\begin{aligned} P\left\{\bigcup_{k=m}^{\infty} \{\omega : |X_{n_k}(\omega) - X(\omega)| > \epsilon\}\right\} &\leq P\left\{\bigcup_{k=m}^{\infty} \left\{\omega : |X_{n_k}(\omega) - X(\omega)| > \frac{1}{k}\right\}\right\} \\ &\leq \sum_{k=m}^{\infty} P\left\{\omega : |X_{n_k}(\omega) - X(\omega)| > \frac{1}{k}\right\} \\ &\leq \sum_{k=m}^{\infty} 2^{-k} \\ &= 2^{1-m} \to 0 \text{ as } m \to \infty. \end{aligned}$$

In other words, $X_{n_k} \to X$ almost surely. □

Theorem. *If $X_n \to X$ in probability and there exists some $Y \in L^p$ such that $|X_n| \leq Y$ for all n, then $|X| \in L^p$ and $X_n \to X$ in L^p.*

Proof. If $Y \in L^p$ and $|X_n| \leq Y$, then $\mathbb{E}|X_n|^p \leq \mathbb{E}(Y^p) < \infty$ so that $|X_n| \in L^p$ as well. Let $\epsilon > 0$. Notice that for each n

$$P\{|X| > Y + \epsilon\} \leq P\{|X| > |X_n| + \epsilon\} = P\{|X| - |X_n| > \epsilon\} \leq P\{|X - X_n| > \epsilon\}$$

and so

$$P\{|X| > Y + \epsilon\} \leq \lim_{n\to\infty} P\{|X - X_n| > \epsilon\} = 0$$

by the assumption that $X_n \to X$ in probability. Since $\epsilon > 0$ is arbitrary, it follows that (taking $\epsilon = 1/m$)

$$P\{|X| > Y\} \leq \lim_{m\to\infty} P\left\{|X| > Y + \frac{1}{m}\right\} = 0$$

which implies that $|X| \leq Y$ almost surely. Hence, $|X| \in L^p$.

To show that $X_n \to X$ in L^p we will derive a contradiction. Suppose not. There must then exist a subsequence n_k with

$$\mathbb{E}(|X_{n_k} - X|^p) \geq \epsilon$$

for all k and for some $\epsilon > 0$. Since $X_n \to X$ in probability, it trivially holds that $X_{n_k} \to X$ in probability. By the previous theorem, there exists a further subsequence n_{k_j} such that $X_{n_{k_j}} \to X$ almost surely as $j \to \infty$. Since $X_{n_{k_j}} - X \to 0$ almost surely as $j \to \infty$ and $|X_{n_{k_j}} - X| \leq 2Y$ we can apply the Dominated Convergence Theorem to conclude

$$\mathbb{E}(|X_{n_{k_j}} - X|^p) \to 0 \quad \text{as } j \to \infty.$$

This contradicts the assumption that $\mathbb{E}(|X_{n_k} - X|^p) \geq \epsilon$. □

Alternate proof that almost sure convergence implies convergence in probability

Lemma. *$X_n \to X$ in probability if and only if*

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = 0.$$

Proof. Suppose that $X_n \to X$ in probability so that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\{|X_n - X| > \epsilon\} = 0.$$

We now note that

$$|X_n - X| \leq 1 + |X_n - X|$$

which implies that

$$\frac{|X_n - X|}{1 + |X_n - X|} \leq 1$$

and so

$$\frac{|X_n - X|}{1 + |X_n - X|} \leq \frac{|X_n - X|}{1 + |X_n - X|} 1_{\{|X_n - X| > \epsilon\}} + \epsilon 1_{\{|X_n - X| \leq \epsilon\}} \leq 1_{\{|X_n - X| > \epsilon\}} + \epsilon.$$

Taking expectations gives

$$\mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) \leq P\{|X_n - X| > \epsilon\} + \epsilon$$

and so

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) \leq \lim_{n\to\infty} P\{|X_n - X| > \epsilon\} + \epsilon = 0 + \epsilon = \epsilon.$$

Since $\epsilon > 0$ was arbitrary, we have

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = 0.$$

Conversely, suppose that

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = 0$$

and let $\epsilon > 0$ be given. We now note that

$$\frac{\epsilon}{1+\epsilon} 1_{\{|X_n - X| > \epsilon\}} \leq \frac{|X_n - X|}{1 + |X_n - X|} 1_{\{|X_n - X| > \epsilon\}} \leq \frac{|X_n - X|}{1 + |X_n - X|}$$

which relied on the fact that $x/(1+x)$ is strictly increasing. Taking expectations and limits yields

$$\frac{\epsilon}{1+\epsilon} \lim_{n\to\infty} P\{|X_n - X| > \epsilon\} \leq \lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = 0.$$

Since $\epsilon > 0$ is given, we have

$$\lim_{n\to\infty} P\{|X_n - X| > \epsilon\} = 0$$

and so $X_n \to X$ in probability. □

Theorem. *If $X_n \to X$ almost surely, then $X_n \to X$ in probability.*

Proof. As noted in the proof of the lemma,

$$\frac{|X_n - X|}{1 + |X_n - X|} \leq 1.$$

Since $X_n \to X$ almost surely, we see that

$$\frac{|X_n - X|}{1 + |X_n - X|} \to 0$$

almost surely. We can now apply the Dominated Convergence Theorem to conclude

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = \mathbb{E}\left(\lim_{n\to\infty} \frac{|X_n - X|}{1 + |X_n - X|}\right) = \mathbb{E}(0) = 0.$$

By the lemma, $X_n \to X$ in probability. □

Lecture #29: Weak Convergence

Reference. Chapter 18, pages 151–163

There is a fifth mode of convergence of random variables which is radically different from the four we studied previously. It is *convergence in distribution* which is also known as *weak convergence* or *convergence in law*.

As the name suggests, it is the weakest of all modes of convergence. Whereas the other four modes require that the random variables are all defined on the same probability space, weak convergence concerns the laws of the random variables.

Recall that if $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ is a random variable, then P^X, the law of X, is given by

$$P^X(B) = P\{X \in B\} \quad \text{for every } B \in \mathcal{B}$$

and defines a probability measure on $(\mathbb{R}, \mathcal{B})$. That is, if X is a random variable, then $(\mathbb{R}, \mathcal{B}, P^X)$ is a probability space.

Now suppose that $X_1, X_2, X_3, \ldots$, and X are all random variables. Except we now no longer need to assume that they are all defined on a common probability space. Suppose that for each n, the random variable $X_n : (\Omega_n, \mathcal{A}_n, P_n) \to (\mathbb{R}, \mathcal{B})$ and $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$.

The laws of each of these random variables induce a probability measure on $(\mathbb{R}, \mathcal{B})$, say P^{X_n}, $n = 1, 2, 3, \ldots$, and P^X.

Definition. We say that X_n *converges weakly to* X (or that X_n *converges in distribution to* X or that X_n *converges in law to* X) if

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x) P^{X_n}(\mathrm{d}x) = \int_{\mathbb{R}} f(x) P^X(\mathrm{d}x)$$

for every bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$.

Remark. This definition is rather unintuitive; it is not very easy to understand what convergence in distribution means. Fortunately, we will prove a number of characterizations of weak convergence which will be, in general, much easier to handle.

Remark. The reason that this is called convergence in law should be clear. We are discussing convergence of the laws of random variables. The reason that it is also called convergence in distribution is the following. Since the law of a random variable is characterized by its distribution function, it should seem clear that convergence of the laws of the random variables is equivalent to convergence of the corresponding distribution functions. This is indeed the case, although there is a technical matter that needs to be addressed. For now we will be content with the following *incorrect* definition.

We say that $X_n \to X$ in distribution if

$$F_{X_n}(x) \to F_X(x) \quad \text{for every } x \in \mathbb{R}.$$

Although this definition is not quite correct, it does explain why weak convergence is also called convergence in distribution.

Remark. As we have already seen, there are differences in terminology between analysis and probability for the same concept. The term *weak convergence* comes from two closely related concepts in functional analysis known as *weak-$*$ convergence* and *vague convergence*.

The first result shows that weak convergence is equivalently phrased in terms of expectations.

Theorem. *Let X_n, $n = 1, 2, 3, \ldots$, be random variables with laws P^{X_n} and let X be a random variable with law P^X. The random variables $X_n \to X$ weakly if and only if*

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

for every bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$.

Proof. By Theorem 9.5, if X_n has distribution P^{X_n} and X has distribution P^X, then

$$\int_{\mathbb{R}} f(x) P^{X_n}(\mathrm{d}x) = \mathbb{E}(f(X_n)) \quad \text{and} \quad \int_{\mathbb{R}} f(x) P^{X}(\mathrm{d}x) = \mathbb{E}(f(X)).$$

This equivalence establishes the theorem. □

The contrapositive of this statement is sometimes useful for showing when random variables do not converge weakly.

Corollary. *If there exists some bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$ such that*

$$\mathbb{E}(f(X_n)) \not\to \mathbb{E}(f(X)),$$

then $X_n \not\to X$ weakly.

Note. Since $f(x) = x$ is NOT bounded, we cannot infer whether or not $X_n \to X$ weakly from the convergence/non-convergence of $\mathbb{E}(X_n)$ to $\mathbb{E}(X)$.

Although the random variables X_n and X need not be defined on the same probability space in order for weak convergence to make sense, if they are defined on a common probability space, then convergence in probability implies convergence in distribution.

Fact. Suppose that X_n, $n = 1, 2, \ldots$, and X are random variables defined on a common probability space $(\Omega, \mathcal{A}, P)$. The facts about convergence are:

$$\begin{array}{lcl} X_n \to X \text{ completely} & \Rightarrow & X_n \to X \text{ almost surely} \\ & & \Downarrow \\ & & X_n \to X \text{ in probability} \quad \Rightarrow \quad X_n \to X \text{ weakly} \\ & & \Uparrow \\ & & X_n \to X \text{ in } L^p \end{array}$$

with no other implications holding in general.

The following example shows just how weak convergence in distribution really is.

Example. Suppose that $\Omega = \{a, b\}$, $\mathcal{A} = 2^\Omega$, and $P\{a\} = P\{b\} = 1/2$. Define the random variables Y and Z by setting

$$Y(a) = 1, \quad Y(b) = 0, \quad \text{and} \quad Z = 1 - Y$$

so that $Z(a) = 0$, $Z(b) = 1$. It is clear that $P\{Y = Z\} = 0$; that is, Y and Z are almost surely unequal. However, Y and Z have the same distribution since

$$P\{Y = 1\} = P\{Z = 0\} = P\{a\} = \frac{1}{2} \quad \text{and} \quad P\{Y = 0\} = P\{Z = 1\} = P\{b\} = \frac{1}{2}.$$

In other words,

$$Y \neq Z \text{ (surely and almost surely)} \quad \text{but} \quad P^Y = P^Z.$$

We can now extend this example to construct a sequence of random variables which converges in distribution but does not converge in probability.

Example (continued). Suppose that the random variable X is defined by $X(a) = 1$, $X(b) = 0$ independently of Y and Z so that $P^X = P^Y = P^Z$. If n is odd, let $X_n = Y$, and if n is even, let $X_n = Z$. Since $P^Y = P^Z = P^X$, it is clear that $P^{X_n} \to P^X$ meaning that $X_n \to X$ in distribution. (In fact, $P^{X_n} = P^X$ meaning that $X_n = X$ in distribution.) However, if $\epsilon > 0$ and n is even, then

$$P\{|X_n - X| > \epsilon\} = P\{|Z - X| > \epsilon\} = P\{Z = 1, X = 0\} + P\{Z = 0, X = 1\} = \frac{1}{2}.$$

Thus, it is not possible for X_n to converge in probability to X. Of course, this example should make sense. By construction, since $Z = 1 - Y$, the observed sequence of X_n will alternate between 1 and 0 or 0 and 1 (depending on whether a or b is first observed).

Example. One of the key theorems of probability and statistics is the Central Limit Theorem. This theorem states that if X_n, $n = 1, 2, 3, \ldots$, are independent and identically distributed L^2 random variables with common mean μ and common variance σ^2, and $S_n = X_1 + \cdots + X_n$, then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \to Z \quad \text{as } n \to \infty$$

weakly where Z is normally distributed with mean 0 and variance 1. In other words, the Central Limit Theorem is a statement about convergence in distribution. As the previous example shows, sometimes we can only achieve results about weak convergence and we must be content in doing so.

Theorem. *If X_n, $n = 1, 2, 3, \ldots$, and X are all defined on a common probability space and if $X_n \to X$ in probability, then $X_n \to X$ weakly.*

Proof. Suppose that $X_n \to X$ in probability and let f be bounded and continuous. Since f is continuous, Theorem 17.5 implies

$$f(X_n) \to f(X) \quad \text{in probability}$$

and Theorem 17.4 implies

$$f(X_n) \to f(X) \quad \text{in } L^1.$$

This means that

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$$

which, by the previous theorem, is equivalent to $X_n \to X$ weakly. □

Example. Let $\Omega = [0,1]$ and let P be the uniform probability measure on Ω. Define the random variables X_n by

$$X_n(\omega) = n \cdot 1_{[0,\frac{1}{n}]}(\omega).$$

We showed last lecture that $X_n \to 0$ in probability. Thus, $X_n \to 0$ in distribution.

Example. Suppose that X_n, $n = 1, 2, 3, \ldots$, has density function

$$f_n(x) = \frac{n}{\pi(1 + n^2x^2)}, \quad -\infty < x < \infty.$$

If X_n are all defined on the same probability space and $\epsilon > 0$, then

$$\begin{aligned} P\{|X_n| > \epsilon\} &= 1 - \int_{-\epsilon}^{\epsilon} f_n(x)\,\mathrm{d}x \\ &= 1 - \int_{-\epsilon}^{\epsilon} \frac{n}{\pi(1 + n^2x^2)}\,\mathrm{d}x \\ &= 1 - \frac{2}{\pi}\arctan(n\epsilon) \\ &\to 1 - \frac{2}{\pi} \cdot \frac{\pi}{2} \quad \text{as } n \to \infty. \end{aligned}$$

In other words,

$$\lim_{n\to\infty} P\{|X_n| > \epsilon\} = 0$$

so that $X_n \to 0$ in probability. Thus, $X_n \to 0$ in distribution. Note that this example does not work if X_n are NOT defined on a common probability space.

Remark. Note that in the previous example, each of the random variables X_n had a density. However, the limiting random variable $X = 0$ does not have a density. If $x \neq 0$, then

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{n}{\pi(1 + n^2x^2)} = \lim_{n\to\infty} \frac{1}{\pi/n + n\pi x^2} = 0.$$

But, if $x = 0$, then

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \frac{n}{\pi} = \infty.$$

That is, the limiting "density function" satisfies

$$f(x) = \begin{cases} 0, & \text{if } x \neq 0, \\ \infty, & \text{if } x = 0. \end{cases}$$

which is not really a function at all. In fact, f is known as a generalized function or a δ-function.

The following theorem gives a sufficient condition for weak convergence in terms of density functions, but as the previous example shows, it is not a necessary condition.

Theorem. *Suppose that X_n, $n = 1, 2, 3, \ldots$, and X are real-valued random variables and have density functions f_n, f, respectively. If $f_n \to f$ pointwise, then $X_n \to X$ in distribution.*

Proof. Since there might be some confusion about the use of f in the statement of weak convergence, assume that g is bounded and continuous. In order to show that $X_n \to X$ weakly, we must show that

$$\lim_{n\to\infty} \int_{\mathbb{R}} g(x) P^{X_n}(\mathrm{d}x) = \int_{\mathbb{R}} g(x) P^{X}(\mathrm{d}x).$$

Since X_n has density function $f_n(x)$, we can write

$$\int_{\mathbb{R}} g(x) P^{X_n}(\mathrm{d}x) = \int_{\mathbb{R}} g(x) f_n(x)\,\mathrm{d}x.$$

Since $f_n \to f$ pointwise, we would like to say that

$$\int_{\mathbb{R}} g(x) f_n(x)\,\mathrm{d}x \to \int_{\mathbb{R}} g(x) f(x)\,\mathrm{d}x. \qquad (*)$$

Assuming $(*)$ is true, we use the fact that $f(x)$ is the density function for X to conclude that

$$\int_{\mathbb{R}} g(x) f(x)\,\mathrm{d}x = \int_{\mathbb{R}} g(x) P^{X}(\mathrm{d}x).$$

In other words, assuming $(*)$ holds,

$$\int_{\mathbb{R}} g(x) P^{X_n}(\mathrm{d}x) \to \int_{\mathbb{R}} g(x) P^{X}(\mathrm{d}x)$$

as required.

The problem now is to establish $(*)$. This is not as obvious as it may seem since pointwise convergence does not necessarily allow us to interchange limits and integrals. We saw a counterexample on a handout from February 11, 2008. However, this time the difference is that f_n, f are density functions and not arbitrary functions. To begin, note that if $g : \mathbb{R} \to \mathbb{R}$ is bounded and continuous, then

$$g(x) f_n(x) \to g(x) f(x)$$

pointwise since $f_n(x) \to f(x)$ pointwise. Let $K = \sup |g(x)|$ and define

$$h_1(x) = g(x) + K \quad \text{and} \quad h_2(x) = K - g(x).$$

Notice that both h_1 and h_2 are positive functions, and so $h_1 f_n$ and $h_2 f_n$ are also necessarily positive functions. Furthermore, $h_i(x) f_n(x) \to h_i(x) f(x)$, $i = 1, 2$, pointwise. We can now apply Fatou's lemma to conclude that

$$\mathbb{E}(h_i(X)) = \int_{\mathbb{R}} h_i(x) f(x)\,\mathrm{d}x \le \liminf_{n\to\infty} \int_{\mathbb{R}} h_i(x) f_n(x)\,\mathrm{d}x = \liminf_{n\to\infty} \mathbb{E}(h_i(X_n)). \qquad (**)$$

Now observe that

$$\mathbb{E}(h_1(X_n)) = \mathbb{E}(g(X_n)) + K \quad \text{and} \quad \mathbb{E}(h_2(X_n)) = K - \mathbb{E}(g(X_n))$$

as well as

$$\mathbb{E}(h_1(X)) = \mathbb{E}(g(X)) + K \quad \text{and} \quad \mathbb{E}(h_2(X)) = K - \mathbb{E}(g(X)).$$

From $(**)$ we see that

$$\mathbb{E}(g(X)) + K = \mathbb{E}(h_1(X)) \leq \liminf_{n\to\infty} \mathbb{E}(h_1(X_n)) = \liminf_{n\to\infty} [\mathbb{E}(g(X_n)) + K]$$

and so

$$\mathbb{E}(g(X)) \leq \liminf_{n\to\infty} \mathbb{E}(g(X_n))$$

and

$$K - \mathbb{E}(g(X)) = \mathbb{E}(h_2(X)) \leq \liminf_{n\to\infty} \mathbb{E}(h_2(X_n)) = \liminf_{n\to\infty} [K - \mathbb{E}(g(X_n))]$$

and so

$$-\mathbb{E}(g(X)) \leq \liminf_{n\to\infty} [-\mathbb{E}(g(X_n))].$$

Using the fact that $\liminf[-a_n] = -\limsup a_n$ gives

$$\mathbb{E}(g(X)) \geq \limsup_{n\to\infty} \mathbb{E}(g(X_n)).$$

Combined we have shown that

$$\limsup_{n\to\infty} \mathbb{E}(g(X_n)) \leq \mathbb{E}(g(X)) \leq \liminf_{n\to\infty} \mathbb{E}(g(X_n))$$

which implies that

$$\lim_{n\to\infty} \mathbb{E}(g(X_n)) = \mathbb{E}(g(X)).$$

This is exactly what is needed to establish $(*)$. □

We end this lecture with an "almost"-converse.

Theorem. *Suppose that X_n, $n = 1, 2, 3, \ldots$, and X are all defined on a common probability space. If $X_n \to X$ weakly and X is almost surely constant, then $X_n \to X$ in probability.*

Proof. Suppose that X is almost surely equal to the constant a. The function

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

is bounded and continuous. Since $X_n \to X$ weakly, we can use Theorem 18.1 to conclude that $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ as $n \to \infty$. That is,

$$\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - a|}{1 + |X_n - a|}\right) = \mathbb{E}(f(X)) = \mathbb{E}(f(a)) = 0.$$

By Theorem 17.1, this implies that $X_n \to a$ in probability. □

Statistics 851 (Winter 2008) March 26, 2008
Prof. Michael Kozdron

Lecture #30: Weak Convergence

Reference. Chapter 18, pages 151–163

Last lecture we introduced the fifth mode of convergence, namely weak convergence or convergence in distribution. As the name suggests, it is the weakest of the five modes. It has the advantage, however, that unlike the other modes, random variables can converge in distribution even if those random variables are not defined on a common probability space. If they happen to be defined on a common probability space, then we proved that convergence in probability implies convergence in distribution.

Formally, suppose that $X_1, X_2, X_3, \ldots$, and X are all random variables. For each n, the random variable $X_n : (\Omega_n, \mathcal{A}_n, P_n) \to (\mathbb{R}, \mathcal{B})$ and $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$.

Definition. We say that X_n *converges weakly to* X (or that X_n *converges in distribution to* X or that X_n *converges in law to* X) if

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x) P^{X_n}(\mathrm{d}x) = \int_{\mathbb{R}} f(x) P^{X}(\mathrm{d}x)$$

for every bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$.

We also indicated last lecture that this definition is difficult to work with. Since the distribution function characterizes the law of a random variable, it seems reasonable that we should be able to deduce weak convergence from convergence of distribution functions.

Theorem. *The random variables $X_n \to X$ in distribution if and only if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all $x \in \mathcal{D}$ where

$$\mathcal{D} = \{x : F_X(x-) = F_X(x)\}$$

denotes the continuity set of F_X.

Proof. See pages 153–155 for a complete proof. □

This theorem says that in order for random variables X_n to converge in distribution to X it is necessary and sufficient for their distribution functions to converge for all those x at which F_X is continuous.

Example. Suppose that X_n, $n = 1, 2, 3, \ldots$, has density function

$$f_n(x) = \frac{n}{\pi(1 + n^2x^2)}, \quad -\infty < x < \infty.$$

We showed last lecture that if X_n are all defined on a common probability space, then $X_n \to 0$ in distribution. Notice that the distribution function of X_n is given by

$$F_{X_n}(x) = P\{X_n \le x\} = \int_{-\infty}^{x} f_n(u)\,\mathrm{d}u = \int_{-\infty}^{x} \frac{n}{\pi(1+n^2u^2)}\,\mathrm{d}u = \frac{1}{2} + \frac{1}{\pi}\arctan(nx)$$

while the distribution function of $X = 0$ is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \ge 0. \end{cases}$$

Notice that if $x = 0$ is fixed, then

$$\lim_{n\to\infty} F_{X_n}(0) = \lim_{n\to\infty} \left[\frac{1}{2} + \frac{1}{\pi}\arctan(0)\right] = \frac{1}{2}$$

which does not equal $F_X(0)$. This does not contradict the convergence in distribution result obtained last lecture since $x = 0$ is not in $\mathcal{D}$, the continuity set of F_X.

We do note that if $x < 0$ is fixed, then

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \left[\frac{1}{2} + \frac{1}{\pi}\arctan(nx)\right] = \frac{1}{2} - \frac{1}{2} = 0$$

while if $x > 0$ is fixed, then

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \left[\frac{1}{2} + \frac{1}{\pi}\arctan(nx)\right] = \frac{1}{2} + \frac{1}{2} = 1.$$

That is, $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathcal{D}$ and so we can conclude directly from the previous theorem that $X_n \to 0$ in distribution. (Recall that last class we showed that $X_n \to 0$ in probability which allowed us to conclude that $X_n \to 0$ in distribution.)

Example. Suppose that X_n, $n = 1, 2, 3, \ldots$, are defined by

$$P\{X_n = n\} = P\{X_n = -n\} = \frac{1}{2}.$$

The distribution function of X_n is given by

$$F_{X_n}(x) = P\{X_n \le x\} = \begin{cases} 0, & \text{if } x < -n, \\ \frac{1}{2}, & \text{if } -n \le x < n, \\ 1, & \text{if } x \ge n. \end{cases}$$

Hence, if $x \in \mathbb{R}$ is fixed, then as $n \to \infty$, we have

$$\lim_{n\to\infty} F_{X_n}(x) = \frac{1}{2}.$$

Since $F(x) = 1/2$, $-\infty < x < \infty$, does not define a legitimate distribution function, we see from the previous theorem that it is not possible for X_n to converge in distribution.

Example. Suppose that $\Omega_n = [-n, n]$, $\mathcal{A}_n = \mathcal{B}(\Omega_n)$, and P_n is the uniform probability measure on Ω_n so that P_n has density function

$$f_n(x) = \frac{1}{2n} \cdot 1_{[-n,n]}(x).$$

Notice that as n gets larger, the height $1/2n$ gets smaller. In other words, as n increases, the uniform distribution needs to spread itself thinner and thinner in order to remain legitimate. Also notice that as $n \to \infty$, we have $f_n(x) \to 0$. If we now define the random variables X_n, $n = 1, 2, 3, \ldots$, by setting

$$X_n(\omega) = \omega \quad \text{for } \omega \in \Omega_n,$$

then it should be clear that X_n does not converge in distribution. Indeed, since X_n is the uniform distribution on $[-n, n]$, the "limiting distribution" should be uniform on $(-\infty, \infty)$. Since $(-\infty, \infty)$ cannot support a uniform distribution, it makes sense that X_n does not converge in distribution. Formally, we can use the previous theorem to prove that this is the case. Notice that the distribution function of X_n is given by

$$F_{X_n}(x) = P_n\{X_n \le x\} = \begin{cases} 0, & \text{if } x < -n, \\ \frac{x+n}{2n}, & \text{if } -n \le x \le n, \\ 1, & \text{if } x > n. \end{cases}$$

Hence, if $x \in \mathbb{R}$ is fixed, then we find

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \frac{x+n}{2n} = \frac{1}{2}.$$

Thus, the only possible limit is $F(x) = 1/2$ which is not a legitimate distribution function and so we conclude that X_n does not converge in distribution.

Remark. Even though the random variables X_n in the previous examples do not converge weakly, we see that the distribution functions still converge (albeit not to a distribution function). As such, some authors say that X_n converges vaguely to the pseudo-distribution function $F(x) = 1/2$. (And such authors then define vague convergence to be weaker than weak convergence.)

Remark. There are two other technical theorems proved in Chapter 18. Theorem 18.6 is a form of the Helly Selection Principle and Theorem 18.8 is a form of Slutsky's Theorem which is sometimes useful in statistics.

We end this lecture with a characterization of convergence in distribution for discrete random variables.

Theorem. *Suppose that X_n, $n = 1, 2, 3, \ldots$, and X are random variables with at most countably many values. It then follows that $X_n \to X$ in distribution if and only if*

$$\lim_{n\to\infty} P\{X_n = j\} = P\{X = j\}$$

for every j in the state space of X_n, $n = 1, 2, 3, \ldots$, and X.

Proof. See pages 161–162. □

Example. Suppose that λ_n, $n = 1, 2, 3, \ldots$, is a sequence of positive real numbers converging to $\lambda > 0$. If X_n has a Poisson(λ_n) distribution so that

$$P\{X_n = j\} = \frac{\lambda_n^j e^{-\lambda_n}}{j!}, \quad j = 0, 1, 2, 3, \ldots,$$

then $X_n \to X$ in distribution where X has a Poisson(λ) distribution

$$P\{X = j\} = \frac{\lambda^j e^{-\lambda}}{j!}, \quad j = 0, 1, 2, 3, \ldots.$$

Example. Suppose that p_n, $n = 1, 2, 3, \ldots$, is a sequence of real numbers with $0 \leq p_n \leq 1$ converging to $p \in [0, 1]$. If N is fixed and X_n has a Binomial(N, p_n) distribution so that

$$P\{X_n = j\} = \binom{N}{j} p_n^j (1 - p_n)^{N-j}, \quad j = 0, 1, \ldots, N,$$

then $X_n \to X$ in distribution where X has a Binomial(N, p) distribution

$$P\{X = j\} = \binom{N}{j} p^j (1 - p)^{N-j}, \quad j = 0, 1, \ldots, N.$$

Lecture #31: Characteristic Functions

Reference. Chapter 13, pages 103–109

In Chapter 13, things are proved for $\mathbb{R}^n$-valued random variables. We only consider $\mathbb{R}$-valued random variables.

Notation. If $\overline{x} = (x_1, \ldots, x_n)$ and $\overline{y} = (y_1, \ldots, y_n)$ are vectors in $\mathbb{R}^n$, then

$$\overline{x} \cdot \overline{y} = \langle \overline{x}, \overline{y} \rangle = \sum_{i=1}^{n} x_i y_i$$

denotes the *dot product* (or *scalar product*) of $\overline{x}$ and $\overline{y}$. In the special case of $\mathbb{R}^1$, the dot product

$$x \cdot y = \langle x, y \rangle = xy$$

reduces to usual multiplication.

Definition. If X is a random variable, then the *characteristic function* of X is given by

$$\varphi_X(u) = \mathbb{E}(e^{iuX}), \quad u \in \mathbb{R}.$$

Note that $\varphi_X : \mathbb{R} \to \mathbb{C}$ is given by $u \mapsto \mathbb{E}(e^{iuX})$.

Recall. The imaginary unit i satisfies $i^2 = -1$ and the complex numbers are the set

$$\mathbb{C} = \{z = x + iy : x \in \mathbb{R}, y \in \mathbb{R}\}.$$

Thus, if $z = x + iy \in \mathbb{C}$, we can identify z with the ordered pair $(x, y) \in \mathbb{R}^2$, and so we sometimes say that $\mathbb{C} \cong \mathbb{R}^2$, i.e., that $\mathbb{C}$ and $\mathbb{R}^2$ are isomorphic.

If $z = x + iy \in \mathbb{C}$, then the *imaginary part* of z is $\Im(z) = y$ and the *real part* of z is $\Re(z) = x$. The *complex conjugate* of z is

$$\overline{z} = x - iy$$

and the *complex modulus* (or absolute value) of z satisfies

$$|z|^2 = z\overline{z} = (x + iy)(x - iy) = x^2 - i^2 y^2 = x^2 + y^2.$$

The *complex exponential* is defined by

$$e^{i\theta} = \cos(\theta) + i \sin(\theta), \quad \theta \in \mathbb{R},$$

and satisfies

$$|e^{i\theta}|^2 = |\cos(\theta) + i \sin(\theta)|^2 = \cos^2(\theta) + \sin^2(\theta) = 1.$$

This previous result leads to the most important fact about characteristic functions, namely that they always exist.

Fact. If X is a random variable, then $\varphi_X(u)$ exists for all $u \in \mathbb{R}$.

Proof. Let $u \in \mathbb{R}$ so that

$$|\varphi_X(u)| = |\mathbb{E}(e^{iuX})| \leq \mathbb{E}|e^{iuX}| = \mathbb{E}(1) = 1.$$

This proves the claim. □

Fact. If X is a random variable, then $\varphi_X(0) = 1$.

Proof. By definition,

$$\varphi_X(0) = \mathbb{E}(e^{i\cdot 0\cdot X}) = \mathbb{E}(1) = 1$$

as required. □

Remark. If you remember moment generating functions from undergraduate probability, then you will see the similarity between characteristic functions and moment generating functions. In fact, the moment generating function of X is given by

$$m_X(t) = \mathbb{E}(e^{tX}), \quad t \in \mathbb{R}.$$

Note that $m_X : \mathbb{R} \to \mathbb{R}$. The problem, however, is that moment generating functions do not always exist.

Example. Suppose that X is exponentially distributed with parameter $\lambda > 0$ so that the density of X is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

The moment generating function of X is

$$m_X(t) = \mathbb{E}(e^{tX}) = \int_0^\infty e^{tx} \cdot \lambda e^{-\lambda x}\, \mathrm{d}x = \lambda \int_0^\infty \exp\{-x(\lambda - t)\}\, \mathrm{d}x = \frac{\lambda}{\lambda - t}.$$

Note that $m_X(t)$ is well-defined only if $t < \lambda$.

Remark. In undergraduate probability you were probably told the facts that

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} m_X(t)\bigg|_{t=0} = \mathbb{E}(X^k)$$

(i.e., the moment generating function generates moments) and that "the moment generating function characterizes the random variable." While the first fact is true provided the moment generating function is well-defined, the second fact is not quite true. It is unfortunate that moment generating functions are taught at all since the characteristic function is better at accomplishing the tasks that the moment generating functions was designed for!

Theorem. *If $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ is a random variable with $\mathbb{E}(|X|^k) < \infty$ for some $k \in \mathbb{N}$, then φ_X has derivatives up to order k which are continuous and satisfy*

$$\frac{\mathrm{d}^k}{\mathrm{d}u^k} \varphi_X(u) = i^k \mathbb{E}(X^k e^{iuX}).$$

In particular,

$$\frac{\mathrm{d}^k}{\mathrm{d}u^k} \varphi_X(u)\bigg|_{u=0} = i^k \mathbb{E}(X^k).$$

Theorem. *The characteristic function characterizes the random variable. That is, if $\varphi_X(u) = \varphi_Y(u)$ for all $u \in \mathbb{R}$, then $P^X = P^Y$. (In other words, X and Y are equal in distribution.)*

We will prove these important results later. For now, we mention a couple of examples where we can explicitly calculate the characteristic function. In general, computing characteristic functions is not as easy as computing moment generating functions since care needs to be shown with complex integrals. The usual technique is to write the real and imaginary parts separately.

Theorem. *If X is a random variable with characteristic function $\varphi(u)$ and $Y = aX + b$ with a, $b \in \mathbb{R}$, then $\varphi_Y(u) = e^{iub}\varphi_X(au)$.*

Proof. By definition,

$$\varphi_Y(u) = \mathbb{E}(e^{iuY}) = \mathbb{E}(e^{iu(aX+b)}) = e^{iub}\mathbb{E}(e^{iuaX}) = e^{iub}\varphi_X(au)$$

and the result is proved. □

Theorem. *If $X_1, X_2, \ldots, X_n$ are independent random variables and*

$$S_n = X_1 + X_2 + \cdots + X_n,$$

then

$$\varphi_{S_n}(u) = \prod_{i=1}^{n} \varphi_{X_i}(u).$$

Furthermore, if $X_1, X_2, \ldots, X_n$ are identically distributed, then

$$\varphi_{S_n}(u) = \left[\varphi_{X_1}(u)\right]^n.$$

Proof. By definition,

$$\varphi_{S_n}(u) = \mathbb{E}(e^{iuS_n}) = \mathbb{E}(e^{iu(X_1+\cdots+X_n)}) = \mathbb{E}(e^{iuX_1}\cdots e^{iuX_n}) = \mathbb{E}(e^{iuX_1})\cdots\mathbb{E}(e^{iuX_n}) = \prod_{i=1}^{n}\varphi_{X_i}(u)$$

where the second-to-last equality follows from the fact that X_i are independent. If X_i are also identically distributed, then $\varphi_{X_i}(u) = \varphi_{X_1}(u)$ for all i and so

$$\varphi_{S_n}(u) = \left[\varphi_{X_1}(u)\right]^n$$

as required. □

Example. Suppose that X is exponentially distributed with parameter $\lambda > 0$ so that the density of X is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

The characteristic function of X is given by

$$\varphi_X(u) = \mathbb{E}(e^{iuX}) = \int_0^\infty e^{iux} \cdot \lambda e^{-\lambda x}\,\mathrm{d}x = \lambda \int_0^\infty \exp\left\{-\lambda x + iux\right\}\,\mathrm{d}x.$$

In order to evaluate this integral, we write

$$\exp\{-\lambda x + iux\} = e^{-\lambda x}\cos(ux) + ie^{-\lambda x}\sin(ux)$$

so that

$$\int_0^\infty \exp\{-\lambda x + iux\}\,\mathrm{d}x = \int_0^\infty e^{-\lambda x}\cos(ux)\,\mathrm{d}x + i\int_0^\infty e^{-\lambda x}\sin(ux)\,\mathrm{d}x.$$

Integration by parts shows that

$$\int_0^\infty e^{-\lambda x}\cos(ux)\,\mathrm{d}x = \frac{1}{\lambda^2+u^2}e^{-\lambda x}\left[-\lambda\cos(ux) + u\sin(ux)\right]\Big|_0^\infty$$

and

$$\int_0^\infty e^{-\lambda x}\sin(ux)\,\mathrm{d}x = \frac{1}{\lambda^2+u^2}e^{-\lambda x}\left[-\lambda\sin(ux) - u\cos(ux)\right]\Big|_0^\infty$$

which combined give

$$\varphi_X(u) = \frac{\lambda^2}{\lambda^2+u^2} - i\frac{\lambda u}{\lambda^2+u^2} = \frac{\lambda}{\lambda - iu}.$$

Note that $\varphi_X(u)$ exists for all $u \in \mathbb{R}$.

Example. If X has a Bernoulli(p) distribution, then

$$\varphi_X(u) = \mathbb{E}(e^{iuX}) = e^{iu\cdot 1}P\{X=1\} + e^{iu\cdot 0}P\{X=0\} = pe^{iu} + 1 - p.$$

Example. If X has a Binomial(n,p) distribution, then

$$\varphi_X(u) = [pe^{iu} + 1 - p]^n.$$

Example. If X has a Uniform$[-a,a]$ distribution, then

$$\varphi_X(u) = \frac{\sin(au)}{au}.$$

Note that $\varphi_X(0) = 1$ as expected.

Lecture #32: The Primary Limit Theorems

The Central Limit Theorem, along with the weak and strong laws of large numbers, are primary results in the theory of probability and statistics. In fact, the Central Limit Theorem has been called the Fundamental Theorem of Statistics.

The Weak and Strong Laws of Large Numbers

Reference. Chapter 20, pages 173–177

Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables. The random variable $\overline{X}_n$ defined by

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j = \frac{X_1 + \cdots + X_n}{n}$$

is sometimes called the *sample mean*, and much of the theory of statistical inference is concerned with the behaviour of the sample mean. In particular, confidence intervals and hypothesis tests are often based on approximate and asymptotic results for the sample mean. The mathematical basis for such results is the law of large numbers which asserts that the sample mean $\overline{X}_n$ converges to the population mean. As we know, there are several modes of convergence that one can consider. The weak law of large numbers provides convergence in probability and the strong law of large numbers provides convergence almost surely.

Of course, convergence almost surely implies convergence in probability, and so the weak law of large numbers follows immediately from the strong law of large numbers. However, the proof serves as a nice warm-up example.

Theorem (Weak Law of Large Numbers)**.** *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed L^2 random variables with common mean $\mathbb{E}(X_1) = \mu$. If $\overline{X}_n$ denotes the sample mean as above, then*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \to \mu \quad \textit{in probability}$$

as $n \to \infty$.

Proof. Write σ^2 for the common variance of $X_1, X_2, \ldots$, and let $S_n = X_1 + \cdots + X_n - n\mu$. In order to prove that $\overline{X}_n \to \mu$ in probability, it suffices to show that

$$\frac{S_n}{n} \to 0 \quad \text{in probability}$$

as $n \to \infty$. Note that $\mathbb{E}(S_n) = 0$ and since $X_1, X_2, \ldots$ are independent, we see that

$$\mathbb{E}(S_n^2) = \mathrm{Var}(S_n) = \sum_{j=1}^{n} \mathrm{Var}(X_j) = n\sigma^2.$$

Therefore, if $\epsilon > 0$ is fixed, then Chebyshev's inequality implies that

$$P\{|S_n| > \epsilon n\} \leq \frac{\mathbb{E}(S_n^2)}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

and so we conclude that

$$\lim_{n\to\infty} P\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

That is, $\overline{X_n} \to \mu$ in probability as $n \to \infty$ and the proof is complete. □

Note that in our proof of the weak law we derived

$$P\{|S_n| > \epsilon n\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since complete convergence implies convergence almost surely, we see that we could establish almost sure convergence if the previous expression were summable in n. Unfortunately, it is not. However, if we relax the hypotheses slightly then we can derive the strong law of large numbers via this technique.

Theorem (Strong Law of Large Numbers)**.** *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed L^4 random variables with common mean $\mathbb{E}(X_1) = \mu$. If $\overline{X}_n$ denotes the sample mean as above, then*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \to \mu \quad \textit{almost surely}$$

as $n \to \infty$.

Proof. As above, write σ^2 for the common variance of the X_j and let

$$S_n = X_1 + \cdots + X_n - n\mu.$$

It then follows from Markov's inequality that

$$P\{|S_n| > \epsilon n\} \leq \frac{\mathbb{E}(S_n^4)}{n^4\epsilon^4}.$$

Our goal now is to estimate $\mathbb{E}(S_n^4)$ and show that the previous expression is summable in n. Therefore, if we write $Y_i = X_i - \mu$, then

$$\begin{aligned} S_n^4 &= (Y_1 + Y_2 + \cdots + Y_n)^4 \\ &= \sum_{j=1}^{n} Y_j^4 + \sum_{i\neq j} Y_i^3 Y_j + \sum_{i\neq j} Y_i^2 Y_j^2 + \sum_{i\neq j\neq k} Y_i^2 Y_j Y_k + \sum_{i\neq j\neq k\neq \ell} Y_i Y_j Y_k Y_\ell. \end{aligned}$$

Since $Y_1, Y_2, \ldots$ are independent with $\mathbb{E}(Y_1) = 0$, we see that

$$\mathbb{E}\left(\sum_{i\neq j} Y_i^3 Y_j\right) = \sum_{i\neq j} \mathbb{E}(Y_i^3 Y_j) = \sum_{i\neq j} \mathbb{E}(Y_i^3)\mathbb{E}(Y_j) = 0$$

as well as

$$\mathbb{E}\left(\sum_{i\neq j\neq k} Y_i^2 Y_j Y_k\right) = \sum_{i\neq j\neq k} \mathbb{E}(Y_i^2)\mathbb{E}(Y_j)\mathbb{E}(Y_k) = 0$$

and

$$\mathbb{E}\left(\sum_{i\neq j\neq k\neq \ell} Y_i Y_j Y_k Y_\ell\right) = \sum_{i\neq j\neq k\neq \ell} \mathbb{E}(Y_i)\mathbb{E}(Y_j)\mathbb{E}(Y_k)\mathbb{E}(Y_\ell) = 0.$$

This gives

$$\mathbb{E}(S_n^4) = \sum_{j=1}^{n} \mathbb{E}(Y_j^4) + \sum_{i\neq j} \mathbb{E}(Y_i^2)\mathbb{E}(Y_j^2).$$

Since $X_j \in L^4$, we see that $Y_j \in L^4$ and $S_n \in L^4$. Thus, if we write $\mathbb{E}(Y_j^4) = M$, then since $\mathbb{E}(Y_j^2) = \sigma^2$, we see that there exists some constant C such that

$$\mathbb{E}(S_n^4) = nM + \binom{4}{2}\frac{n(n-1)}{2}\sigma^4 \leq Cn^2.$$

Note that we can take $C = M + 3\sigma^4$ since

$$nM + \binom{4}{2}\frac{n(n-1)}{2}\sigma^4 = nM + 3n^2\sigma^4 - 3n\sigma^4 \leq nM + 3n^2\sigma^4 \leq n^2(M + 3\sigma^4).$$

Combined with the above, we see that

$$P\{|S_n| > \epsilon n\} \leq \frac{\mathbb{E}(S_n^4)}{n^4\epsilon^4} \leq \frac{Cn^2}{n^4\epsilon^4} = \frac{C}{\epsilon^4}\cdot\frac{1}{n^2}$$

and so

$$\sum_{n=1}^{\infty} P\{|S_n| > \epsilon n\} \leq \frac{C}{\epsilon^4}\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty.$$

In other words, $\overline{X}_n \to \mu$ completely as $n \to \infty$ and so we see that $\overline{X}_n \to \mu$ almost surely as $n \to \infty$. □

Remark. The weakest possible hypothesis for the strong law of the large numbers is to just assume that $X_1, X_2, \ldots$ are independent and identically distributed with $\mathbb{E}|X_1| < \infty$. That is, to assume that $X_j \in L^1$. It turns out that the theorem is true under this hypothesis, but the proof is much more detailed. We will return to this point later.

The Central Limit Theorem

Reference. Chapter 21, pages 181–185

Last lecture we defined characteristic functions and stated their most important properties. If $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ is a random variable, then the characteristic function of X is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ given by

$$\varphi_X(u) = \mathbb{E}(e^{iuX}), \quad u \in \mathbb{R}.$$

As shown last lecture, $|\varphi_X(u)| \leq 1$ and so characteristic functions always exist. The two most important facts about characteristic functions are the following.

- If $\mathbb{E}|X|^k < \infty$, then $\varphi_X(u)$ has derivatives up to order k and

$$\frac{\mathrm{d}^k}{\mathrm{d}u^k}\varphi_X(u) = i^k\mathbb{E}(X^k e^{iuX}).$$

- If X and Y are random variables with $\varphi_X(u) = \varphi_Y(u)$ for all u, then $P^X = P^Y$ so that X and Y are equal in distribution.

There is one more extremely useful result which tells us when we can deduce convergence in distribution from convergence of characteristic functions. This result is a special case of a more general result due to Lévy.

Theorem (Lévy's Continuity Theorem)**.** *Suppose that X_n, $n = 1, 2, 3, \ldots$, is a sequence of random variables with corresponding characteristic functions φ_n. Suppose further that*

$$g(u) = \lim_{n\to\infty} \varphi_n(u) \quad \textit{for all } u \in \mathbb{R}.$$

If g is continuous at 0, then g is the characteristic function of some random variables X (that is, $g = \varphi_X$) and $X_n \to X$ weakly.

Having discussed convergence in distribution and characteristic functions, we are now in a position to state and prove the Central Limit Theorem. We will defer the proofs of the two previous facts to a later lecture in order to first prove the Central Limit Theorem since the proof serves as a nice illustration of the ideas we have recently discussed.

Theorem. *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed L^2 random variables. If we write $\mathbb{E}(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$ for their common mean and variance, respectively, then as $n \to \infty$,*

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to Z$$

in distribution where Z is normally distributed with mean 0 and variance 1.

In order to prove this result, we need to use Exercise 14.4 which we state as a Lemma.

Lemma. *If X is an L^2 random variable with $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) = \sigma^2$, then*

$$\varphi_X(u) = 1 - \frac{1}{2}\sigma^2u^2 + o(u^2) \quad \textit{as } u \to 0.$$

Recall. A function $h(t)$ is said to be *little-o of $g(t)$*, written $h(t) = o(g(t))$, if

$$\lim_{t\to 0+} \frac{h(t)}{g(t)} = 0.$$

Proof of the Central Limit Theorem. For each $n = 1, 2, 3, \ldots$, let

$$Y_n = \frac{X_n - \mu}{\sigma}$$

so that $Y_1, Y_2, \ldots,$ are independent and identically distributed with $\mathbb{E}(Y_1) = 0$ and $\mathrm{Var}(Y_1) = 1$. Furthermore, if we define S_n by

$$S_n = \frac{Y_1 + \cdots + Y_n}{\sqrt{n}},$$

then

$$S_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

and so the theorem will be proved if we can show $S_n \to Z$ in distribution. Using our results from last lecture on the characteristic functions of linear transformations, we find

$$\varphi_{S_n}(u) = \mathbb{E}(e^{iuS_n}) = \mathbb{E}(e^{i\frac{u}{\sqrt{n}}(Y_1+\cdots+Y_n)}) = \varphi_{Y_1}(u/\sqrt{n})\cdots\varphi_{Y_n}(u/\sqrt{n}) = \left[\varphi_{Y_1}(u/\sqrt{n})\right]^n.$$

Using the previous lemma, we find

$$\varphi_{S_n}(u) = \left[\varphi_{Y_1}(u/\sqrt{n})\right]^n = \left[1 - \frac{u^2}{2n} + o(u^2/n)\right]^n$$

and so taking the limit as $n \to \infty$ gives

$$\lim_{n\to\infty} \varphi_{S_n}(u) = \lim_{n\to\infty} \left[1 - \frac{u^2}{2n} + o(u^2/n)\right]^n = \lim_{n\to\infty} \left[1 - \frac{u^2}{2n}\right]^n = \exp\left\{-\frac{u^2}{2}\right\}$$

by the definition of e as a limit. Thus, if $g(u) = \exp\{-u^2/2\}$, then g is clearly continuous at 0 and so g must be the characteristic function of some random variable. We now note that if Z is a normal random variable with mean 0 and variance 1, then the characteristic function of Z is $\varphi_Z(u) = \exp\{-u^2/2\}$. Thus, by Lévy's continuity theorem, we see that $S_n \to Z$ in distribution. In other words,

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \to Z$$

in distribution as $n \to \infty$ and the theorem is proved. □

Lecture #33: Further Results on Characteristic Functions

Reference. Chapter 13 pages 103–109; Chapter 14 pages 111–113

Last lecture we proved the Central Limit Theorem whose proof required an analysis of characteristic functions. The purpose of this lecture is to prove several of the facts about characteristic functions that we have been using.

The first result tells us that characteristic functions can be used to calculate moments. Recall that if X is a random variable, then its characteristic function $\varphi_X : \mathbb{R} \to \mathbb{C}$ is defined by

$$\varphi_X(u) = \mathbb{E}(e^{iuX}).$$

Theorem. *If $X \in L^1$, then*

$$\frac{\mathrm{d}}{\mathrm{d}u}\varphi_X(u) = i\mathbb{E}(Xe^{iuX}).$$

Proof. Using the definition of derivative, we find

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}u}\varphi_X(u) = \frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}(e^{iuX}) &= \frac{\mathrm{d}}{\mathrm{d}u}\int_{\mathbb{R}} e^{iux}P^X(\mathrm{d}x) \\
&= \lim_{h\to 0}\frac{1}{h}\left[\int_{\mathbb{R}} e^{i(u+h)x}P^X(\mathrm{d}x) - \int_{\mathbb{R}} e^{iux}P^X(\mathrm{d}x)\right] \\
&= \lim_{h\to 0}\frac{1}{h}\int_{\mathbb{R}} e^{iux}(e^{ixh}-1)P^X(\mathrm{d}x) \\
&= \lim_{h\to 0}\int_{\mathbb{R}} e^{iux}\left(\frac{e^{ixh}-1}{h}\right)P^X(\mathrm{d}x).
\end{aligned}$$

We would now like to interchange the integral and the limit. In order to do so, we will use the Dominated Convergence Theorem. To begin, note that

$$\left|e^{iux}\left(\frac{e^{ixh}-1}{h}\right)\right| \leq \left|\frac{e^{ixh}-1}{h}\right| \leq \frac{2|ixh|}{h} = 2|x|$$

using Taylor's Theorem; that is, $|e^{\theta}-1| \leq 2|\theta|$. (You are asked to prove this in Exercise 15.14.) Also note that

$$\int_{\mathbb{R}} |x|P^X(\mathrm{d}x) = \mathbb{E}(|X|) < \infty$$

by the assumption that $X \in L^1$. We can now apply the Dominated Convergence Theorem to conclude that

$$\lim_{h\to 0}\int_{\mathbb{R}} e^{iux}\left(\frac{e^{ixh}-1}{h}\right)P^X(\mathrm{d}x) = \int_{\mathbb{R}} e^{iux}\lim_{h\to 0}\left(\frac{e^{ixh}-1}{h}\right)P^X(\mathrm{d}x) = i\int_{\mathbb{R}} xe^{iux}P^X(\mathrm{d}x)$$

and the proof is complete. □

It now follows immediately that if $X \in L^k$, then φ_X has derivatives up to order k which satisfy

$$\frac{\mathrm{d}^k}{\mathrm{d}u^k}\varphi_X(u) = i^k \mathbb{E}(X^k e^{iuX}).$$

Theorem. *If $X_n \to X$ in distribution, then $\varphi_{X_n}(u) \to \varphi_X(u)$.*

Proof. Since $X_n \to X$ in distribution, we know that

$$\int_{\mathbb{R}} g(x) P^{X_n}(\mathrm{d}x) \to \int_{\mathbb{R}} g(x) P^X(\mathrm{d}x)$$

for every bounded, continuous function g. In other words,

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

for every such g. If we choose $g(x) = e^{iux}$, then we see that g is continuous and satisfies $|g(x)| \leq 1$, and so

$$\lim_{n\to\infty} \mathbb{E}(e^{iuX_n}) = \mathbb{E}(e^{iuX}).$$

Since $\varphi_{X_n}(u) = \mathbb{E}(e^{iuX_n})$ and $\varphi_X(u) = \mathbb{E}(e^{iuX})$, the result follows. □

Remark. The object that we call the characteristic function, namely

$$\varphi_X(u) = \mathbb{E}(e^{iuX}) = \int_{\mathbb{R}} e^{iux} P^X(\mathrm{d}x),$$

is known in functional analysis as the *Fourier transform* of the measure P^X. It is sometimes written as

$$\hat{P}^X(u) = \int_{\mathbb{R}} e^{iux} P^X(\mathrm{d}x).$$

In particular, theorems about uniqueness of characteristic functions can be phrased in terms of Fourier transforms and proved using results from functional analysis.

Theorem. *If X, Y are random variables with $\varphi_X(u) = \varphi_Y(u)$ for all u, then $P^X = P^Y$. That is, X and Y are equal in distribution.*

Proof. The proof of this theorem requires the Stone-Weierstrass theorem and is given by Theorem 14.1 on pages 111–112. □

Example. Suppose that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent. Determine the distribution of $X_1 + X_2$.

Solution. This problem can be solved using characteristic functions. We know that the characteristic function of a $\mathcal{N}(\mu, \sigma^2)$ random variable is given by

$$\varphi(u) = \exp\left\{ i\mu u - \frac{\sigma^2 u^2}{2} \right\}.$$

We also know that

$$\varphi_{X_1+X_2}(u) = \varphi_{X_1}(u)\varphi_{X_2}(u)$$

since X_1 and X_2 are independent. Therefore,

$$\varphi_{X_1+X_2}(u) = \exp\left\{i\mu_1 u - \frac{\sigma_1^2 u^2}{2}\right\} \exp\left\{i\mu_2 u - \frac{\sigma_2^2 u^2}{2}\right\} = \exp\left\{i(\mu_1+\mu_2)u - \frac{(\sigma_1^2+\sigma_2^2)u^2}{2}\right\}$$

which implies that $X_1 + X_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.

Lecture #34: Conditional Expectation

Reference. Chapter 23 pages 197–207

In order to motivate the definition and construction of conditional expectation, we will recall the definition from undergraduate probability. Suppose that (X,Y) are jointly distributed continuous random variables with joint density function $f(x,y)$. We define the conditional density of Y given $X=x$ by setting

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

where

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$$

is the marginal density for X. The conditional expectation of Y given $X=x$ is then defined to be

$$\mathbb{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\,\mathrm{d}y = \frac{\int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y}{f_X(x)} = \frac{\int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y}{\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y}.$$

Example. Suppose that (X,Y) are continuous random variables with joint density function

$$f(x,y) = \begin{cases} 12xy, & \text{if } 0<y<1 \text{ and } 0<x<y^2<1, \\ 0, & \text{otherwise.} \end{cases}$$

We find

$$f_X(x) = \int_{\sqrt{x}}^{1} 12xy\,\mathrm{d}y = 6x(1-x), \quad 0<x<1,$$

so that

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{2y}{1-x}, \quad \sqrt{x}<y<1.$$

Thus, if $0<x<1$ is fixed, then the conditional expectation of Y given $X=x$ is

$$\mathbb{E}(Y|X=x) = \int_{\sqrt{x}}^{1} y\cdot\frac{2y}{1-x}\,\mathrm{d}y = \frac{2-2x^{3/2}}{3(1-x)}.$$

In general (and as seen in the previous example), the conditional expectation depends on the value of X observed so that $\mathbb{E}(Y|X=x)$ can be regarded as a function of x. That is, we will write

$$\mathbb{E}(Y|X=x) = \varphi(x)$$

for some function φ. Furthermore, we see that if we view the conditional expectation not as a function of the observed x but rather as a function of the random X, we have

$$\mathbb{E}(Y|X) = \varphi(X).$$

Thus, the conditional expectation is really a random variable.

Example (continued). We have

$$\varphi(x) = \frac{2 - 2x^{3/2}}{3(1-x)}$$

and so

$$\mathbb{E}(Y|X) = \varphi(X) = \frac{2 - 2X^{3/2}}{3(1-X)}.$$

And so motivated by our experience from undergraduate probability, we assert that the conditional expectation has the following properties.

Claim. The conditional expectation $\mathbb{E}(Y|X)$ is a function of the random variable X and so it too is a random variable. That is, $\mathbb{E}(Y|X) = \varphi(X)$ for some function of X.

Claim. The random variable $\mathbb{E}(Y|X)$ is measurable with respect to (the σ-algebra generated by) X.

We now return to undergraduate probability to develop our next properties of conditional expectation.

Example. Suppose that (X, Y) are continuous random variables with joint density function $f(x, y)$. Consider the conditional expectation $\mathbb{E}(Y|X) = \varphi(X)$ defined by $\varphi(x) = \mathbb{E}(Y|X = x)$. Since $\mathbb{E}(Y|X)$ is a random variable, we can calculate its expectation. That is,

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(\varphi(X)) &= \int_{-\infty}^{\infty} \varphi(x) f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \mathbb{E}(Y|X=x) f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f_{Y|X=x}(y)\,\mathrm{d}y f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\,\frac{f(x,y)}{f_X(x)}\cdot f_X(x)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\right]\,\mathrm{d}y = \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y = \mathbb{E}(Y).
\end{aligned}$$

We will now show that

- if $g : \mathbb{R} \to \mathbb{R}$ is a function, then $\mathbb{E}(g(X)\,Y|X) = g(X)\mathbb{E}(Y|X)$ (taking out what is known), and
- $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ if X and Y are independent.

By definition, to verify the equality $\mathbb{E}(g(X)\,Y|X) = g(X)\mathbb{E}(Y|X)$ means that we must verify that $\mathbb{E}(g(X)\,Y|X = x) = g(x)\mathbb{E}(Y|X = x)$. Therefore, we find

$$\mathbb{E}(g(X)\,Y|X=x) = \int_{-\infty}^{\infty} g(x) y f_{Y|X=x}(y)\,\mathrm{d}y = g(x)\int_{-\infty}^{\infty} y f_{Y|X=x}(y)\,\mathrm{d}y = g(x)\mathbb{E}(Y|X=x).$$

If X and Y are independent, then

$$\begin{aligned}\mathbb{E}(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)\,\mathrm{d}y = \int_{-\infty}^{\infty} y\frac{f(x,y)}{f_X(x)}\,\mathrm{d}y &= \int_{-\infty}^{\infty} y\frac{f_X(x)f_Y(y)}{f_X(x)}\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} y f_Y(y)\,\mathrm{d}y \\ &= \mathbb{E}(Y).\end{aligned}$$

Thus, we are led to two more assertions.

Claim. The conditional expectation $\mathbb{E}(Y|X)$ has expected value

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y).$$

Claim. If $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, then

$$\mathbb{E}(g(X)\,Y|X) = g(X)\mathbb{E}(Y|X).$$

Hence, we are now ready to define the conditional expectation, and we state the fundamental result which is originally due to Kolmogorov in 1933.

Theorem. *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space, and let Y be a real-valued L^1 random variable. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. There exists a random variable Z such that*

(i) Z is $\mathcal{F}$-measurable, i.e., $Z : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$,

(ii) $\mathbb{E}(|Z|) < \infty$, i.e., $\int_\Omega |Z|\,\mathrm{d}P < \infty$, and

(iii) for every set $A \in \mathcal{F}$, we have

$$\mathbb{E}(Z1_A) = \mathbb{E}(Y1_A).$$

Moreover, Z is almost surely unique. That is, if $\tilde{Z}$ is another random variable with these properties, then $P\{\tilde{Z} = Z\} = 1$. We call Z (a version of) the conditional expectation of Y given $\mathcal{F}$ *and write*

$$Z = \mathbb{E}(Y|\mathcal{F}).$$

Remark. This definition is given for any sub-σ-algebra of $\mathcal{A}$. In particular, if X is another random variable on $(\Omega, \mathcal{A}, P)$, then $\sigma(X) \subseteq \mathcal{A}$ and so it makes sense to discuss

$$\mathbb{E}(Y|\sigma(X)) =: \mathbb{E}(Y|X).$$

Thus, our notation is consistent with undergraduate usage.

We end with one more result from undergraduate probability.

Example. Suppose that (X, Y) are continuous random variables with joint density $f(x, y)$. If A is an event of the form $A = \{a \le X \le b\}$, then $\mathbb{E}(\mathbb{E}(Y|X)1_A) = \mathbb{E}(Y1_A)$.

Solution. Since $\mathbb{E}(Y|X)$ is $\sigma(X)$-measurable, we can write $\mathbb{E}(Y|X) = \varphi(X)$ for some function φ. Thus, by the definition of expectation,

$$\mathbb{E}(\mathbb{E}(Y|X)1_A) = \mathbb{E}(\varphi(X)1_A) = \int_{-\infty}^{\infty} \varphi(x)1_A(x)\, f_X(x)\,\mathrm{d}x = \int_a^b \varphi(x)\, f_X(x)\,\mathrm{d}x.$$

However, by the definition of conditional expectation as above,

$$\varphi(x) = \mathbb{E}(Y|X = x) = \frac{\int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y}{f_X(x)}.$$

Substituting in gives

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y|X)1_A) = \int_a^b \varphi(x)\, f_X(x)\,\mathrm{d}x &= \int_a^b \left(\frac{\int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y}{f_X(x)}\right) f_X(x)\,\mathrm{d}x \\ &= \int_a^b \int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y\,\mathrm{d}x \qquad (*)\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}(Y1_A) = \int_{-\infty}^{\infty} y1_A(x) f_Y(y)\,\mathrm{d}y &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y1_A(x) f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y1_A(x) f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} 1_A(x) \int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_a^b \int_{-\infty}^{\infty} y f(x,y)\,\mathrm{d}y\,\mathrm{d}x. \qquad (**)\end{aligned}$$

Thus, comparing $(*)$ and $(**)$ we conclude that

$$\mathbb{E}(\mathbb{E}(Y|X)1_A) = \mathbb{E}(Y1_A)$$

as required.

Lecture #35: Conditional Expectation

Reference. Chapter 23 pages 197–207

Recall that last lecture we introduced the concept of conditional expectation and stated the fundamental result originally due to Kolmogorov in 1933.

Theorem (Definition of Conditional Expectation)**.** *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space, and let Y be a real-valued L^1 random variable. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. There exists a random variable Z such that*

(i) Z is $\mathcal{F}$-measurable, i.e., $Z : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B})$,

(ii) $\mathbb{E}(|Z|) < \infty$, i.e., $\int_\Omega |Z| \, \mathrm{d}P < \infty$, and

(iii) for every set $A \in \mathcal{F}$, we have

$$\mathbb{E}(Z 1_A) = \mathbb{E}(Y 1_A).$$

Moreover, Z is almost surely unique. That is, if $\tilde{Z}$ is another random variable with these properties, then $P\{\tilde{Z} = Z\} = 1$. We call Z (a version of) the conditional expectation of Y given $\mathcal{F}$ *and write*

$$Z = \mathbb{E}(Y|\mathcal{F}).$$

The proof that the conditional expectation is almost surely unique is relatively straightforward.

Theorem. *Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space and $Y : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$ is an L^1 random variable. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. If Z and $\tilde{Z}$ are versions of the conditional expectation $\mathbb{E}(Y|\mathcal{F})$, then $P(Z = \tilde{Z}) = 1$. That is, the conditional expectation $\mathbb{E}(Y|\mathcal{F})$ is almost surely unique.*

Proof. Let Z and $\tilde{Z}$ be versions of $\mathbb{E}(Y|\mathcal{F})$. Therefore, Z, $\tilde{Z} \in L^1(\Omega, \mathcal{F}, P)$ and

$$\mathbb{E}((Z - \tilde{Z})1_A) = 0 \quad \text{for every } A \in \mathcal{F}.$$

Suppose that Z and $\tilde{Z}$ are not almost surely equal and assume (without loss of generality) that we have

$$P\{Z > \tilde{Z}\} > 0.$$

Since

$$\{Z > \tilde{Z} + n^{-1}\} \uparrow \{Z > \tilde{Z}\}$$

we see that there exists an n such that

$$P\{Z > \tilde{Z} + n^{-1}\} > 0.$$

However, Z and $\tilde{Z}$ are both $\mathcal{F}$ measurable and so it follows that $\{Z > \tilde{Z} + n^{-1}\} \in \mathcal{F}$. However, if we choose $A = \{Z > \tilde{Z} + n^{-1}\}$ then

$$\mathbb{E}((Z - \tilde{Z})1_A) = \mathbb{E}((Z - \tilde{Z})1_{\{Z > \tilde{Z} + n^{-1}\}}) \geq n^{-1} P\{Z > \tilde{Z} + n^{-1}\} > 0.$$

This contradicts the fact that $\mathbb{E}((Z - \tilde{Z})1_A) = 0$ for *every* $A \in \mathcal{F}$, and so we are forced to conclude that $Z = \tilde{Z}$ almost surely. □

The proof that the conditional expectation exists is a little more involved and requires the results from Chapter 22 on L^2 and Hilbert spaces. We will not discuss this material in STAT 851 but rather assume it as given and accept that the conditional expectation exists.

We will now prove a number of results which give the most important properties of conditional expectation. In particular, these show that conditional expectations often behave like usual expectations.

Theorem. *If $Z = \mathbb{E}(Y|\mathcal{F})$, then $\mathbb{E}(Z) = \mathbb{E}(Y)$.*

Proof. This follows from (iii) by noting that $\Omega \in \mathcal{F}$ since $\mathcal{F}$ is a σ-algebra. That is,

$$\mathbb{E}(Z) = \mathbb{E}(Z1_\Omega) = \mathbb{E}(Y1_\Omega) = \mathbb{E}(Y)$$

and the proof is complete. □

Theorem. *If the random variable Y is $\mathcal{F}$-measurable, then $\mathbb{E}(Y|\mathcal{F}) = Y$ almost surely.*

Proof. This follows immediately from the definition of conditional expectation. □

Theorem (Linearity of Conditional Expectation)**.** *If Z_1 is a version of $\mathbb{E}(Y_1|\mathcal{F})$ and Z_2 is a version of $\mathbb{E}(Y_2|\mathcal{F})$, then $\alpha Z_1 + \beta Z_2$ is a version of $\mathbb{E}(\alpha Y_1 + \beta Y_2|\mathcal{F})$. Informally, this says that if $Z_1 = \mathbb{E}(Y_1|\mathcal{F})$ and $Z_2 = \mathbb{E}(Y_2|\mathcal{F})$, then*

$$\alpha Z_1 + \beta Z_2 = \mathbb{E}(\alpha Y_1 + \beta Y_2|\mathcal{F})$$

almost surely.

Proof. This result is also immediate from the definition of conditional expectation. □

The proof of the next theorem is similar to the proof above of the almost sure uniqueness of conditional expectation.

Theorem (Monotonicity of Conditional Expecation)**.** *If $Y \geq 0$ almost surely, then $\mathbb{E}(Y|\mathcal{F}) \geq 0$ almost surely.*

Proof. Let Z be a version of $\mathbb{E}(Y|\mathcal{F})$. Assume that Z is not almost surely greater than or equal to 0 so that $P\{Z < 0\} > 0$. Thus, there exists some n such that

$$P\{Z < -n^{-1}\} > 0.$$

If we set $A = \{Z < -n^{-1}\}$, then $A \in \mathcal{F}$ so that

$$0 \leq \mathbb{E}(Y1_A) = \mathbb{E}(Z1_A) < -n^{-1} P\{A\} < 0.$$

This is a contradiction and so we are forced to conclude that $Z \geq 0$ almost surely. □

Corollary. *If $Y \geq X$ almost surely, then $\mathbb{E}(Y|\mathcal{F}) \geq \mathbb{E}(X|\mathcal{F})$ almost surely.*

Theorem (Tower Property of Conditional Expectation)**.** *If $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, then*

$$\mathbb{E}(\,\mathbb{E}(Y|\mathcal{F})\,|\,\mathcal{G}) = \mathbb{E}(Y|\mathcal{G})$$

almost surely.

Proof. This result, too, is essentially immediate from the definition of conditional expectation. □

Theorem ("Taking out what is known")**.** *If X is $\mathcal{F}$-measurable and bounded, then*

$$\mathbb{E}(XY|\mathcal{F}) = X\mathbb{E}(Y|\mathcal{F})$$

almost surely.

Proof. This can be proved using the standard machine applied to X. It is left as an exercise. □

Finally, we state without proof the analogues of the main results about interchanging limits and expectations. Since conditional expectations are random variables, these are statements about almost sure limits.

Theorem (Monotone Convergence)**.** *If $Y_n \geq 0$ for all n and $Y_n \uparrow Y$, then*

$$\lim_{n\to\infty} \mathbb{E}(Y_n|\mathcal{F}) = \mathbb{E}(Y|\mathcal{F})$$

almost surely.

Theorem (Fatou's Lemma)**.** *If $Y_n \geq 0$ for all n, then*

$$\mathbb{E}\left(\liminf_{n\to\infty} Y_n\,|\,\mathcal{F}\right) \leq \liminf_{n\to\infty} \mathbb{E}(Y_n|\mathcal{F})$$

almost surely.

Theorem (Dominated Convergence)**.** *Suppose $X \in L^1$ and $|Y_n(\omega)| \leq X(\omega)$ for all n. If $Y_n \to Y$ almost surely, then*

$$\lim_{n\to\infty} \mathbb{E}(Y_n|\mathcal{F}) = \mathbb{E}(Y|\mathcal{F})$$

almost surely.

Example. Suppose that the random variable Y has a binomial distribution with parameters $p \in (0,1)$ and $n \in \mathbb{N}$. It is well-known that $\mathbb{E}(Y) = pn$. On the other hand, suppose that the parameter N is a random integer. For instance, assume that N has a Poisson distribution with parameter $\lambda > 0$. We now find that the conditional expectation $\mathbb{E}(Y|N) = pN$. Since N has mean $\mathbb{E}(N) = \lambda$, we conclude that

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|N)) = \mathbb{E}(pN) = p\mathbb{E}(N) = p\lambda.$$

We leave it as an exercise to show that

$$\mathbb{E}(N|Y) = Y + (1-p)\lambda.$$

Note that

$$\mathbb{E}(N) = \mathbb{E}(\mathbb{E}(N|Y)) = \mathbb{E}(Y + (1-p)\lambda) = \mathbb{E}(Y) + (1-p)\lambda = p\lambda + (1-p)\lambda = \lambda$$

as expected.

Lecture #36: Introduction to Martingales

Reference. Chapter 24 pages 211–216

Suppose that $(\Omega, \mathcal{A}, P)$ is a probability space. An increasing sequence of sub-σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{A}$ is often called a *filtration*.

Example. If $X_0, X_1, X_2, \ldots$ is a sequence of random variables on $(\Omega, \mathcal{A}, P)$ and we let $\mathcal{F}_j = \sigma(X_0, X_1, \ldots, X_j)$, then $\{\mathcal{F}_n,\, n = 0, 1, 2, \ldots\}$ is a filtration often called the *natural filtration*.

Definition. A sequence $X_0, X_1, X_2, \ldots$ of random variables is said to be a *martingale* with respect to the filtration $\{\mathcal{F}_n\}$ if for every $n = 0, 1, 2, \ldots$,

(i) $\mathbb{E}|X_n| < \infty$,

(ii) X_n is $\mathcal{F}_n$-measurable, and

(iii) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$.

Remark. If the filtration used is the natural one, then we sometimes write the third condition as

$$\mathbb{E}(X_{n+1}|X_0, X_1, \ldots, X_n) = X_n$$

instead.

Remark. The first two conditions in the definition of martingale are technically important. However, it is really the third condition which "defines" a martingale.

Theorem. *If* $\{X_n,\, n = 1, 2, \ldots\}$ *is a martingale, then* $\mathbb{E}(X_n) = \mathbb{E}(X_0)$ *for every* $n = 0, 1, 2, \ldots$.

Proof. Since

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) = \mathbb{E}(X_n),$$

we can use induction to conclude that

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(X_n) = \mathbb{E}(X_{n-1}) = \cdots = \mathbb{E}(X_0)$$

as required. □

Example. Suppose that $Y_1, Y_2, \ldots$ are independent, identically distributed random variables with $P\{Y_1 = 1\} = P\{Y = -1\} = 1/2$. Let $X_0 = 0$, and for $n = 1, 2, \ldots$, define $X_n = Y_1 + Y_2 + \cdots + Y_n$. The sequence $\{X_n, n = 0, 1, 2, \ldots\}$ is called a *(simple, symmetric) random walk (starting at* 0*)*. Random walks are the quintessential example of a discrete time *stochastic process*, and are the subject of much current research. We begin by computing $\mathbb{E}(X_n)$ and $\mathrm{Var}(X_n)$. Note that

$$(Y_1 + Y_2 + \cdots + Y_n)^2 = Y_1^2 + Y_2^2 + \cdots + Y_n^2 + \sum_{i \neq j} Y_i Y_j.$$

Since $\mathbb{E}(Y_1) = 0$ and $\mathrm{Var}(Y_1) = \mathbb{E}(Y_1^2) = 1$, we find

$$\mathbb{E}(X_n) = \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \cdots + \mathbb{E}(Y_n) = 0$$

and

$$\begin{aligned}\mathrm{Var}(X_n) = \mathbb{E}(X_n^2) = \mathbb{E}(Y_1 + Y_2 + \cdots + Y_n)^2 &= \mathbb{E}(Y_1^2) + \mathbb{E}(Y_2^2) + \cdots + \mathbb{E}(Y_n^2) + \sum_{i \neq j} \mathbb{E}(Y_i Y_j) \\ &= 1 + 1 + \cdots + 1 + 0 \\ &= n\end{aligned}$$

since $\mathbb{E}(Y_i Y_j) = \mathbb{E}(Y_i)\mathbb{E}(Y_j)$ when $i \neq j$ because of the assumed independence of $Y_1, Y_2, \ldots$.

We now show that the random walk $\{X_n, n = 0, 1, 2, \ldots\}$ is a martingale with respect to the natural filtration. To see that $\{X_n\}$ satisfies the first two conditions, note that $X_n \in L^2$ and so we necessarily have $X_n \in L^1$ for each n. Furthermore, X_n is by definition measurable with respect to $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$. As for the third condition, notice that

$$\begin{aligned}\mathbb{E}(X_{n+1}|\mathcal{F}_n) &= \mathbb{E}(Y_{n+1} + X_n|\mathcal{F}_n) \\ &= \mathbb{E}(Y_{n+1}|\mathcal{F}_n) + \mathbb{E}(X_n|\mathcal{F}_n).\end{aligned}$$

Since Y_{n+1} is independent of $X_1, X_2, \ldots, X_n$ we conclude that

$$\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = \mathbb{E}(Y_{n+1}) = 0.$$

If we condition on $X_1, X_2, \ldots, X_n$ then X_n is *known* and so

$$\mathbb{E}(X_n|\mathcal{F}_n) = X_n.$$

Combined we conclude

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \mathbb{E}(Y_{n+1}|\mathcal{F}_n) + \mathbb{E}(X_n|\mathcal{F}_n) = 0 + X_n = X_n$$

which proves that $\{X_n, n = 0, 1, 2, \ldots\}$ is a martingale.

Next we show that $\{X_n^2 - n, n = 0, 1, 2, \ldots\}$ is also a martingale with respect to the natural filtration. In other words, if we let $M_n = X_n^2 - n$, then we must show that $\mathbb{E}(M_{n+1}|\mathcal{F}_n) = M_n$. Clearly M_n is $\mathcal{F}_n$-measurable and satisfies

$$\mathbb{E}|M_n| \leq \mathbb{E}(X_n^2) + n = 2n < \infty.$$

As for the third condition, notice that

$$\begin{aligned}\mathbb{E}(X_{n+1}^2|\mathcal{F}_n) &= \mathbb{E}((Y_{n+1} + X_n)^2|\mathcal{F}_n) \\ &= \mathbb{E}(Y_{n+1}^2|\mathcal{F}_n) + 2\mathbb{E}(Y_{n+1}X_n|\mathcal{F}_n) + \mathbb{E}(X_n^2|\mathcal{F}_n).\end{aligned}$$

However,

- $\mathbb{E}(Y_{n+1}^2|\mathcal{F}_n) = \mathbb{E}(Y_{n+1}^2) = 1$,

- $\mathbb{E}(Y_{n+1}X_n|\mathcal{F}_n) = X_n\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = X_n\mathbb{E}(Y_{n+1}) = 0$, and
- $\mathbb{E}(X_n^2|\mathcal{F}_n) = X_n^2$

from which we conclude that

$$\mathbb{E}(X_{n+1}^2|\mathcal{F}_n) = X_n^2 + 1.$$

Therefore,

$$\begin{aligned}\mathbb{E}(M_{n+1}|\mathcal{F}_n) &= \mathbb{E}(X_{n+1}^2 - (n+1)|\mathcal{F}_n)\\ &= \mathbb{E}(X_{n+1}^2|\mathcal{F}_n) - (n+1)\\ &= X_n^2 + 1 - (n+1)\\ &= X_n^2 - n\\ &= M_n\end{aligned}$$

and so we conclude that $\{M_n, n = 0, 1, 2, \ldots\} = \{X_n^2 - n, n = 0, 1, 2, \ldots\}$ is a martingale.

Gambler's Ruin for Random Walk

Suppose that $\{X_n, n = 0, 1, 2, \ldots\}$ is a martingale. As shown above,

$$\mathbb{E}(X_n) = \mathbb{E}(X_{n-1}) = \cdots = \mathbb{E}(X_1) = \mathbb{E}(X_0)$$

which, in other words, says that a martingale has *stable expectation*. The fact that

$$\mathbb{E}(X_n) = \mathbb{E}(X_0) \qquad (*)$$

is extremely useful as the following example shows.

Example. Suppose that $\{S_n, n = 0, 1, 2, \ldots\}$ is a simple random walk starting from 0. As we have already shown, both $\{S_n, n = 0, 1, 2, \ldots\}$ and $\{S_n^2 - n, n = 0, 1, 2, \ldots\}$ are martingales. Hence,

$$\mathbb{E}(S_n^2 - n) = \mathbb{E}(S_0^2 - 0) = 0 \quad \text{and so} \quad \mathbb{E}(S_n^2) = n.$$

Suppose now that T is a stopping time. The question of whether or not we obtain an expression like $(*)$ by replacing n with T is very deep.

Example. Suppose that $\{S_n, n = 0, 1, 2, \ldots\}$ is a simple random walk starting from 0. Let T denote the first time that the simple random walk reaches 1. Since S_n is a martingale, we know that

$$\mathbb{E}(S_n) = \mathbb{E}(S_0) = 0.$$

However, by definition, $S_T = 1$ and so

$$\mathbb{E}(S_T) = 1 \neq \mathbb{E}(S_0).$$

In order to obtain $\mathbb{E}(X_T) = \mathbb{E}(X_0)$, we need more structure on T. The following result provides such a structure.

Theorem (Doob's Optional Stopping Theorem)**.** *Suppose that* $\{X_n, n = 0, 1, 2, \ldots\}$ *is a martingale with the property that there exists a* $K < \infty$ *with* $|X_n| < K$ *for all* n*. (That is, suppose that* $\{X_n, n = 0, 1, \ldots\}$ *is uniformly bounded.) If* T *is a stopping time with* $P\{T < \infty\} = 1$*, then*

$$\mathbb{E}(X_T) = \mathbb{E}(X_0).$$

Example. Suppose that $\{S_n, n = 0, 1, 2, \ldots\}$ is a simple random walk except that it starts at $S_0 = x$. Let T denote the first time that the SRW reaches level N or 0. Note that $|S_n| \leq N$ for all n and that this stopping time T satisfies $P\{T < \infty\} = 1$. Therefore, by the optional stopping theorem,

$$\mathbb{E}(S_T) = \mathbb{E}(S_0) = x.$$

However, we can calculate

$$\mathbb{E}(S_T) = 0 \cdot P\{S_T = 0\} + N \cdot P\{S_T = N\}.$$

This implies

$$N \cdot P\{S_T = N\} = x \quad \text{or} \quad P\{S_T = N\} = \frac{x}{N}$$

and so

$$P\{S_T = 0\} = 1 - \frac{x}{N} = \frac{N - x}{N}.$$

Suppose that $Y_1, Y_2, \ldots$ are independent and identically distributed random variables with $P\{Y_1 = 1\} = p$, $P\{Y_1 = -1\} = 1 - p$ for some $0 < p < 1/2$.

Define the *biased random walk* $\{S_n, n = 0, 1, 2, \ldots\}$ by setting $S_0 = 0$ and

$$S_n = Y_1 + \cdots + Y_n = \sum_{j=1}^{n} Y_j$$

for $n = 1, 2, 3, \ldots$.

Unlike the simple random walk, the biased random walk is NOT a martingale. Notice that

$$\mathbb{E}(S_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n + Y_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n|\mathcal{F}_n) + \mathbb{E}(Y_{n+1}|\mathcal{F}_n) = S_n + \mathbb{E}(Y_{n+1})$$

where the last equality follows by "taking out what is known" and using the fact that Y_{n+1} is independent of S_n. However,

$$\mathbb{E}(Y_{n+1}) = 1 \cdot P\{Y_{n+1} = 1\} + (-1) \cdot P\{Y_1 = -1\} = p - (1 - p) = 2p - 1$$

so that

$$\mathbb{E}(S_{n+1}|\mathcal{F}_n) = S_n + (2p - 1).$$

However, if we define $\{X_n, n = 0, 1, 2, \ldots\}$ by setting $X_n = S_n - n(2p - 1)$, then $\{X_n, n = 0, 1, 2, \ldots\}$ *is* a martingale.

Exercise. Show that $\{Z_n, n = 0, 1, 2, \ldots\}$ defined by setting

$$Z_n = \left(\frac{1 - p}{p}\right)^{S_n}$$

is a martingale.

Now, suppose that we are playing a casino game. Obviously, the game is designed to be in the house's favour, and so we can assume that the probability you win on a given play of the game is p where $0 < p < 1/2$. Furthermore, if we assume that the game pays even money on each play and that the plays are independent, then we can model our "fortune" after n plays by $\{S_n, n = 0, 1, 2, \ldots\}$ where S_0 represents our initial fortune.

Example. Consider the standard North American roulette game that has 38 numbers of which 18 are black, 18 are red, and 2 are green. A bet on "black" pays even money. That is, a bet of \$1 on black pays \$1 plus your original \$1 if black does, in fact, appear. However, the probability that you win on a bet of "black" is $p = 18/38 \approx 0.4737$.

Example. The casino game of craps includes a number of sophisticated bets. One simple bet that pays even money is the "pass line bet." The probability that you win on a "pass line bet" is $p = 244/495 \approx 0.4899$.

Let's suppose that we enter the casino with $\$x$ and that we decide to stop playing either when we go broke or when we win $\$N$ (whichever happens first).

Hence, our "fortune process" is $\{S_n, n = 0, 1, 2, \ldots\}$ where $S_0 = x$.

If we define the stopping time $T = \min\{n : S_n = 0 \text{ or } N\}$, then since both the martingales $\{X_n, n = 0, 1, 2, \ldots\}$ and $\{Z_n, n = 0, 1, 2, \ldots\}$ are uniformly bounded, we can apply the optional sampling theorem.

Thus, $\mathbb{E}(X_0) = \mathbb{E}(X_T)$ implies $x = \mathbb{E}(X_T) = \mathbb{E}(S_T - T(2p-1))$ and so

$$\mathbb{E}(T) = \frac{\mathbb{E}(S_T) - x}{2p - 1}$$

since $\mathbb{E}(X_0) = \mathbb{E}(S_0) = x$. However, we cannot calculate $\mathbb{E}(S_T)$ directly. That is, although

$$\mathbb{E}(S_T) = 0 \cdot P\{S_T = 0\} + N \cdot P\{S_T = N\},$$

we do not have any information about S_T.

Fortunately, we can use the martingale $\{Z_n, n = 0, 1, 2, \ldots\}$ to figure out $\mathbb{E}(S_T)$. Applying the optional stopping theorem implies $\mathbb{E}(Z_0) = \mathbb{E}(Z_T)$ and so

$$\left(\frac{1-p}{p}\right)^x = \mathbb{E}\left(\left(\frac{1-p}{p}\right)^{S_T}\right).$$

Now,

$$\begin{aligned}
\mathbb{E}\left(\left(\frac{1-p}{p}\right)^{S_T}\right) &= \left(\frac{1-p}{p}\right)^0 P\{S_T = 0\} + \left(\frac{1-p}{p}\right)^N P\{S_T = N\} \\
&= P\{S_T = 0\} + \left(\frac{1-p}{p}\right)^N (1 - P\{S_T = 0\})
\end{aligned}$$

implying that

$$\left(\frac{1-p}{p}\right)^x = P\{S_T = 0\} + \left(\frac{1-p}{p}\right)^N (1 - P\{S_T = 0\}).$$

Solving for $P\{S_T = 0\}$ gives

$$P\{S_T = 0\} = \frac{\left(\frac{1-p}{p}\right)^x - \left(\frac{1-p}{p}\right)^N}{1 - \left(\frac{1-p}{p}\right)^N}.$$

We therefore find that

$$P\{S_T = N\} = 1 - \frac{\left(\frac{1-p}{p}\right)^x - \left(\frac{1-p}{p}\right)^N}{1 - \left(\frac{1-p}{p}\right)^N} = \frac{1 - \left(\frac{1-p}{p}\right)^x}{1 - \left(\frac{1-p}{p}\right)^N}$$

which implies that

$$\mathbb{E}(T) = \frac{NP\{S_T = N\} - x}{2p-1} = \frac{1}{2p-1}\left(\frac{N - N\left(\frac{1-p}{p}\right)^x}{1 - \left(\frac{1-p}{p}\right)^N} - x\right).$$

Example. Suppose that $p = 18/38$, $x = 10$ and $N = 20$. Substituting into the above formulæ gives

$$P\{S_T = 0\} = \frac{10\ 000\ 000\ 000}{13\ 486\ 784\ 401} \approx 0.7415$$

and

$$\mathbb{E}(T) = \frac{1\ 237\ 510\ 963\ 810}{13\ 486\ 784\ 401} \approx 91.76.$$