Graduate Level Probability Review

Michael J. Neely
https://viterbi-web.usc.edu/~mjneely/

An updated version of these notes is on the following clickable link:
https://ee.usc.edu/stochastic-nets/docs/probability-review.pdf

These notes provide a focused probability review.[1] There are about 200 exercises, some of which the author has given in past homework assignments and past graduate screening exams on the probability topic (EE 503).[2] Students taking the EE 503 screening exam can also refer to textbooks such as [1][2]. Screening exam students are expected to: Demonstrate good reasoning; Solve problems involving events, random variables, conditioning, and expectation; Prove basic results such as the law of total probability and the fact that if events A and B are independent, then A and B^c are also independent.

Appendices B, C, D, E contain advanced material that is not intended for the screening exam, but that may be of interest (see also advanced probability texts such as [3][4][5]).

I. Foundations

Probability theory involves sets, sets of sets, and mappings between sets. For example, the set Ω of all *outcomes* of a probability experiment is called the *sample space*. The set $\mathcal{F}$ of all *events* is a set of subsets of Ω that has a particular *sigma algebra* structure (a sigma algebra is also called a *sigma field*). A *probability measure* is a function that maps the set $\mathcal{F}$ to the set $[0,1]$ and satisfies particular axioms. A random variable is a function that maps the set Ω to the set $\mathbb{R}$ and that has a certain measurability property with respect to $\mathcal{F}$. Thus, set theory is important for the foundations of probability.

A. Sets

A set is a collection of objects called *elements*. The elements can have any structure: They can be numbers, vectors, colors, letters, or even sets themselves. The *empty set* is the set that contains no elements and is denoted ϕ.

Let A and B be two nonempty sets. A function $f : A \to B$ is a mapping that assigns each $a \in A$ to a value $f(a) \in B$. A function $f : A \to B$ is said to be *injective* (also called *one-to-one*) if for all elements $a, b \in A$ we have

$$f(a) = f(b) \implies a = b$$

A function $f : A \to B$ is called *surjective* (also called *onto*) if for all $b \in B$ there is at least one $a \in A$ such that $f(a) = b$. A function $f : A \to B$ is called *bijective* if it is both injective and surjective. A bijective function is also called a *one-to-one correspondence*.

A set is *finite* if its number of elements is a nonnegative integer (the empty set ϕ is a finite set with zero elements). That is, a set is finite if and only if it is either empty or can be put into one-to-one correspondence with the set $\{1, ..., n\}$ for some positive integer n. A set is *infinite* if it is not finite. Let $\mathbb{N}$ denote the (infinite) set of positive integers:

$$\mathbb{N} = \{1, 2, 3, ...\}$$

A set A is said to be *countably infinite* if it can be put into one-to-one correspondence with $\mathbb{N}$. That is, set A is countably infinite if there exists a bijective function $f : \mathbb{N} \to A$. From this definition, it follows that a set A is countably infinite if and only if its elements can be *listed*, so that

$$A = \{f(i)\}_{i=1}^{\infty} = \{f(1), f(2), f(3), \ldots\}$$

This means the list contains only elements of A, and each element of A is on the list once and only once.

A countably infinite set is indeed an infinite set. An infinite set that is not countably infinite is called *uncountably infinite*. Such sets are so large that their elements cannot be listed. The following can be shown.

- Countably infinite sets include $\mathbb{N}$ (the positive integers), $\mathbb{Z}$ (the integers), $\mathbb{Q}$ (the rational numbers). For $k \in \mathbb{N}$, the sets $\mathbb{N}^k, \mathbb{Z}^k, \mathbb{Q}^k$ are also countably infinite. For example, $\mathbb{N}^2 = \mathbb{N} \times \mathbb{N} = \{(a, b) : a, b \in \mathbb{N}\}$ is countably infinite.
- Uncountably infinite sets include $\mathbb{R}$, $\mathbb{R}^k$ (for any positive integer k), the intervals $[0,1]$, $[0,1)$, $(0,1)$, $(0,1]$, and the set $\{0,1\}^{\mathbb{N}}$ of all infinite binary sequences:

$$\{0,1\}^{\mathbb{N}} = \{(b_1, b_2, b_3, ...) : b_i \in \{0,1\} \quad \forall i \in \mathbb{N}\}$$

[1] Please contact me if you spot any typos or mistakes so I can fix them and update [teh] PDF file.

[2] For example, Exercises 17, 18 in Section I-K and Exercises 1,3,4,5 in Section III-R were originally developed for oral screening exams. In the Fall 2023 exam, I asked the Section I-K question 17 verbatim, then asked students to define mutual independence and show it holds for $A_1, A_2^c, A_3^c, A_4, A_5^c$. I did not give the "sequential complement" hint that I give in question 18. Valid approaches some students used were: (i) Make an induction argument; (ii) First consider only three events to gain intuition. The most elegant approach is to make an observation about sequential complements, as suggested in question 18.

B. Complements and De Morgan's law

Let Ω be a nonempty set. If $A \subseteq \Omega$ we define the complement of A with respect to Ω, written A^c, by

$$A^c = \{\omega \in \Omega : \omega \notin A\} \tag{1}$$

It is clear that $(A^c)^c = A$. Let ϕ denote the *empty set*, defined as the set with no elements. It holds that $\phi \subseteq \Omega$. Also, $\phi^c = \Omega$ and $\Omega^c = \phi$.

If $\{A_i\}_{i=1}^{\infty}$ is a sequence of subsets of Ω, so $A_i \subseteq \Omega$ for all $i \in \{1, 2, 3, \ldots\}$, then *De Morgan's law* states:

$$(\cup_{i=1}^{\infty} A_i)^c = \cap_{i=1}^{\infty} A_i^c$$

More generally, let I be an arbitrary nonempty set (possibly finite, countably infinite, or uncountably infinite). Set I shall be called an *index set* because its elements i are used as subscripts. Specifically, for each $i \in I$, let $A_i \subseteq \Omega$. Then

$$(\cup_{i \in I} A_i)^c = \cap_{i \in I} A_i^c$$

C. Sigma algebra

Let Ω be a nonempty set. Let $\mathcal{F}$ be a set of subsets of Ω. We often use the word *collection* when talking about a set of sets, so $\mathcal{F}$ can also be called a *collection* of subsets of Ω (the words "collection" and "set" have the same meaning here). We say that $\mathcal{F}$ is a *sigma algebra on* Ω if the following three properties hold:

- $\Omega \in \mathcal{F}$.
- If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
- If $\{A_i\}_{i=1}^{\infty}$ satisfy $A_i \in \mathcal{F}$ for all $i \in \{1, 2, 3, \ldots\}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The second bullet assumes A^c is defined with respect to Ω, as in (1). These properties can be used to show: $\phi \in \mathcal{F}$; the union of a finite or countably infinite number of sets in $\mathcal{F}$ is another set in $\mathcal{F}$; the intersection of a finite or countably infinite number of sets in $\mathcal{F}$ is another set in $\mathcal{F}$. The smallest sigma algebra on Ω is the 2-element set $\mathcal{F} = \{\phi, \Omega\}$. The largest sigma algebra on Ω is $\mathcal{F} = Pow(\Omega)$, where $Pow(\Omega)$ is called the *power set* of Ω and is defined as the set of all subsets of Ω. A sigma algebra is also called a *sigma field*.

For an example, let

$$\Omega = \{red, blue, green\}$$

The power set of Ω has 8 elements:

$$Pow(\Omega) = \{\phi, \Omega, \{red\}, \{blue\}, \{green\}, \{red, blue\}, \{red, green\}, \{blue, green\}\}$$

Define two smaller collections of subsets of Ω

$$\mathcal{F}_1 = \{\phi, \Omega, \{red\}, \{red, blue\}\}$$
$$\mathcal{F}_2 = \{\phi, \Omega, \{red\}, \{blue, green\}\}$$

The collection $\mathcal{F}_1$ is not a sigma algebra on Ω because $\{red\} \in \mathcal{F}_1$ but $\{red\}^c = \{blue, green\} \notin \mathcal{F}_1$, so property 2 of a sigma algebra is not satisfied. On the other hand, it can be shown that $\mathcal{F}_2$ is a sigma algebra on Ω. This is done by showing the three properties required of a sigma algebra are satisfied.

D. Axioms of probability

Let Ω be a nonempty set and let $\mathcal{F}$ be a sigma algebra on Ω. A function $P : \mathcal{F} \to \mathbb{R}$ is called a *probability measure* if it satisfies the following three axioms of probability:

1) $P[A] \geq 0$ for all $A \in \mathcal{F}$.
2) $P[\Omega] = 1$.
3) If $\{A_i\}_{i=1}^{\infty}$ is a sequence of *disjoint* sets in $\mathcal{F}$, so that $A_i \cap A_j = \phi$ for all $i \neq j$, then

$$P[\cup_{i=1}^{\infty} A_i] = \sum_{i=1}^{\infty} P[A_i]$$

In the third axiom, we are allowed to evaluate the probability function at $\cup_{i=1}^{\infty} A_i$ because $\mathcal{F}$ is a sigma algebra and so $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. The third axiom of probability is called *countable summability*. It can be shown that countable summability implies the following *finite summability* result: If $\{A_i\}_{i=1}^{\infty}$ is a sequence of disjoint sets in $\mathcal{F}$ then

$$P[\cup_{i=1}^{n} A_i] = \sum_{i=1}^{n} P[A_i] \quad \forall n \in \{1, 2, 3, ...\} \tag{2}$$

This can be shown by using countable summability and defining $A_i = \phi$ for all $i \geq n + 1$. While countable summability implies finite summability, finite summability does not imply countable summability (see Appendix E).

The axioms of probability can be used to show that if $A \in \mathcal{F}$ then[3]

$$P[A] + P[A^c] = 1 \tag{3}$$

Since Axiom 1 implies probabilities cannot be negative, we have $P[A] \geq 0$ and $P[A^c] \geq 0$ so (3) implies

$$0 \leq P[A] \leq 1 \quad \forall A \in \mathcal{F}$$

Therefore, the function $P : \mathcal{F} \rightarrow \mathbb{R}$ shall henceforth be viewed as a function $P : \mathcal{F} \rightarrow [0, 1]$. Since $\phi = \Omega^c$, the formula (3) implies $P[\Omega] + P[\phi] = 1$, and since Axiom 1 implies $P[\Omega] = 1$ we obtain

$$P[\phi] = 0$$

E. Probability space

A *probability space* is defined by a triplet $(\Omega, \mathcal{F}, P)$ where Ω is a nonempty set called the *sample space*; $\mathcal{F}$ is a sigma algebra on Ω, and $P : \mathcal{F} \rightarrow \mathbb{R}$ is a probability measure.

- Elements of Ω are called *outcomes*.
- Sets in $\mathcal{F}$ are called *events*. Every event $A \in \mathcal{F}$ has a probability value $P[A]$.
- Subsets of Ω that are not in $\mathcal{F}$ are not events and are not assigned probabilities. They are called *nonmeasurable sets*.

Unless otherwise specified, when working on a particular probability problem, all events are assumed to be *on the same probability space*. This means there is one probability space $(\Omega, \mathcal{F}, P)$ and all events of interest are sets in $\mathcal{F}$.

F. Intuitive interpretation

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let A be an event. An event is a subset of Ω that is in the collection $\mathcal{F}$. We often say things such as "if event A is true" or "if event A is false." How can a subset be "true" or "false"? It cannot. When we say "event A is true" we are imagining a situation where nature "randomly" selects an outcome $\omega \in \Omega$. With this thought experiment, "event A is true" has the interpretation "the randomly selected outcome ω is in the set A." Probability theory does not rely on such thought experiments. In fact, probability theory never defines the concept of "randomly choosing an outcome." However, such thought experiments are important for using probability theory to understand real world scenarios such as flipping coins or rolling dice, where the probability measure P is thoughtfully constructed to assign values to each event A according to our intuitive understanding of the "likelihood" of that event. The three axioms of probability also correspond to this intuition.

G. Consequences of the axioms

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Recall that a set $A \subseteq \Omega$ is an *event* if and only if $A \in \mathcal{F}$. The axioms of probability can be used to prove the following:

- If A and B are events then
$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \tag{4}$$
- Subset bound: If A and B are events that satisfy $A \subseteq B$, then $P[A] \leq P[B]$.
- Union bound: If $\{A_i\}_{i=1}^{\infty}$ are events then:[4]
$$P[\cup_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} P[A_i] \quad \forall n \in \{1, 2, 3, ...\} \tag{5}$$
and
$$P[\cup_{i=1}^{\infty} A_i] \leq \sum_{i=1}^{\infty} P[A_i] \tag{6}$$

The axioms can also be used to prove a result called *continuity of probability*. To state this result, let $\{A_n\}_{n=1}^{\infty}$ be a sequence of subsets of Ω. Let A be another subset of Ω.

Definition 1: We say $A_n \nearrow A$ if

$$A_n \subseteq A_{n+1} \quad \forall n \in \{1, 2, 3, ...\}$$
$$\cup_{n=1}^{\infty} A_n = A$$

[3]If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ (because $\mathcal{F}$ is a sigma algebra). Since A, A^c are disjoint sets in $\mathcal{F}$, Axiom 3 implies $P[A \cup A^c] = P[A] + P[A^c]$. Since $\Omega = A \cup A^c$, Axiom 2 implies $1 = P[\Omega] = P[A \cup A^c]$.

[4]The finite version of the union bound in (5) can be proven from (4) by induction. The countably infinite version of the union bound in (6) *cannot be proven using only* (4) (see Appendix E). The countable version of the union bound requires countable summability and/or continuity of probability.

Definition 2: We say $A_n \searrow A$ if

$$A_n \supseteq A_{n+1} \quad \forall n \in \{1, 2, 3, ...\}$$
$$\cap_{n=1}^{\infty} A_n = A$$

Theorem 1: (Continuity of probability) Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events and let A be a subset of Ω.
a) If $A_n \nearrow A$ then A is an event and

$$\lim_{n\to\infty} P[A_n] = P[A]$$

b) If $A_n \searrow A$ then A is an event and

$$\lim_{n\to\infty} P[A_n] = P[A]$$

Proof: To prove part (a), assume $\{A_n\}_{n=1}^{\infty}$ are events and $A_n \nearrow A$. Then $A = \cup_{n=1}^{\infty} A_n$ is a countable union of sets in $\mathcal{F}$ and so (by properties of the sigma algebra $\mathcal{F}$) must also be in $\mathcal{F}$. That is, A must be an event. Define the events

$$\tilde{A}_1 = A_1$$
$$\tilde{A}_n = A_n \cap A_{n-1}^c \quad \forall n \in \{2, 3, 4, \ldots\}$$

It can be shown that $\{\tilde{A}_i\}_{i=1}^{\infty}$ are disjoint. For $n \geq 2$ we have $A_{n-1} \subseteq A_n$ and so

$$A_n = A_{n-1} \cup (A_n \cap A_{n-1}^c) = A_{n-1} \cup \tilde{A}_n$$

and so by induction on n it holds that

$$A_n = \cup_{i=1}^{n} \tilde{A}_i \quad \forall n \in \{1, 2, 3, \ldots\} \tag{7}$$

Therefore

$$A = \cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} \cup_{i=1}^{n} \tilde{A}_i = \cup_{i=1}^{\infty} \tilde{A}_i \tag{8}$$

Equalities (7) and (8) represent A_n and A as countable unions of disjoint events, so by Axiom 3

$$P[A_n] = \sum_{i=1}^{n} P[\tilde{A}_i] \quad \forall n \in \{1, 2, 3, \ldots\}$$
$$P[A] = \sum_{i=1}^{\infty} P[\tilde{A}_i]$$

By definition of an infinite sum, $\sum_{i=1}^{\infty} P[\tilde{A}_i] = \lim_{n\to\infty} \sum_{i=1}^{n} P[\tilde{A}_i]$. Thus

$$P[A] = \lim_{n\to\infty} \sum_{i=1}^{n} P[\tilde{A}_i] = \lim_{n\to\infty} P[A_n]$$

This proves part (a). Part (b) can be obtained from part (a) and De Morgan's law, so $A_n \searrow A \implies A_n^c \nearrow A^c$. □

A special case of continuity of probability is that if $\{A_i\}_{i=1}^{\infty}$ is a sequence of events then

$$\lim_{n\to\infty} P[\cup_{i=1}^{n} A_i] = P[\cup_{i=1}^{\infty} A_i]$$
$$\lim_{n\to\infty} P[\cap_{i=1}^{n} A_i] = P[\cap_{i=1}^{\infty} A_i]$$

H. Conditioning and the law of total probability

Fix a $(\Omega, \mathcal{F}, P)$ as a probability space. If A and B are events with $P[B] > 0$ then the *conditional probability of A given B* is defined

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

We say that a sequence of events $\{B_i\}_{i=1}^{\infty}$ *partition the sample space* if they are disjoint and their union is Ω:

$$B_i \cap B_j = \phi \quad \forall i \neq j$$
$$\cup_{i=1}^{\infty} B_i = \Omega$$

The *law of total probability* states that if A is an event and $\{B_i\}_{i=1}^{\infty}$ are events that partition the sample space then

$$\boxed{P[A] = \sum_{i=1}^{\infty} P[A \cap B_i] = \sum_{i:P[B_i]>0} P[A|B_i]P[B_i]}$$

This formula is proven by observing

$$A = A \cap \Omega = A \cap (\cup_{i=1}^{\infty} B_i) = \cup_{i=1}^{\infty} (A \cap B_i)$$

and so

$$P[A] = P\left[\cup_{i=1}^{\infty} (A \cap B_i)\right]$$

Now apply the third axiom of probability to the sequence of disjoint events $\{A \cap B_i\}_{i=1}^{\infty}$.

I. Independence of events

Fix a $(\Omega, \mathcal{F}, P)$ as a probability space.

- Two events A and B are said to be *independent* if

$$P[A \cap B] = P[A]P[B]$$

- Fix n as a positive integer. A finite collection of events $A_1, A_1, ..., A_n$ are said to be *mutually independent* if

$$P[\cap_{i\in I} A_i] = \prod_{i\in I} P[A_i]$$

for all subsets $I \subseteq \{1, ..., n\}$. Pairwise independence does not imply mutual independence (see item 25 in Section I-K).

- A countably infinite collection of events $\{A_i\}_{i=1}^{\infty}$ are *mutually independent* if $\{A_i\}_{i=1}^{n}$ is mutually independent for every positive integer n.

J. Countable versus uncountable sample spaces

A set that is finite or countably infinite is called a *countable set* (also called a *discrete set*). When the sample space Ω is a countable set, we typically use the sigma algebra $\mathcal{F} = Pow(\Omega)$, so all subsets of Ω are events (using $\mathcal{F} = Pow(\Omega)$ is so common that it is often unstated). When Ω is countable, every probability measure $P : Pow(\Omega) \to [0, 1]$ has the form

$$P[A] = \sum_{\omega\in A} P[\{\omega\}] \quad \forall A \subseteq \Omega \tag{9}$$

so the probability measure is determined by the *probability masses* $P[\{\omega\}]$ for all $\omega \in \Omega$.[5] To prove (9), fix $A \subseteq \Omega$. Since $A = \cup_{\omega\in A}\{\omega\}$ is a countable union of disjoint events, we have by Axiom 3

$$P[A] = P[\cup_{\omega\in A}\{\omega\}] = \sum_{\omega\in A} P[\{\omega\}] \tag{10}$$

The derivation (10) does not work when A is uncountably infinite because: (i) $\cup_{\omega\in A}\{\omega\}$ is not a countable union; (ii) We do not define summation over uncountably many terms.[6] When working with uncountably infinite sample spaces Ω we typically *do not use* $\mathcal{F} = Pow(\Omega)$, *not because we do not want to, but because (in most cases of interest) nobody knows how to do it without contradicting the axioms of probability.*[7] This is why the theory of sigma algebras is important for the foundations of probability. This issue mainly arises when random variables are studied and motivates the *Borel sigma algebra over* $\mathbb{R}$, *denoted* $\mathcal{B}(\mathbb{R})$ (see Appendix D). It can be shown that any probability space that houses a random variable with a continuous CDF must have an uncountably infinite sample space (see item 35 in Section III-H). Some very good probability textbooks (including [2]) avoid sigma algebras by *pretending* we can always use $\mathcal{F} = Pow(\Omega)$, even when Ω is uncountably infinite. This does not create significant problems because the standard Borel sigma algebras on $\mathbb{R}$ and $\mathbb{R}^n$ are so extensive that they (arguably) contain all sets of "practical interest," so contradictory cases that may exist tend not to arise.[8]

K. Exercises on foundations and axioms

1) Show that $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$ for $x \in \mathbb{R}$ is neither injective nor surjective.
2) Let $f : \mathbb{N}^2 \to \mathbb{N}$ be a bijective function. Define $g : \mathbb{N}^3 \to \mathbb{N}$ by $g(a, b, c) = f(f(a, b), c)$ for all $a, b, c \in \mathbb{N}$. Show that g is a bijective function.
3) Define $\Omega = \{a, b, c\}$. Explain why the collection of sets $\{\phi, \Omega, \{b\}\}$ is not a sigma algebra on Ω. Add the smallest number of sets to make it a sigma algebra. (The resulting collection is called the *sigma algebra on* Ω *generated by* $\{\phi, \Omega, \{b\}\}$.)
4) Define $\Omega = \{a, b, c\}$. Is the collection of sets $\{\phi, \Omega, \{a\}, \{b, c\}, \{c\}, \{a, b\}\}$ a sigma algebra on Ω? If not, add the smallest number of sets to make it a sigma algebra.
5) Suppose $\Omega = \{a, b, c, d\}$ and $\mathcal{F} = Pow(\Omega)$. Suppose $P[\{a, b\}] = 0.4$, $P[\{b, c, d\}] = 0.6$, $P[\{a, c\}] = 0.7$. Find the probability masses $P[\{a\}], P[\{b\}], P[\{c\}], P[\{d\}]$.
6) Suppose $\Omega = \{green, red, blue, purple\}$ and $\mathcal{F} = Pow(\Omega)$. Suppose $P[\{blue, purple\}] = 0.3$, $P[\{red, blue, purple\}] = 0.3$, $P[\{green, blue\}] = 0.7$. Find the probability masses $P[\{green\}], P[\{red\}], P[\{blue\}], P[\{purple\}]$.

[5] It can be shown the function $P : Pow(\Omega) \to [0, 1]$ in (9) is a probability measure if and only if the values $P[\{\omega\}]$ are nonnegative and sum to 1.

[6] Here is a related fact that explains why any attempt to define the summation of an uncountably infinite number of positive real numbers would need to define the sum as ∞: Let A be an uncountably infinite set and let $f : A \to (0, \infty)$ be a function. Then, there must be a countably infinite subset $B \subseteq A$ such that $\sum_{b\in B} f(b) = \infty$. Indeed, for each positive integer i define $A_i = \{a \in A : f(a) \geq 1/i\}$ and note that $A = \cup_{i=1}^{\infty} A_i$, so at least one of the sets A_i must be infinite (and this infinite set must contain a countably infinite subset B).

[7] See chapter 3 in [3] for two impossibility theorems that prove nonexistence of a probability measure $P : Pow((0, 1]) \to [0, 1]$ with $P[\{\omega\}] = 0$ for all $\omega \in (0, 1]$ (one that uses the axiom of choice and a translation invariance assumption, another that uses the axiom of choice and the continuum hypothesis).

[8] Counter-intuitively, even though nonBorel subsets of $\mathbb{R}$ tend not to arise in practical problems, there are more nonBorel subsets of $\mathbb{R}$ than Borel subsets of $\mathbb{R}$. Specifically, under mild axiom of choice assumptions, it can be proven that $\mathcal{B}(\mathbb{R})$ can be put into one-to-one correspondence with $\mathbb{R}$, while the nonBorel subsets of $\mathbb{R}$ can be put into one-to-one correspondence with $Pow(\mathbb{R})$, see [6].

7) In a probability triplet $(\Omega, \mathcal{F}, P)$, the set Ω is the
   a) set of probabilities
   b) set of events
   c) sample space
   d) set of expectations
8) In a probability triplet $(\Omega, \mathcal{F}, P)$, the collection $\mathcal{F}$ is the
   a) set of outcomes
   b) set of events
   c) sample space
   d) set of expectations
9) For a probability space $(\Omega, \mathcal{F}, P)$, every event
   a) is a subset of Ω
   b) is a set in the collection $\mathcal{F}$
   c) has a probability
   d) all of the above
10) For a probability space $(\Omega, \mathcal{F}, P)$, if $A \subseteq \Omega$ but A is not an event, then
   a) $P[A]$ does not exist
   b) $A \notin \mathcal{F}$
   c) $A^c \notin \mathcal{F}$
   d) all of the above
11) Two events A and B can be used to partition the sample space into how many disjoint events? Draw the Venn diagram.
12) Fix $n \in \mathbb{N}$ and let $A_1, ..., A_n$ be events. These can be used to partition the sample space into how many disjoint events?
13) Let A and B be events. Prove $A \subseteq B \implies P[A] \leq P[B]$.
14) Let A and B be events. Argue that $A = (A \cap B) \cup (A \cap B^c)$.
15) Let A and B be events. Prove $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.
16) Let A and B be events. Prove that $P[A \cap B] \geq P[A] - P[B^c]$.
17) Suppose A and B are independent events. Prove that A and B^c are independent events.
18) Fix $n \in \mathbb{N}$ and suppose $A_1, \ldots, A_n$ are mutually independent events. Show $A_1, ..., A_{n-1}, A_n^c$ are mutually independent. Argue that we can sequentially complement any set in the list $A_1, ..., A_n$ and the resulting events are mutually independent.
19) Suppose A and B are events that satisfy $A \subseteq B$. Suppose $P[B] = 0$. Prove $P[A] = 0$.
20) Suppose A and B are events. Suppose $P[A] = 1$. Prove that A and B are independent.
21) Suppose A and B are events that satisfy $P[A] = P[B] = 1$. Show that $P[A \cap B] = 1$.
22) Suppose $\{A_i\}_{i=1}^{\infty}$ are events that satisfy $P[A_i] = 0$ for all $i \in \{1, 2, 3, \ldots\}$. Prove that $P[\cup_{i=1}^{\infty} A_i] = 0$.
23) Suppose events A and B satisfy $A \subseteq B$. Assuming $P[B] > 0$, compute $P[A|B]$. Assuming $P[A] > 0$, compute $P[B|A]$.
24) Suppose A and B are *disjoint* events that satisfy $P[A] > 0$ and $P[B] > 0$. Are A and B independent?
25) (Pairwise independent but not mutually independent) Suppose we flip a fair coin twice. We have sample space $\Omega = \{(H,H), (H,T), (T,H), (T,T)\}$ with all four outcomes equally likely. Define

$$A = \{\text{first flip is heads}\} = \{(H,T), (H,H)\}$$
$$B = \{\text{second flip is heads}\} = \{(H,H), (T,H)\}$$
$$C = \{\text{first and second flips are different}\} = \{(H,T), (T,H)\}$$

   Show that A, B, C are pairwise independent but not mutually independent.
26) Let $\{A_i\}_{i=1}^{\infty}$ be mutually independent events with $P[A_i] = \frac{1}{i+1}$ for $i \in \mathbb{N}$. Compute $P[\cup_{i=1}^{n} A_i]$ and $P[\cup_{i=1}^{\infty} A_i]$.
27) Let $\{A_i\}_{i=1}^{\infty}$ be mutually independent events with $P[A_i] = \frac{1}{(i+1)^2}$ for $i \in \mathbb{N}$. Compute $P[\cup_{i=1}^{n} A_i]$ and $P[\cup_{i=1}^{\infty} A_i]$.
28) (Arbitrarily long run lengths of heads) We independently flip a fair coin an infinite number of times. Let H_i be the event that flip i is heads. Thus, $\{H_i\}_{i=1}^{\infty}$ are mutually independent and $P[H_i] = 1/2$ for all $i \in \{1, 2, 3, ...\}$. For $m \in \{1, 2, 3, ...\}$, let A_m be the event that we have a run length of at least m consecutive heads somewhere in our infinite sequence of flips. Let $A = \cap_{m=1}^{\infty} A_m$ be the event that we have arbitrarily long run lengths of heads. Show that $P[A] = 1$.
29) Fix probability space $(\Omega, \mathcal{F}, P)$. Suppose that $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$. Suppose $\{A_i\}_{i=1}^{\infty}$ is a sequence of mutually independent events such that $P[A_i] = 1/2$ for all $i \in \{1, 2, 3, ...\}$.
   a) Show that $P[\{\omega\}] = 0$ for all $\omega \in \Omega$. [You can use the result of items 12 and 18 in this list of questions.]
   b) Show that Ω is an uncountably infinite set.
30) Fix probability space $(\Omega, \mathcal{F}, P)$. For each $\omega \in \Omega$ define

$$p(\omega) = \inf_{A \in \mathcal{F}: \omega \in A} \{P[A]\}$$

   a) In the special case when $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$, show that $P[\{\omega\}] = p(\omega)$ for all $\omega \in \Omega$.
   b) In the general case, show that for all $\omega \in \Omega$ there is an $A \in \mathcal{F}$ such that $\omega \in A$ and $P[A] = p(\omega)$.

L. Exercises on conditioning and law of total probability

1) We play team A, B, C, or D equally likely. Let A, B, C, D denote the events that we play team A, B, C, D, respectively. Let W denote the event that we win. Our conditional probability of winning depends on which team we play as follows:

$$P[W|A] = 0.9, \quad P[W|B] = 0.5, \quad P[W|C] = 0.4, \quad P[W|D] = 0$$

   a) Compute $P[W]$
   b) Compute $P[A|W]$
   c) Compute $P[D|W]$
   d) Compute $P[W|A \cup B]$

2) Let $\{B_i\}_{i=1}^{\infty}$ be events that partition the sample space. Let C be an event that satisfies $P[C] > 0$. Let A be an event. Prove the law of conditioned total probability:

$$P[A|C] = \sum_{i=1}^{\infty} P[A|B_i \cap C]P[B_i|C]$$

   where the summation only sums over indices i that satisfy $P[B_i \cap C] > 0$.

3) We have a countably infinite number of biased coins, where coin $i \in \{1, 2, 3, ...\}$ has $P[Heads] = 1/i$. We first have a computer generate a random positive integer X with $P[X = i] = (1/2)^i$ for $i \in \{1, 2, 3, ...\}$. Given that $X = i$, we flip coin i.
   a) Compute $P[Heads]$
   b) Compute $P[X = 1|Heads]$
   c) Compute $P[X = 2|Heads]$
   d) Compute $P[X \in \{1, 2\}|Heads]$
   e) Compute $P[Heads|X \in \{1, 2\}]$

4) Suppose that at breakfast time we either eat cereal, oatmeal, or rice. Suppose:
   - Having milk and having a side dish are independent.
   - $P[\{\text{have milk}\}] = 3/4$, $P[\{\text{have side dish}\}] = 1/2$.
   - Given we only have milk (no side dish): We eat cereal with prob $3/4$ and oatmeal with prob $1/4$.
   - Given we have milk and side dish: We eat cereal with prob $1/2$, rice with prob $1/4$, oatmeal with prob $1/4$.
   - Given we only have side dish (no milk): We eat oatmeal with prob $1/2$ and rice with prob $1/2$.
   - Given we have no milk and no side dish: We eat oatmeal with prob 1.

   Let C, O, R be events of eating [C]ereal, [O]atmeal, [R]ice, respectively. Let M, S be events of having [M]ilk and having a [S]ide dish, respectively.
   a) What is the conditional probability we eat oatmeal, given we only have a side dish (no milk)?
   b) Compute the probability that we eat cereal.
   c) Compute the probability that we eat rice. Then, compute the probability we eat oatmeal.
   d) Compute the conditional probability we have a side dish, given we eat cereal.
   e) Compute the probability we have milk and eat cereal. Explain this result in relation to part (b).

5) (Cascaded events) Let A, B, C be events with $P[A \cap B] > 0$. Prove $P[A \cap B \cap C] = P[A]P[B|A]P[C|A \cap B]$. Guess a generalized formula for $P[A_1 \cap ... \cap A_n]$ when $A_1, ..., A_n$ are events with $P[\cap_{i=1}^{n-1} A_i] > 0$.

II. Combinatorics

Let Ω be a finite sample space. Define $\mathcal{F} = Pow(\Omega)$. For each $A \subseteq \Omega$, define $|A|$ as the number of elements in set A. Clearly $|A| \leq |\Omega|$ for all $A \subseteq \Omega$. Consider the probability measure $P : Pow(\Omega) \to [0, 1]$ where *all outcomes are equally likely*, so that

$$P[\{\omega\}] = \theta \quad \forall \omega \in \Omega$$

for some value $\theta \in [0, 1]$. Since $\Omega = \cup_{\omega \in \Omega}\{\omega\}$ is a finite disjoint union we have

$$1 = P[\Omega] = \sum_{\omega \in \Omega} P[\{\omega\}] = \theta \cdot |\Omega|$$

and so $\theta = 1/|\Omega|$. Thus, for any set $A \subseteq \Omega$ we have

$$P[A] = P[\cup_{\omega \in A}\{\omega\}] = \sum_{\omega \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|}$$

In conclusion, in a finite sample space where all outcomes are equally likely, the probability of any event is just the size of the event divided by the size of the sample space. This is one motivation for the field of *combinatorics* which is the theory for counting finite sets.

A. Multiplication principle

Fix n as a positive integer. Let $k_1, k_2, \ldots, k_n$ also be positive integers. Imagine a procedure that involves n steps, as follows:

- Step 1: Make one of k_1 choices.
- Step 2: For each particular choice of step 1, there are k_2 new choices to make at step 2.
- Step 3: For each particular choice for step 1 and step 2, there are k_3 new choices to make at step 3.
- $\cdots$
- Step n: For each particular choice for steps $1, 2, ..., n-1$, there are k_n new choices to make at step n.

Note that the precise choices at each new step i may depend on the previous steps, but the number of options at step i is always k_i. All possible choices can be represented by a tree with n levels. The total number of possible combinations of choices is given by the number of leaves at the bottom level of the tree, which is

$$\text{Number of choice combinations} = k_1 \cdot k_2 \cdot \ldots \cdot k_n = \prod_{i=1}^{n} k_i$$

For example, if $k_i = 2$ for all $i \in \{1, \ldots, n\}$ then the number of possible choice combinations is 2^n.

B. Permutations, sampling with and without replacement

Fix n as a positive integer. The multiplication principle (and its consequences) can be used to derive

- The fact that there are 2^n binary vectors of size n.
- The $n!$ formula for the number of permutations of n objects.
- The $n(n-1)\cdots(n-(k-1))$ formula for choosing k objects from a bin of n, with regard to ordering.
- The binomial coefficient $\binom{n}{k}$.
- The multinomial coefficient $\frac{n!}{k_1!k_2!\cdots k_m!}$.
- The stars and bars formula.

These are useful for solving problems about dealing cards, choosing balls in bins, sampling with and without replacement, counting choices with and without regard to order (see [1]). It is useful to create and solve your own combinatorics problems.

III. Random variables

Fix a probability space $(\Omega, \mathcal{F}, P)$. A *random variable* is a function $X : \Omega \to \mathbb{R}$ that satisfies the following *measurability property*:[9]

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F} \quad \forall x \in \mathbb{R} \tag{11}$$

The measurability property (11) automatically holds in the special case when $\mathcal{F}$ is the set of all subsets of Ω. Therefore, if $\mathcal{F} = Pow(\Omega)$ then every function $X : \Omega \to \mathbb{R}$ is a random variable.

A. A function $X : \Omega \to \mathbb{R}$ that is not a random variable

It is easy to construct an example function $X : \Omega \to \mathbb{R}$ that is *not* a random variable if we use a finite sample space Ω and a sigma algebra other than $Pow(\Omega)$. For example, consider

$$\Omega = \{blue, red, green, black\}$$
$$\mathcal{F} = \{\phi, \Omega, \{blue, red, green\}, \{black\}\}$$

It can be shown that $\mathcal{F}$ is a sigma algebra on Ω, but $\mathcal{F}$ does not include all subsets of Ω. Define a function $X : \Omega \to \mathbb{R}$ by

$$X(blue) = 0$$
$$X(red) = 0$$
$$X(green) = 1.3$$
$$X(black) = 7$$

Then

$$\{X \leq 1\} = \{blue, red\} \notin \mathcal{F}$$

and so the function $X : \Omega \to \mathbb{R}$ is not a random variable. Of course, if we changed the sigma algebra to $\mathcal{F} = Pow(\Omega)$ then the same function $X : \Omega \to \mathbb{R}$ defined above would be a random variable.

[9] In Appendix D it is shown that $X : \Omega \to \mathbb{R}$ satisfies (11) if and only if $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ for all sets $B \in \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ is the standard Borel sigma algebra on $\mathbb{R}$ defined in Appendix D.

B. Working with random variables

A probability experiment has one probability space $(\Omega, \mathcal{F}, P)$. The probability experiment may have multiple stages and may involve many random variables $Y, Z, \{X_i\}_{i=1}^{\infty}$. These random variables are housed by the space $(\Omega, \mathcal{F}, P)$, which means

$$Y : \Omega \to \mathbb{R}$$
$$Z : \Omega \to \mathbb{R}$$
$$X_i : \Omega \to \mathbb{R} \quad \forall i \in \{1, 2, 3, \ldots\}$$

and their measurability property is with respect to the same sigma algebra $\mathcal{F}$. The particular outcome $\omega \in \Omega$ determines the particular numerical values of all random variables on the space:

$$Y(\omega), Z(\omega), X_1(\omega), X_2(\omega), X_3(\omega), X_4(\omega), \ldots$$

Unless otherwise stated, when multiple random variables are considered together, it is implicitly assumed they are *on the same probability space*. This means there is a single probability space $(\Omega, \mathcal{F}, P)$ that houses them. The probability that "Y is less than or equal to 4.5 and Z is less than or equal to -2" only makes sense if Y and Z are determined by the same probability experiment, meaning they are both on the same probability space.

C. Cumulative Distribution Function (CDF)

Fix a probability space $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$ be a random variable. It is convenient to use the compressed notation $\{X \leq x\}$ to represent a subset of Ω

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\}$$

where x is a given real number. Since X is a random variable, the measurability property (11) implies that for each $x \in \mathbb{R}$, the set $\{X \leq x\}$ is a valid event, hence it has a probability. The *cumulative distribution function (CDF)* F_X is defined by collecting these probabilities into a function. Indeed, define $F_X : \mathbb{R} \to [0, 1]$ by

$$F_X(x) = P[X \leq x] \quad \forall x \in \mathbb{R}$$

where $P[X \leq x]$ is compressed notation for $P[\{X \leq x\}]$. It can be proven that for $a, b \in \mathbb{R}$ with $a < b$:

$$P[X \in (a, b]] = P[X \leq b] - P[X \leq a] = F_X(b) - F_X(a)$$
$$P[X = b] = \lim_{n\to\infty} P[X \in (b - 1/n, b]] = F_X(b) - \lim_{n\to\infty} F_X(b - 1/n)$$

Every CDF has four fundamental properties: Nondecreasing; right-continuous; $\lim_{x\to\infty} F_X(x) = 1$; $\lim_{x\to-\infty} F_X(x) = 0$ (see items 26, 27, 28, 29 in Section III-H). It can be shown that any function $F : \mathbb{R} \to \mathbb{R}$ with these four properties is the CDF of some random variable on some probability space.

D. Joint CDF

Fix a probability space $(\Omega, \mathcal{F}, P)$. If Y and Z are random variables, the measurability property (11) ensures

$$\{Y \leq 4.5\} \in \mathcal{F}, \quad \{Z \leq -2\} \in \mathcal{F}$$

Since the intersection of two sets in the sigma algebra $\mathcal{F}$ is again a set in $\mathcal{F}$, we conclude

$$\{Y \leq 4.5\} \cap \{Z \leq -2\} \in \mathcal{F}$$

Since $\{Y \leq 4.5\} \cap \{Z \leq -2\}$ is an event, it has a probability. For simplicity of notation we often remove the braces and replace the intersection symbol with a comma (which means "and"):

$$P[Y \leq 4.5, Z \leq -2] = P[\{Y \leq 4.5\} \cap \{Z \leq -2\}] = P[\{\omega \in \Omega : Y(\omega) \leq 4.5 \text{ and } Z(\omega) \leq -2\}]$$

Fix n as a positive integer. Let $X_1, \ldots, X_n$ be random variables. Let $x_1, \ldots, x_n$ be real numbers. For each $i \in \{1, \ldots, n\}$, since X_i is a random variable, the measurability property (11) ensures $\{X_i \leq x_i\}$ is an event. The intersection of a finite number of events is again an event, so

$$\cap_{i=1}^{n} \{X_i \leq x_i\} \in \mathcal{F}$$

So $\cap_{i=1}^{n} \{X_i \leq x_i\}$ must have a probability. The *joint cumulative distribution function (joint CDF)* $F_{X_1,\ldots,X_n}$ is defined by collecting these probabilities into a function. Define $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0, 1]$ by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P[\cap_{i=1}^{n} \{X_i \leq x_i\}] \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

It can be proven that the joint CDF $F_{X_1,\ldots,X_n}$ determines

$$P[(X_1, \ldots, X_n) \in A] \quad \forall A \in \mathcal{B}(\mathbb{R}^n)$$

where $\mathcal{B}(\mathbb{R}^n)$ is defined in Appendix D. In particular, if the joint CDF is known, then for any Borel measurable set $A \subseteq \mathbb{R}^n$, the probability that $(X_1, \ldots, X_n)$ lies in the set A can (in principle) be computed, even if $(\Omega, \mathcal{F}, P)$ is unknown.

E. Independence of random variables

Fix $(\Omega, \mathcal{F}, P)$. Fix n as a positive integer. Random variables $X_1, X_2, ..., X_n$ are defined as *mutually independent* if

$$P[\cap_{i=1}^n \{X_i \leq x_i\}] = \prod_{i=1}^n P[X_i \leq x_i] \quad \forall (x_1, \ldots, x_n) \in \mathbb{R}^n$$

This is equivalent to the joint CDF factoring into a product of marginal CDFs:

$$F_{X_1,...X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n) \quad \forall (x_1, ..., x_n) \in \mathbb{R}^n$$

Random variables in an infinite sequence $\{X_i\}_{i=1}^\infty$ are defined to be *mutually independent* if $X_1, \ldots, X_n$ are mutually independent for each positive integer n.

It can be shown that if $X_1, \ldots, X_n$ are mutually independent random variables then:

- For any sets $A_1, \ldots, A_n$ that are Borel measurable subsets of $\mathbb{R}$ (defined in Appendix D) and all $I \subseteq \{1, \ldots, n\}$ we have
$$P[\cap_{i\in I}\{X_i \in A_i\}] = \prod_{i\in I} P[X_i \in A_i]$$
- For any Borel measurable functions $g_i : \mathbb{R} \to \mathbb{R}$ (defined in Appendix D), if $\mathbb{E}[g_i(X_i)]$ is finite for $i \in \{1, \ldots, n\}$ then
$$\mathbb{E}\left[\prod_{i=1}^n g_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[g_i(X_i)]$$
In particular $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$ whenever X_1 and X_2 are independent random variables with finite expectations.

F. Operations on measurable functions

Basic operations on random variables preserve the measurability property (11). Specifically, fix $(\Omega, \mathcal{F}, P)$ as a probability space. The sum or product of a finite number of random variables on this space produces another random variable on this space.[10] A linear combination of random variables on this space is another random variable on this space. For example, if $\{X_i\}_{i=1}^\infty$ is a sequence of random variables, then for each positive integer n we can define $M_n : \Omega \to \mathbb{R}$ by

$$M_n = \frac{1}{n}\sum_{i=1}^n X_i$$

meaning that

$$M_n(\omega) = \frac{1}{n}\sum_{i=1}^n X_i(\omega) \quad \forall \omega \in \Omega$$

and it can be shown that $\{\omega \in \Omega : M_n(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. So M_n satisfies (11) and thus M_n is a random variable.

More generally, it is known that continuous functions preserve measurability (see Appendix D). Indeed, fix n as a positive integer and let $h : \mathbb{R}^n \to \mathbb{R}$ be any continuous function. Let $X_1, \ldots, X_n$ be random variables. Define $Y = h(X_1, \ldots, X_n)$, so

$$Y(\omega) = h(X_1(\omega), \ldots, X_n(\omega)) \quad \forall \omega \in \Omega$$

Appendix D shows that Y satisfies the measurability property (11) and hence Y is a random variable. For example:

- If $h : \mathbb{R} \to \mathbb{R}$ is the continuous function $h(x) = e^x$ for all $x \in \mathbb{R}$, then $h(X_1) = e^{X_1}$. So e^{X_1} is a random variable.
- If $h : \mathbb{R}^2 \to \mathbb{R}$ is the continuous function $h(x_1, x_2) = x_1x_2$ for all $(x_1, x_2) \in \mathbb{R}^2$, then $h(X_1, X_2) = X_1X_2$. So the product of two random variables is a random variable.
- If $h : \mathbb{R}^3 \to \mathbb{R}$ is the continuous function $h(x_1, x_2, x_3) = 7x_1\cos(x_2 + x_3)$ then $h(X_1, X_2, X_3) = 7X_1\cos(X_2 + X_3)$. So $7X_1\cos(X_2 + X_3)$ is a random variable.

G. Summarizing exercises 1 (multiple choice)

1) Let $\Omega = \{red, green, blue\}$. Define $X : \Omega \to \mathbb{R}$ by $X(red) = 0$, $X(green) = 1.2$, $X(blue) = -4$. Then
   a) $\{X \leq 1\} = \{red, blue\}$
   b) $\{X^2 = 16\} = \{blue\}$
   c) $\{2X > 0\} = \{green\}$
   d) All of the above
2) Let X be a random variable. What is $\{X \leq 8.7\}$?
   a) $\{\omega \in \Omega : X(\omega) \leq 8.7\}$
   b) A set of outcomes
   c) An event
   d) All of the above

[10] Fix random variables X and Y. If $t \in \mathbb{R}$ then $\{X + Y < t\} = \cup_{\{(a,b)\in\mathbb{Q}^2 : a+b<t\}}\{X \leq a\} \cap \{Y \leq b\}$ is a countable union of events, hence an event.

3) Let X be a random variable. Which is always true?
   a) $\{X \leq 5\} = \{X \leq 5.5\}$
   b) $\{X \leq 5\} \supseteq \{X \leq 5.5\}$
   c) $\{X \leq 5\} \subseteq \{X \leq 5.5\}$
   d) X must take integer values
4) In a probability triplet $(\Omega, \mathcal{F}, P)$, if $\mathcal{F}$ is the set of all subsets of Ω then
   a) all functions $X : \Omega \to \mathbb{R}$ are random variables.
   b) we cannot define random variables
   c) all events have one outcome
   d) some events are not in $\mathcal{F}$
5) If X is a random variable then $\{X \leq 5.6\}$ is
   a) an event
   b) an outcome
   c) a real number
   d) a function of ω
6) If $\{X \leq 5.6\}$ is an event, then $\{X > 5.6\}$ is an event because
   a) all subsets of Ω are events
   b) all events are subsets of Ω
   c) the number 5.6 is special
   d) the complement of an event is also an event
7) If $\{X \leq 5.6\}$ and $\{Y \leq 8.3\}$ are both events then $\{X \leq 5.6\} \cap \{Y \leq 8.3\}$ is an event because
   a) the intersection of two events is also an event
   b) the union of events is also an event
   c) all subsets of Ω are events
   d) $8.3 - 5.6 = 2.7$
8) If A is an event and $P[A] = 0$ then
   a) A must be the empty set
   b) A is not necessarily the empty set
   c) We can apply the central limit theorem to the set A
   d) We can take expectations of the set A
9) If X and Y are random variables on some probability space, then $X + Y$ is a random variable because
   a) a sum of two random variables is again a random variable
   b) a linear combination of two random variables is again a random variable
   c) a continuous function of two random variables is again a random variable
   d) all of the above
10) If X and Y are random variables that have the same CDF function, then they must be independent.
   a) True
   b) False
   c) Unknown
   d) Electoral college
11) If Y and Z are random variables then $\{Y + e^Z \leq Z\} \in \mathcal{F}$ because
   a) all subsets of Ω are events
   b) $Y + e^Z - Z$ is a random variable and hence satisfies the measurability property (11)
   c) Y is a function of Z
   d) Y and Z are functions of ω
12) Random variables X and Y are independent if and only if
   a) they are uncorrelated
   b) they are jointly Gaussian
   c) for all $x, y \in \mathbb{R}$ the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent
   d) they converge to a constant
13) Let $\{X_i\}_{i=1}^{\infty}$ be independent and identically distributed (i.i.d.) random variables on some probability space. If the probability experiment produces a particular outcome $\omega^* \in \Omega$ then
   a) the first random variable takes the value $X_1(\omega^*)$ but the values of $\{X_2, X_3, X_4, \ldots\}$ remain unknown
   b) the random variables take the values $\{X_1(\omega^*), X_2(\omega^*), X_3(\omega^*), X_4(\omega^*), \ldots\}$
   c) the single-outcome set $\{\omega^*\}$ must be an event
   d) if $\{\omega^*\} \in \mathcal{F}$ then $P[\{\omega^*\}] > 0$

H. Summarizing exercises 2

Fix a probability space $(\Omega, \mathcal{F}, P)$. Let X, Y, Z be random variables on this space.

1) Does X necessarily have a CDF?
2) Do X and Y necessarily have a joint CDF $F_{X,Y}$?
3) Do X, Y, Z necessarily have a joint CDF $F_{X,Y,Z}$?
4) Is $4X - Y + 2.1Z$ a random variable?
5) Is $X + 2Y + 8Z$ a random variable?
6) Is $X + e^Y$ a random variable?
7) Is $\frac{YZ}{X^2+1}$ a random variable?
8) Is $\{X \leq 3\} \in \mathcal{F}$?
9) Is $\{X > 3\} \in \mathcal{F}$?
10) Is $\{X \in (2, 5]\} \in \mathcal{F}$?
11) Is $\{\frac{1}{3}(X + Y + Z) \leq 8\} \in \mathcal{F}$?
12) Is $\{e^X + \cos(Y + Z) > 2X\} \in \mathcal{F}$?
13) Is $\{\cos(X + Y) \leq 0\} \cup \{e^{4(X+Z)} > 3\} \in \mathcal{F}$?
14) Is $\cap_{i=1}^{\infty}\{\cos(iX + Y) \leq \frac{Z}{i}\} \in \mathcal{F}$?
15) Fix $A \in \mathcal{F}$. Is the following *indicator function* 1_A a random variable? Define $1_A : \Omega \to \mathbb{R}$ by
$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{else} \end{cases}.$$
16) For $A \in \mathcal{F}$, is $X1_A + Y$ a random variable?
17) For $A \in \mathcal{F}$ and $B \in \mathcal{F}$, is $X1_A + e^Y 1_B + XY$ a random variable?
18) Fix $C \subseteq \Omega$ such that $C \notin \mathcal{F}$. Is the indicator function 1_C a random variable?
19) Argue that for any $a, b \in \mathbb{R}$ we have $\{X \in (a, b]\} \in \mathcal{F}$.
20) Argue that $\{X = 2.3\} \in \mathcal{F}$ by observing $\{X = 2.3\} = \cap_{i=1}^{\infty}\{X \in (2.3 - 1/i, 2.3]\}$.
21) Is $\{X < 2.3\} \in \mathcal{F}$?
22) Is $\{X \in \{2.3, 4.1, 9\}\} \in \mathcal{F}$?
23) Is $\{XY \leq 8\} \cup \{XY = 9.1\} \in \mathcal{F}$?
24) Is $\{XY \leq 7Z\} \cap \{YZ > 4\}^c \in \mathcal{F}$?
25) Is $\cup_{i=1}^{\infty}\{X = 1/i\} \in \mathcal{F}$?
26) Suppose X is a random variable. Show the CDF is nondecreasing: $x \leq y \implies F_X(x) \leq F_X(y)$.
27) Suppose $X : \Omega \to \mathbb{R}$ is a random variable. Use $\{X \leq n\} \nearrow \Omega$ to prove $\lim_{n\to\infty} F_X(n) = 1$. Does this also mean that $\lim_{x\to\infty} F_X(x) = 1$? In other words, does the answer change if we take the limit over integers or over real numbers?
28) Suppose $X : \Omega \to \mathbb{R}$ is a random variable. Argue that $\{X \leq -n\} \searrow \phi$. Use this to conclude $\lim_{n\to\infty} F_X(-n) = 0$.
29) Suppose X is a random variable. Fix $x \in \mathbb{R}$. Use $\{X \leq x + 1/n\} \searrow \{X \leq x\}$ to show the CDF is right-continuous: $\lim_{n\to\infty} F_X(x + 1/n) = F_X(x)$ for all $x \in \mathbb{R}$.
30) Suppose X is a random variable that has a continuous CDF. Show that $P[X = x] = 0$ for all $x \in \mathbb{R}$. Hint: Fix $x \in \mathbb{R}$ and observe $\{X = x\} \subseteq \{X \in (x - \delta, x]\}$ for all $\delta > 0$.
31) Create a function $h : \mathbb{R} \to \mathbb{R}$ such that $\lim_{n\to\infty} h(n) = 1$ (the limit is taken over the positive integers $n \in \{1, 2, 3, ...\}$), but $\lim_{x\to\infty} h(x)$ does not exist (the limit is taken over real numbers x). Is your h nondecreasing?
32) Fix $(\Omega, \mathcal{F}, P)$ and let X, Y be random variables. From first principles, prove $\{XY = 0\} \in \mathcal{F}$.
33) Fix $(\Omega, \mathcal{F}, P)$ and let X be a random variable. Define $Y = aX + b$ for some $a, b \in \mathbb{R}$. From first principles, prove Y is a random variable. That is, prove $\{Y \leq t\} \in \mathcal{F}$ for all $t \in \mathbb{R}$. You can treat the cases $a > 0, a = 0, a < 0$ separately. [Note that $\{b \leq t\}$ is either Ω or ϕ depending on if the constants b, t satisfy $b \leq t$ or $b > t$.]
34) Let X be a random variable. Show the CDF $F_X(x)$ has at most countably many points of discontinuity. [Hint: One way is to use the footnote of Section I-J and the fact that the CDF is continuous at a point $x \in \mathbb{R}$ if and only if $P[X = x] = 0$.]
35) Suppose $(\Omega, \mathcal{F}, P)$ is a probability space such that all single-outcome subsets of the sample space are events, that is, $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$. Suppose there is a random variable $X : \Omega \to \mathbb{R}$ that has a continuous CDF.
    a) Show that $P[\{\omega\}] = 0$ for all $\omega \in \Omega$.
    b) Show that Ω is an uncountably infinite set.
36) Suppose random variables X and Y satisfy $P[X = Y] = 1$. Show X and Y have the same CDF.
37) Suppose X and Y are random variables and $P[X = 4.2] = 1$. Show X and Y are independent.
38) (Throwing away an event of probability 0 has no significant impact) Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X : \Omega \to \mathbb{R}$ be a random variable. Let $V \in \mathcal{F}$ be a nonempty event such that $P[V] = 0$. Define $\tilde{\Omega} = \Omega \cap V^c$, $\tilde{\mathcal{F}} = \{A \in \mathcal{F} : A \subseteq \tilde{\Omega}\}$, $\tilde{P} : \tilde{\Omega} \to [0, 1]$ by $\tilde{P}[A] = P[A]$ for all $A \in \tilde{\mathcal{F}}$.
    a) Show that $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ is a probability space.
    b) Define $\tilde{X} : \tilde{\Omega} \to \mathbb{R}$ by $\tilde{X}(\omega) = X(\omega)$ for all $\omega \in \tilde{\Omega}$. Show that $\tilde{P}[\tilde{X} \in A] = P[X \in A]$ for all $A \in \mathcal{B}(\mathbb{R})$. [If you prefer, you can simply show $\tilde{P}[\tilde{X} \leq x] = P[X \leq x]$ for all $x \in \mathbb{R}$.] So X and $\tilde{X}$ have the same distribution.

I. Probability desnsity function (PDF)

If X is a random variable then a *probability density function (PDF)* for X is a nonnegative function $f_X(x)$ defined for all $x \in \mathbb{R}$ that can be integrated and that satisfies

$$P[X \leq x] = \int_{-\infty}^{x} f_X(t)dt \quad \forall x \in \mathbb{R} \tag{12}$$

Taking $x \to \infty$ in the above equality proves that every PDF satisfies $1 = \int_{-\infty}^{\infty} f_X(t)dt$. It turns out that every nonnegative function $f : \mathbb{R} \to \mathbb{R}$ that can be integrated and that satisfies $\int_{-\infty}^{\infty} f(t)dt = 1$ is a PDF for some random variable. It can be shown that if a random variable has a PDF then: (i) Its CDF must be continuous everywhere; (ii) $P[X \in A] = \int_{x \in A} f_X(x)dx$ for all $A \in \mathcal{B}(\mathbb{R})$ (defined in Appendix D).

The fundamental theorem of calculus can be used with (12): If a PDF function f_X is continuous at a particular point $x \in \mathbb{R}$ then the CDF F_X is differentiable at x and $F_X'(x) = f_X(x)$. One way to find a PDF for a random variable is to first find the CDF and then take a derivative at points where the CDF is differentiable. This can often be done at all but a finite or countably infinite number of points and typically reveals a nonnegative function that can be integrated back to get the CDF.

Every random variable has a unique CDF. Not all random variables have PDFs. For example, a random variable with a discontinuous CDF cannot have a PDF. When a PDF exists then it is not unique. That is because we can change the PDF value at a (single) particular point $x \in \mathbb{R}$ to any nonnegative number without affecting the value of any integral, so the modified function is also a PDF for the same random variable X.

Famous PDFs are:

- $X \sim Exp(\lambda)$. Fix $\lambda > 0$. A random variable X is *exponentially distributed* with parameter λ if it has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

  If desired, we can modify the above PDF to make $f_X(0) = 0$, rather than $f_X(0) = \lambda$, and the CDF is unchanged. So this is still a valid PDF for a random variable $X \sim Exp(\lambda)$.
- $X \sim \text{Unif}[a, b]$. Fix $a, b \in \mathbb{R}$ with $a < b$. A random variable X is *uniformly distributed* over $[a, b]$ if it has PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

  If desired, we can modify the above PDF to make $f_X(a) = f_X(b) = 0$ and the CDF is unchanged.
- $X \sim N(\mu, \sigma^2)$. Fix μ, σ^2 such that $\mu \in \mathbb{R}$ and $0 < \sigma^2 < \infty$. A random variable X has a *Gaussian distribution* with parameters μ, σ^2, also called a *normal distribution*, if it has PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}$$

Caveat: The PDF value $f_X(x)$ at a given point $x \in \mathbb{R}$ is a nonnegative real number. It is not a probability. It is easy to size the parameters of the exponential, normal, and uniform PDFs to find real numbers x for which $f_X(x) > 1$. For example, if $X \sim \text{Unif}[0, 1/10]$ then the uniform PDF given above yields $f_X(x) = 10$ for $0 \leq x \leq 1/10$.

J. Expectation

Expectation and variance are reviewed in Appendix A, including the *law of total expectation* and *linearity of expectation*:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}[X|B_i] P[B_i]$$
$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

where X is a random variable with finite expectation; $\{B_i\}_{i=1}^{\infty}$ is a collection of events that partition the sample space; $X_1, ..., X_n$ are (possibly dependent) random variables with finite expectations; $a_1, ..., a_n$ are real numbers.

K. What kind of thing is this?

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X be a random variable that satisfies $-1000 \leq X \leq 1000$ surely.

1) What does $P[X \leq 5.2]$ represent?
   a) The probability that X is larger than 4
   b) The probability that X is less than or equal to 5.2
   c) The probability that X is an integer
   d) The expectation of X
2) What does $P[\{X > 4\} \cup \{X < 2\}]$ represent?
   a) The probability that X is greater than 4

b) The probability that X is less than $\sqrt{2}$
c) The probability that X is either greater than 4 or less than 2
d) The probability that X is a prime number

3) What does $P[X]$ represent?
a) The probability that X
b) The P function turns X into an event and then computes its probability
c) The expectation of X
d) Nonsense

4) We take probabilities of
a) events
b) random variables
c) expectations
d) sigma algebras

5) We take expectations of
a) outcomes
b) events
c) axioms
d) random variables

6) What is $\mathbb{E}[X]$?
a) a random variable
b) a probability
c) a function of x
d) a real number

7) What is $4 + \cos(X)$?
a) a probability
b) a real number
c) a random variable
d) nonsense

8) What is $\mathbb{E}[4 + \cos(X)|X \geq w]$?
a) a function of w
b) a probability
c) a random variable
d) a sigma algebra

9) What is $\mathbb{E}[8\cos(X)]$?
a) a real number
b) an axiom
c) a function of x
d) a sigma algebra

10) What is $P[X \leq 5]$?
a) a real number
b) a random variable
c) a function of x
d) nonsense

11) What is $P[X \leq x]$?
a) the variance of X
b) a random variable
c) a function of x
d) a real number

12) What is $P[\cos(X) \leq t]$?
a) a random variable
b) a function of t
c) a sigma algebra
d) nonsense

13) What is $P[\cos(X)]$?
a) a real number
b) a random variable
c) a function of x
d) nonsense

L. Indicator random variable

Let A be an event. Define the binary-valued random variable 1_A by

$$1_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{else} \end{cases}$$

In particular, $1_A(\omega) = 1$ if and only if $\omega \in A$. The random variable 1_A is called an *indicator random variable* or *indicator function*.

M. Bernoulli trials

Fix $p \in [0,1]$. Let $\{A_i\}_{i=1}^{\infty}$ be sequence of mutually independent events with

$$P[A_i] = p \quad \forall i \in \{1,2,3,...\}$$

This sequence shall be called *i.i.d. Bernoulli trials*. Intuitively we can imagine a sequence of independent "trials" or "mini-experiments," where each trial i can either be *successful* or *not successful*. The event A_i is the event that trial i is successful. That is, trial i is called a *success* if and only if event A_i is true.

- Bernoulli random variable ($X \sim Bern(p)$): Define X as an indicator function that is 1 if and only if the first Bernoulli trial is successful:

  $$X = 1_{A_1}$$

  Then

  $$\begin{aligned} &X \in \{0,1\} \\ &P[X=1] = p, P[X=0] = 1-p \\ &\mathbb{E}[X] = p \\ &Var(X) = p(1-p) \end{aligned}$$

- Binomial random variable ($X \sim Binom(n,p)$): Fix n as a positive integer. Let X be the number of successes within the first n trials:

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

  Then

  $$\begin{aligned} &X \in \{0,1,2,...,n\} \\ &P[X=k] = \binom{n}{k} p^k (1-p)^{n-k} \quad \forall k \in \{0,1,...,n\} \\ &\mathbb{E}[X] = np \\ &Var(X) = np(1-p) \end{aligned}$$

- Geometric random variable ($X \sim Geom(p)$): Assume $p > 0$. Define X as the trial index of the first success. Then[11]

  $$\begin{aligned} &X \in \{1,2,3,...\} \\ &P[X=k] = p(1-p)^{k-1} \quad \forall k \in \{1,2,3,...\} \\ &\mathbb{E}[X] = 1/p \\ &Var(X) = \frac{1-p}{p^2} \end{aligned}$$

[11]Using the concept of a *random element* in Appendix D, one might view X as an *extended random variable* $X : \Omega \to \mathbb{R} \cup \{\infty\}$ if there are some outcomes $\omega \in \Omega$ for which $X(\omega) = \infty$ (meaning there are *never any successes*). These outcomes, taken together, have collective probability $P[X=\infty] = 1 - \sum_{i=1}^{\infty} P[X=i] = 1 - \sum_{i=1}^{\infty} p(1-p)^{i-1} = 0$. Alternatively, one could modify the sample space Ω by removing all outcomes in the zero-probability event $\{X = \infty\}$. This does not change any probability or expectation because $P[X=\infty] = 0$ (see also item 38 in Section III-H).

N. Famous random variables

Fix a probability space $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$ be a random variable. Define $S_X \subseteq \mathbb{R}$ as the set of all possible values the random variable can take, also called the *image* of the random variable:

$$S_X = \{X(\omega) \in \mathbb{R} : \omega \in \Omega\}$$

- Discrete type: A random variable X is of the *discrete type* if S_X is a finite or countably infinite set.[12] The CDF of a discrete type random variable is completely characterized by its *probability mass function* (PMF) $P[X = x]$ for $x \in S_X$. Some famous discrete type random variables are Bernoulli, geometric, binomial, Poisson, discrete uniform. Note that the sample space Ω does not need to be a finite or countably infinite set.

- Continuous type: A random variable X is of the *continuous type* if its CDF $F_X(x)$ is a continuous function for all $x \in \mathbb{R}$, and if there is a probability density function (PDF). It can be shown that if X has a continuous CDF then it cannot be of the discrete type. Some famous continuous type random variables are continuous uniform, exponential, Gaussian ("normal"), Cauchy, student's-t.

An example of a random variable X that is neither discrete nor continuous type is any random variable with a CDF $F_X(x)$ that is continuous and strictly increasing over some interval (a, b), but discontinuous at some point outside (a, b). Often, the CDF of such a random variable is the integral of a *generalized PDF* that contains impulses at the points where the CDF has discontinuous jumps. There are also (rare) examples of random variables with continuous CDF functions but with no PDF.

O. Computing expectations

Assume:

- $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ are measurable functions.
- X and Y are discrete random variables that take values in the finite or countably infinite sets S_X and S_Y, respectively.
- U and V are continuous random variables with joint PDF $f_{U,V}(u, v)$ and marginal PDFs $f_U(u)$ and $f_V(v)$.
- $\mathbb{E}[h(X)]$, $\mathbb{E}[g(X, Y)]$, $\mathbb{E}[h(U)]$, $\mathbb{E}[g(U, V)]$, $\mathbb{E}[g(X, V)]$ are finite.

It can be shown:[13]

$$\begin{aligned}
\mathbb{E}[h(X)] &= \sum_{x \in S_X} h(x) P[X = x] \\
\mathbb{E}[g(X, Y)] &= \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) P[X = x, Y = y] \\
\mathbb{E}[h(U)] &= \int_{-\infty}^{\infty} h(u) f_U(u) du \\
\mathbb{E}[g(U, V)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) f_{U,V}(u, v) du dv \\
\mathbb{E}[g(X, V)] &= \sum_{x \in S_X} \left[\int_{-\infty}^{\infty} g(x, v) f_{V|X=x}(v|x) dv \right] P[X = x] \qquad (13) \\
&= \int_{-\infty}^{\infty} \left[\sum_{x \in S_X} g(x, v) P[X = x | V = v] \right] f_V(v) dv \qquad (14)
\end{aligned}$$

[12] It is possible for a random variable $X : \Omega \to \mathbb{R}$ to have the same distribution as a discrete random variable, even though it has an uncountable image set S_X. For example, define $\Omega = \mathbb{R}$, $\mathcal{F} = Pow(\mathbb{R})$, $P : Pow(\mathbb{R}) \to [0, 1]$ by $P[A] = (1/2)1_{\{0 \in A\}} + (1/2)1_{\{1 \in A\}}$ for all $A \subseteq \mathbb{R}$. The resulting triplet $(\Omega, \mathcal{F}, P)$ can be shown to be a valid probability space. Since $\mathcal{F} = Pow(\Omega)$, all functions $X : \Omega \to \mathbb{R}$ are random variables. Define the identity function $X : \Omega \to \mathbb{R}$ by $X(\omega) = \omega$. Then $S_X = \{X(\omega) : \omega \in \Omega\} = \mathbb{R}$ (an uncountably infinite set), but $P[X \in \{0, 1\}] = 1$ and X has the $Bern(1/2)$ distribution because $P[X = 0] = P[X = 1] = 1/2$. This is also a rare example where the sample space Ω is an uncountably infinite set, but we can define the sigma algebra to be $Pow(\Omega)$ (with no contradictions). This is a valid (but overly complicated) way to build a random variable with a $Bern(1/2)$ distribution. Some would call this random variable X a Bernoulli random variable because it takes values in the set $\{0, 1\}$ almost surely; others might object because it does not take values in the set $\{0, 1\}$ *surely*. The distinction is a minor one and does not affect any probability calculations (see also item 38 in Section III-H).

[13] For example, define $Z = g(X, Y)$ for the discrete random variables X, Y. Let S_Z be the discrete set of values Z can take. Now $\mathbb{E}[Z] = \sum_{z \in S_Z} zP[Z = z] = \sum_{z \in S_Z} z \left(\sum_{(x,y) \in S_X \times S_Y : g(x,y) = z} P[X = x, Y = y] \right) = \sum_{z \in S_Z} \sum_{(x,y) \in S_X \times S_Y : g(x,y) = z} g(x, y) P[X = x, Y = y] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) P[X = x, Y = y]$. The sums are absolutely convergent and hence are finite and can be switched and summed in any order. In general, the sum of a countable number of nonnegative numbers always exists and is independent of order (possibly being ∞). If A is a finite or countably infinite set and $f : A \to \mathbb{R}$ is a function then $\sum_{a \in A} f(a)$ is finite (and yields the same value regardless of summation order) if and only if $\sum_{a \in A} |f(a)| < \infty$.

If $\{B_i\}_{i=1}^{\infty}$ are events that partition the sample space and W is any random variable with finite expectation then

$$\mathbb{E}[W] = \sum_{i=1}^{\infty} \mathbb{E}[W|B_i]\, P[B_i]$$

Similarly if W has finite expectation:

$$\begin{aligned}
\mathbb{E}[W] &= \sum_{x\in S_X} \mathbb{E}[W|X=x]\, P[X=x] \\
\mathbb{E}[W] &= \int_{-\infty}^{\infty} \mathbb{E}[W|U=u]\, f_U(u)du \\
\mathbb{E}[W] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{E}[W|U=u,V=v]\, f_{U,V}(u,v)dudv \\
\mathbb{E}[W] &= \sum_{x\in S_X} \left[\int_{-\infty}^{\infty} \mathbb{E}[W|X=x,V=v]\, f_{V|X=x}(v|x)dv\right] P[X=x] \qquad (15)\\
&= \int_{-\infty}^{\infty} \left[\sum_{x\in S_X} \mathbb{E}[W|X=x,V=v]\, P[X=x|V=v]\right] f_V(v)dv \qquad (16)
\end{aligned}$$

The above formulas can be applied, for example, to compute $\mathbb{E}[f(R,S)]$ (for some given measurable function f and some given random variables R,S) by defining $W = f(R,S)$ and using the fact that a measurable function of a finite number of random variables is again a random variable.

The formulas (13), (14), (15), (16) are given in the above form for simplicity. They implicitly assume the conditional expectations are well defined and that summations and integrations are taken only over regions of positive support. For example, it is understood that the formulas (13) and (15) can be modified to $\sum_{x\in S_X: P[X=x]>0}$. The formulas can be formally derived by using, for any event A such that $P[A]>0$ and any $v\in\mathbb{R}$ such that the PDF of V is continuous at v and $f_V(v)>0$:

$$P[A|V=v] = \frac{f_{V|A}(v)P[A]}{f_V(v)} \quad , \quad f_{V|A}(v) = \frac{P[A|V=v]f_V(v)}{P[A]}$$

and so when X is discrete and $P[X=x]>0$:

$$f_{V|X=x}(v|x) = \frac{P[X=x|V=v]f_V(v)}{P[X=x]} \qquad (17)$$

For example, temporarily ignoring technical issues of "divide by zero" and "PDF continuity," we have

$$\begin{aligned}
\mathbb{E}[g(X,V)] &= \sum_{x\in S_X} \mathbb{E}[g(X,V)|X=x]\, P[X=x] \\
&= \sum_{x\in S_X} \mathbb{E}[g(x,V)|X=x]\, P[X=x] \\
&= \sum_{x\in S_X} \left[\int_{-\infty}^{\infty} g(x,v) f_{V|X=x}(v|x)dv\right] P[X=x] \qquad (18)\\
&\overset{(a)}{=} \sum_{x\in S_X} \left[\int_{-\infty}^{\infty} g(x,v)\frac{P[X=x|V=v]f_V(v)}{P[X=x]}dv\right] P[X=x] \\
&= \sum_{x\in S_X} \left[\int_{-\infty}^{\infty} g(x,v) P[X=x|V=v] f_V(v)dv\right] \\
&\overset{(b)}{=} \int_{-\infty}^{\infty} \left[\sum_{x\in S_X} g(x,v) P[X=x|V=v]\right] f_V(v)dv \qquad (19)
\end{aligned}$$

where equality (a) uses (17); equality (b) switches the order of summation and integration and is formally justified by the Fubini-Tonelli theorem of measure theory (and the fact that $\mathbb{E}[g(X,V)]$ is finite). Observe that (18) is identical to formula (13); (19) is identical to formula (14). The derivation of formulas (15) and (16) is similar. It should be emphasized that the above derivation is "stylized" because it ignores conditions for (17) to make sense ($P[X=x]>0$, $f_V(v)>0$, the PDF f_V is continuous at v). In most cases when these formulas are applied, the PDFs or conditional PDFs are continuous or piecewise continuous over a region of positive support. The integrals and summations can be taken over appropriate terms and intervals.

The formula (17) treats the case when X is discrete type and V is continuous type. When U,V are continuous with joint PDF $f_{U,V}(u,v)$ the formula becomes $f_{V|U}(v|u) = \frac{f_{U|V}(u|v)f_V(v)}{f_U(u)}$ via the definition $f_{U|V}(u|v) = \frac{f_{U,V}(u,v)}{f_V(v)}$.

P. Exercises (computing expectations)

The following multiple choice problems can be solved by computer. For example, to verify that $\sum_{i=1}^{\infty} i(1/2)^i = 2$ see: `https://www.wolframalpha.com/input?i=sum+i*%281%2F2%29%5Ei%2C+i%3D1..%5Cinfty`

The problems ask you to compute expectations. Assume X, Y, U, V are mutually independent random variables and

- X and Y are discrete type with $X \sim Geom(1/2)$, $Y \sim$ Unif$\{1, 2, 3\}$.
- U and V are continuous type with $U \sim Exp(2)$ and $V \sim$ Unif$[0, 1]$.

1) $\mathbb{E}[X]$
   a) 3
   b) 2
   c) 2.01
   d) -4

2) $\mathbb{E}[XY]$
   a) 4
   b) 3.01
   c) 3
   d) 0.5

3) $\mathbb{E}[X\cos(X)]$
   a) 0.8763390100
   b) -0.3222562303
   c) 2.5489534553
   d) 0.7427489902

4) $\mathbb{E}[\cos(X+Y)]$
   a) 0.5001128203
   b) -0.0034884486
   c) 0.4784639274
   d) -0.3820557339

5) $\mathbb{E}\left[\frac{X+\cos(XY^2)}{1+X^2}\right]$
   a) 0.2117893563
   b) -1.3378566408
   c) 0.3256249016
   d) 0.3899440031

6) $\mathbb{E}[U]$
   a) 3
   b) 2
   c) 1/2
   d) $1/\pi$

7) $\mathbb{E}\left[U^2+U\right]$
   a) 3
   b) 4
   c) 5
   d) 1

8) $\mathbb{E}\left[U\cos(U)\sin(U^2)\right]$
   a) 0.334671
   b) 0.224671
   c) 0.114671
   d) 0.004671

9) $\mathbb{E}[\cos(U+V)]$
   a) 0.48930

b) 0.58930
c) 0.68930
d) 0.10334

10) $\mathbb{E}\left[\frac{U}{1+V^2}\right]$
a) 0.39270
b) 0.12124
c) 0.00001
d) 0.29938

11) $\mathbb{E}\left[\cos(X+V)\right]$
a) -0.11874
b) -0.24863
c) 0.34998
d) 0

12) $\mathbb{E}\left[\frac{X}{1+V^2}\right]$
a) 1.5708
b) 1.9903
c) 1.1001
d) 2.3202

13) Suppose $\mathbb{E}\left[W|X=k\right]=k^2$ for all relevant k. $\mathbb{E}\left[W\right]=?$
a) 5.5
b) 6
c) 6.5
d) 7

14) Suppose $\mathbb{E}\left[W|X=k\right]=\cos(k+4.5)$ for all relevant k. $\mathbb{E}\left[W\right]=?$
a) 0.101136
b) 0.205982
c) -0.490835
d) 0.573531

15) Suppose $\mathbb{E}\left[W^2|X=n,Y=m\right]=\exp(-(n+m))$ for all relevant n,m. $\mathbb{E}\left[W^2\right]=?$
a) 0.2098740
b) 0.0415488
c) 0.0539852
d) 0.0659552

16) Suppose $\mathbb{E}\left[W|U=u,V=v\right]=uv$ for all relevant u,v. $\mathbb{E}\left[W\right]=?$
a) 1/2
b) 1/4
c) 1/8
d) 1/16

17) Suppose $\mathbb{E}\left[W|U=u,V=v\right]=\cos(u+v^2)$ for all relevant u,v. $\mathbb{E}\left[W\right]=?$
a) 0.599512
b) 0.699512
c) 0.399512
d) 0.299512

18) Suppose $\mathbb{E}\left[e^W|X=n,V=v\right]=nv$ for all relevant n,v. $\mathbb{E}\left[e^W\right]=?$
a) 0.89887
b) 0.78994
c) 1
d) 0.93032

19) Suppose $\mathbb{E}[W|X=n,V=v]=\cos(n+v)$ for all relevant n,v. $\mathbb{E}[W]=$?
    a) 0.36561
    b) -0.18879
    c) 0.19454
    d) -0.24863

20) Give a formula for $\mathbb{E}[r(U,V,W)]$ given a measurable function r and a joint PDF $f_{U,V,W}(u,v,w)$.
21) Let A be an event and suppose $P[A|X=x]$ is known for all relevant x. Give a formula for $P[A]$. Design and solve a problem of this type.
22) Let A be an event and suppose $P[A|X=x,Y=y]$ is known for all relevant x,y. Give a formula for $P[A]$. Design and solve a problem of this type.
23) Let A be an event and suppose $P[A|U=u]$ is known for all relevant u. Give a formula for $P[A]$. Design and solve a problem of this type.
24) Let A be an event and suppose $P[A|U=u,V=v]$ is known for all relevant u,v. Give a formula for $P[A]$. Design and solve a problem of this type.
25) Let A be an event and suppose $P[A|X=x,V=v]$ is known for all relevant x,v. Give a formula for $P[A]$. Design and solve a problem of this type.
26) Let X and Y be independent discrete random variables with finite image sets S_X and S_Y. Prove $\mathbb{E}[XY]=\mathbb{E}[X]\mathbb{E}[Y]$ by the fact $P[X=x,Y=y]=P[X=x]P[Y=y]$ for all $(x,y)\in S_X\times S_Y$.
27) Let X be a discrete random variable with finite image set S_X. Let $g:\mathbb{R}\to\mathbb{R}$ be *any* function and define $Y=g(X)$. Argue that Y is a discrete random variable with a finite image set S_Y. Show that $\sum_{y\in S_Y} yP[Y=y]=\sum_{x\in S_X} g(x)P[X=x]$.
28) Let X,Y be discrete random variables with finite image sets S_X,S_Y. Let $h:\mathbb{R}^2\to\mathbb{R}$ be *any* function and define $Z=h(X,Y)$. Argue that Z is a discrete random variable with a finite image set S_Z. Show that $\sum_{z\in S_Z} zP[Z=z]=\sum_{x\in S_X}\sum_{y\in S_Y} h(x,y)P[X=x,Y=y]$.

Q. Make your own examples

1) Give an example of random variables X,Y that are dependent.
2) Give an example of random variables X,Y such that $X\sim$ Unif$[0,1]$, Y is continuous type, and $\mathbb{E}[Y|X=x]=3x$ for all $x\in[0,1]$. Are X and Y independent?
3) Give an example of random variables X,Y such that $X\sim$ Unif$[0,1]$, Y is discrete type, and $\mathbb{E}[Y|X=x]=3x$ for all $x\in[0,1]$. Are X and Y independent?
4) Give an example of random variables X,Y such that $X\sim$ Unif$[0,1]$, and $\mathbb{E}[Y|X=x]=x^2$ for all $x\in[0,1]$. Are X and Y independent?
5) Give an example of a random variable X such that $\mathbb{E}[X]$ is finite but $\mathbb{E}\left[X^2\right]=\infty$.
6) Give an example of a random variable such that $\mathbb{E}[X]=DNE$ (Does Not Exist).
7) Give an example of a continuous type random variable with $\mathbb{E}[X]=1$ and $Var(X)=3$.
8) Give an example of a discrete type random variable with $\mathbb{E}[X]=1$ and $Var(X)=3$.
9) Give an example of a random variable X with $\mathbb{E}[X]=0$, $\mathbb{E}\left[X^2\right]=8$, $\mathbb{E}\left[X^4\right]=65$.
10) Is it possible to construct a random variable X with $\mathbb{E}[X]=0$, $\mathbb{E}\left[X^2\right]=8$, $\mathbb{E}\left[X^4\right]=60$?
11) Give an example of *dependent* random variables X,Y such that $X\in\{0,1\}$ and $\mathbb{E}[Y|X=x]=0$ for all $x\in\{0,1\}$.
12) Give an example of random variables X,Y that are uncorrelated but not independent.
13) Give an example of something interesting that involves random variables.

R. Manufacturing random variables

Fix $(\Omega,\mathcal{F},P)$. Suppose X and U are independent random variables such that $X\sim Bern(1/2)$ and $U\sim$ Unif$[0,1]$. We want to manufacture new random variables as measurable functions of X and U.

1) Build a random variable $Y\sim Bern(1/7)$.
2) Build a random variable Z with $\mathbb{E}[Z]=0$ and $Var(Z)=100$.
3) Build a random variable $R\sim Bern(1/2)$ such that $\mathbb{E}[XR]=1/8$.
4) Build a random variable $W\sim Bern(1/2)$ such that $\mathbb{E}[XW]$ is maximized.
5) Build a random variable $H\sim Bern(1/4)$ such that $\mathbb{E}[XH]$ is maximized.
6) Build a random variable J such that $\mathbb{E}[XJ]=8$ and $\mathbb{E}[UJ]=-1$.
7) Build a random variable K such that $\mathbb{E}[K]=42$ and X,U,K are mutually independent.
8) Build a random variable $A\sim$ Unif$[0,1]$ such that $\mathbb{E}[XA]=1/8$.

S. Expectation inequalities

If X and Y are random variables with finite expectations then it can be shown

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

As a consequence, if X is a nonnegative random variable (so $X(\omega) \geq 0$ for all $\omega \in \Omega$) then for all $\epsilon > 0$ we have

$$\left(X \geq \epsilon 1_{\{X \geq \epsilon\}}\right) \implies \left(\mathbb{E}[X] \geq \epsilon \mathbb{E}\left[1_{\{X \geq \epsilon\}}\right]\right)$$

which proves the *Markov inequality* for nonnegative random variables (for any $\epsilon > 0$):

$$P[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}$$

If X has mean μ and finite second moment then the Markov inequality implies the following *Chebyshev inequality*: For all $\epsilon > 0$ we have

$$P[|X-\mu| \geq \epsilon] = P[|X-\mu|^2 \geq \epsilon^2] \leq \frac{\mathbb{E}\left[|X-\mu|^2\right]}{\epsilon^2} = \frac{Var(X)}{\epsilon^2} \tag{20}$$

If X and Y are random variables with finite second moments $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[Y^2\right]$, then XY has finite expectation and

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

where the second inequality is a form of the *Cauchy-Schwarz inequality* for random variables X, Y with finite second moments:

$$\mathbb{E}[XY]^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right] \tag{21}$$

which can be proven by first assuming $0 < \mathbb{E}\left[Y^2\right] < \infty$, observing the function $f(a) = \mathbb{E}\left[(X-aY)^2\right]$ is always nonnegative, and then finding the minimizer $a^* \in \mathbb{R}$ (the case $\mathbb{E}\left[Y^2\right] = 0$ can be treated separately). The Cauchy-Schwarz inequality (21) is met with equality when $X = aY$ for some $a \in \mathbb{R}$. More generally, the Cauchy-Schwarz inequality (21) is met with equality if and only if either $P[Y=0] = 1$ or $P[X = aY] = 1$ for some $a \in \mathbb{R}$.

T. Using words to define random variables

Sometimes a probability problem is presented with given random variables, so existence of a probability space $(\Omega, \mathcal{F}, P)$ that houses those random variables can be assumed. Other times we have the opposite situation: We want to formulate and solve our *own* probability problem. In this case, the problem and the random variables of interest should be formulated with clear and consistent words. For example, suppose we independently flip 5 fair coins. Using clear and consistent words, we define[14]

- $X =$ Number of flips that are HEADS
- $Y =$ Number of flips that are TAILS
- For $i \in \{1,2,3,4,5\}$ define $H_i = 1$ if flip i is HEADS (else define $H_i = 0$).

The coin flips are assumed to be *ordered* so we can distinguish flip 1, flip 2, and so on. The random variables $\{H_i\}_{i=1}^5$ are known to be i.i.d. $Bernoulli(1/2)$ because the flips were stated to be both *independent* and *fair*. While the sample space Ω and the sigma algebra $\mathcal{F}$ have not been specified, *it is implicitly assumed these things exist* and that X, Y, H_i are valid random variables (meaning they satisfy (11)). This existence is enough to apply the properties of probability and expectation:

$$X + Y = 5 \implies \mathbb{E}[X] + \mathbb{E}[Y] = 5 \tag{22}$$

$$X = \sum_{i=1}^5 H_i \implies \mathbb{E}[H] = \sum_{i=1}^5 \mathbb{E}[H_i] = \sum_{i=1}^5 P[H_i = 1] = \frac{5}{2} \tag{23}$$

$$Var(X) = Var\left(\sum_{i=1}^5 H_i\right) = \sum_{i=1}^5 Var(X_i) = \frac{5}{4} \tag{24}$$

where (24) has used the fact that $\{H_i\}_{i=1}^5$ are mutually independent (and hence pairwise uncorrelated).

The equations $X + Y = 5$ and $X = \sum_{i=1}^5 H_i$ indeed mean

$$X(\omega) + Y(\omega) = 5 \quad \forall \omega \in \Omega$$

$$X(\omega) = \sum_{i=1}^5 H_i(\omega) \quad \forall \omega \in \Omega$$

[14] An example of words that are not clear and consistent is this: "Let X be the number of heads, let Y be the number of tails, and assume X and Y are independent." The assumption about independence of X and Y is not consistent with the fact $X + Y = 5$. It is fine to define X as the number of heads and Y as the number of tails, but the joint distribution of (X, Y) must be consistent with the rest of the problem. In this case, the joint distribution of (X, Y) must be *derived* from the basic description of the problem, not *defined*.

There are infinitely many probability spaces $(\Omega, \mathcal{F}, P)$ that fit this coin flipping problem. For example, consider the finite sample space

$$\Omega = \{(z_1, \ldots, z_5) : z_i \in \{H, T\} \quad \forall i \in \{1, ..., 5\}\}$$

with $\mathcal{F} = Pow(\Omega)$. Define $P : Pow(\Omega) \to [0, 1]$ as the probability measure where all 2^5 outcomes $\omega \in \Omega$ are equally likely. Using $\omega = (z_1, \ldots, z_5)$, define $X : \Omega \to \mathbb{R}$, $Y : \Omega \to \mathbb{R}$, $H_i : \Omega \to \mathbb{R}$ for $i \in \{1, \ldots, 5\}$ by

$$X(\omega) = \sum_{i=1}^{5} 1_{\{z_i = H\}} \quad \forall \omega \in \Omega$$
$$Y(\omega) = 5 - X(\omega) \quad \forall \omega \in \Omega$$
$$H_i(\omega) = 1_{\{z_i = H\}} \quad \forall \omega \in \Omega$$

where notation 1_A denotes a function that is 1 if condition A is true, and 0 else. These functions X, Y, H_i are indeed functions from Ω to $\mathbb{R}$. The measurability requirement (11) is automatically satisfied because $\mathcal{F} = Pow(\Omega)$, so X, Y, H_i are indeed valid random variables. Other probability spaces $(\Omega, \mathcal{F}, P)$ could be used for this same problem, including spaces that contain results for an infinite sequence of coin flips (where we only care about the first five flips), or spaces that include results about flipping coins and rolling dice (where we only care about flipping coins).

What if we have a problem that requires an uncountably infinite sample space Ω? Then we typically *cannot* use $\mathcal{F} = Pow(\Omega)$. For example, consider throwing a dart at a circular dart board of radius 1. Suppose the location of the dart is uniformly distributed over the board. We can define the uncountably infinite sample space of all possible locations:

$$\Omega = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\} \tag{25}$$

It can be shown that there is a sigma algebra on $\mathbb{R}^2$, called $\mathcal{B}(\mathbb{R}^2)$, which consists of all subsets of $\mathbb{R}^2$ with a well defined *area* (possibly having area 0 or ∞) and a function $\mu : \mathcal{B}(\mathbb{R}^2) \to [0, \infty) \cup \{\infty\}$, called the *standard Borel measure on* $\mathbb{R}^2$, that satisfies countable summability and such that for all $A \in \mathcal{B}(\mathbb{R}^2)$ the value $\mu(A)$ is the area of A. Also, it holds that

$$\Omega \in \mathcal{B}(\mathbb{R}^2)$$
$$\mu(\Omega) = \pi$$

For our dartboard problem we can formally define $\mathcal{F}$ and $P : \mathcal{F} \to [0, 1]$ by

$$\mathcal{F} = \{A \in \mathcal{B}(\mathbb{R}^2) : A \subseteq \Omega\}$$
$$P[A] = \frac{\mu(A)}{\pi} \quad \forall A \in \mathcal{F}$$

It can be shown that $\mathcal{F}$ is a sigma algebra on Ω and P is a probability measure. For each $A \in \mathcal{F}$, the value $P[A]$ is equal to the ratio of the area of set A to the total area of the dart board (which is π). The random location of the dart can be represented by a pair of random variables (X, Y) where $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are defined (with $\omega = (x, y)$) by

$$X(x, y) = x \quad \forall (x, y) \in \Omega$$
$$Y(x, y) = y \quad \forall (x, y) \in \Omega$$

These X and Y functions are continuous functions from a Borel measurable subset of $\mathbb{R}^2$ to the real number line, and hence can be shown to be Borel measurable. In particular, it can be shown that they satisfy the measurability requirement (11) so they are valid random variables. Cases where three darts are independently thrown at the same dart board can be treated by using an advanced concept of *product space* that allows taking the probability space $(\Omega, \mathcal{F}, P)$ that was constructed for one dart, and then expanding it to form a new probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ that allows three "independent copies" of the single dart experiment. Specifically

$$\tilde{\Omega} = \Omega \times \Omega \times \Omega = \{(\omega_1, \omega_2, \omega_3) : \omega_i \in \Omega \quad \forall i \in \{1, 2, 3\}\}$$

while definitions for $\tilde{\mathcal{F}}$ and $\tilde{P}$ are a bit more complicated (see Appendix D and [7][3]).[15]

What if we define our own problem and our own random variables with words such as "let X be the number of devices connected to the communication network at a certain time" or "let Y be the number of people in a given population that have a certain disease." How can we know if there exists a probability space $(\Omega, \mathcal{F}, P)$ that fits the problem of interest? How can we know if X and Y truly satisfy the measurability requirement (11)? Fortunately, most practical problems indeed have probability spaces $(\Omega, \mathcal{F}, P)$ that fit (and that make our functions valid random variables), so we can often assume existence of $(\Omega, \mathcal{F}, P)$ without explicitly stating what it is. In more detail, we have two rules of thumb:

[15]For this product space, $\tilde{\mathcal{F}}$ and $\tilde{P}$ are often denoted $\tilde{\mathcal{F}} = \mathcal{F} \otimes \mathcal{F} \otimes \mathcal{F}$ and $\tilde{P} = P \otimes P \otimes P$, where the notation $\otimes$ acts as a reminder that product spaces are being used. The set $\mathcal{F} \otimes \mathcal{F} \otimes \mathcal{F}$ is defined as the sigma algebra generated by the collection of sets of the type $A \times B \times C$ where $A, B, C \in \mathcal{F}$. It can be shown there is a unique probability measure $\tilde{P} : \tilde{\mathcal{F}} \to [0, 1]$ that satisfies $\tilde{P}[A \times B \times C] = P[A]P[B]P[C]$ for all $A, B, C \in \mathcal{F}$. See details in [7][3].

1) Suppose the problem can be described by a finite or countably infinite sample space Ω. Then we can use $\mathcal{F} = Pow(\Omega)$. This means that all functions from Ω to $\mathbb{R}$ automatically satisfy the measurability requirement (11) and hence are random variables. In this case, the probability measure is the one that specifies the probability mass $P[\{\omega\}]$ for all $\omega \in \Omega$.
2) Suppose the problem cannot be described by a finite or countably infinite sample space, but it can (in principle) be simulated on a computer using a sequence of i.i.d. Unif$[0,1]$ random variables that are the result of sequential calls to a random number generator. As argued in Appendix D, the problem can be housed in the probability space $([0,1], \mathcal{B}([0,1]), \mu)$, so that $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}([0,1])$, and $P = \mu$.

IV. Convergence

A. Convergence surely, almost surely, in probability, in mean square

Fix a probability space $(\Omega, \mathcal{F}, P)$. Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of random variables. Let Y be another random variable.

Definition 3: We say $Y_n \to Y$ *surely* if

$$\lim_{n\to\infty} Y_n(\omega) = Y(\omega) \quad \forall \omega \in \Omega$$

Definition 4: We say $Y_n \to Y$ *almost surely* (also called *convergence with probability 1*) if

$$P\left[\left\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right\}\right] = 1$$

Definition 5: We say $Y_n \to Y$ *in probability* if for all $\epsilon > 0$ we have

$$\lim_{n\to\infty} P[|Y_n - Y| \geq \epsilon] = 0$$

Definition 6: We say $Y_n \to Y$ *in mean square* if $\mathbb{E}\left[Y_n^2\right] < \infty$ for all $n \in \{1,2,3,\ldots\}$ and

$$\lim_{n\to\infty} \mathbb{E}\left[(Y_n - Y)^2\right] = 0$$

The set $\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}$ used in Definition 4 is formally shown to be in $\mathcal{F}$ in Appendix C. With respect to these types of convergence, it can be shown that

$$\text{surely} \implies \text{almost surely} \implies \text{in probability}$$

The Chebyshev inequality can be used to show:[16]

$$\text{mean square} \implies \text{in probability}$$

Finally, it can be shown that if $Y_n \to Y$ in probability then a property called *convergence in distribution* holds:

$$\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$$

for all $y \in \mathbb{R}$ at which F_Y is continuous, where $F_{Y_n}(y) = P[Y_n \leq y]$ and $F_Y(y) = P[Y \leq y]$ are the CDF functions for Y_n and Y, respectively.

B. Law of large numbers (LLN)

Random variables $\{X_i\}_{i=1}^{\infty}$ are said to be *independent and identically distributed* (i.i.d.) if they are mutually independent and if all random variables have the same marginal CDF, so that

$$F_{X_i}(x) = F_{X_1}(x) \quad \forall x \in \mathbb{R}, \forall i \in \{1,2,3,\ldots\}$$

Theorem 2: (Law of Large Numbers – LLN) If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. random variables with finite mean μ (so $\mathbb{E}[X_i] = \mu$ for all $i \in \{1,2,3,\ldots\}$), then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu \quad \text{almost surely}$$

Proof: A proof for the special case when the random variables have a finite variance is given in Appendix B. A full proof that allows $Var(X_i) = \infty$ is given in [8]. □

[16]If there are real numbers $c > 0, \epsilon > 0$ such that $\mathbb{E}\left[|Y_n|^{2+\epsilon}\right] \leq c$ for all $n \in \{1,2,3,...\}$, then mean square convergence is *equivalent* to in probability convergence. This can be seen for the special case when $Y_n \to 0$ in probability by fixing $\delta > 0$, defining $A_n = \{|Y_n| > \delta\}$, and then $\mathbb{E}\left[Y_n^2\right] \leq \delta^2 + \mathbb{E}\left[Y_n^2|A_n\right] P[A_n]$. Jensen's inequality for the concave function $f(t) = t^{\frac{2}{2+\epsilon}}$ for all $t \geq 0$ gives $\mathbb{E}\left[Y_n^2|A_n\right] = \mathbb{E}\left[(|Y_n|^{2+\epsilon})^{\frac{2}{2+\epsilon}}|A_n\right] \leq \mathbb{E}\left[|Y_n|^{2+\epsilon}|A_n\right]^{\frac{2}{2+\epsilon}}$. Counter-examples exist when there are no such c, ϵ values (see item 2 in Section IV-H).

The above theorem requires the random variables X_i to have a finite mean, but allows them to have a possibly infinite variance. This version of the LLN is often called the *strong law of large numbers* because it treats convergence almost surely. The almost sure convergence in Theorem 2 means that if we define A as the set of all outcomes in the sample space for which the limit converges to μ, then $P[A] = 1$. That is

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) = \mu\right\} \tag{26}$$

and $P[A] = 1$. Defining $A^c = \Omega \setminus A$ as the complement of set A, it follows that $P[A^c] = 0$. The set A^c contains all outcomes for which the limit either does not exist, or the limit exists but is not equal to μ. The set A^c may contain uncountably many outcomes, so it is remarkable that the LLN ensures $P[A^c] = 0$.

Appendix B shows that convergence almost surely implies the weaker *convergence in probability*: For all $\epsilon > 0$ we have

$$\lim_{n\to\infty} P\left[\left|\mu - \frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq \epsilon\right] = 0 \tag{27}$$

In many applications the i.i.d. random variables $\{X_i\}_{i=1}^{\infty}$ have both finite mean and finite variance, so $\mathbb{E}[X_i] = \mu$ and $Var(X_i) = \sigma^2$ for all $i \in \{1, 2, 3, \ldots\}$. In this case it holds by *linearity of expectation* and basic properties of variance that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu \quad \forall n \in \{1, 2, 3, \ldots\}$$

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \overset{(a)}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n} \quad \forall n \in \{1, 2, 3, \ldots\}$$

where step (a) uses the fact that the variance of a sum of pairwise uncorrelated random variables is the sum of variances. It follows that the mean of $\frac{1}{n}\sum_{i=1}^{n} X_i$ is always μ, while its variance converges to 0 as $n \to \infty$. This provides a great deal of intuition for the LLN. Further, the fact that the variance converges to zero can be used, together with the Markov/Chebyshev inequality, to directly prove the weak convergence result given in (27). More work is required to prove the stronger result for almost sure convergence given in Theorem 2 (see Appendix B and [8]).

C. Applying the LLN twice

If A and B are events that satisfy $P[A] = 1$ and $P[B] = 1$, the axioms of probability can be used to prove $P[A \cap B] = 1$. This fact impacts the following important example. Assume:

- $\{X_i\}_{i=1}^{\infty}$ are i.i.d. random variables with $\mathbb{E}[X_i] = 5.7$ for all $i \in \{1, 2, 3, ...\}$.
- $\{T_i\}_{i=1}^{\infty}$ are i.i.d. random variables with $\mathbb{E}[T_i] = 17.1$ for all $i \in \{1, 2, 3, ...\}$.
- Each T_i is surely positive. That is, $T_i(\omega) > 0$ for all $\omega \in \Omega$, $i \in \{1, 2, 3, ...\}$.

In this example, the random variables $\{X_i\}_{i=1}^{\infty}$ are assumed to be i.i.d. amongst themselves; the random variables $\{T_i\}_{i=1}^{\infty}$ are i.i.d. amongst themselves. However, there can be arbitrary dependencies between the X_i and T_j variables. For example, we might have $T_i = \cos(X_i) + 17.1$ for all $i \in \{1, 2, 3, \ldots\}$.

For each positive integer n, define the random variable Z_n by

$$Z_n = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} T_i}$$

The random variables T_i are assumed to be surely positive to avoid the undefined case of division by zero.

Does Z_n converge to anything? If so, does it converge almost surely? To answer this question, let us first fix a particular outcome $\omega^* \in \Omega$. Then $\{X_i(\omega^*)\}_{i=1}^{\infty}$, $\{T_i(\omega^*)\}_{i=1}^{\infty}$, and $\{Z_n(\omega^*)\}_{n=1}^{\infty}$ are just sequences of real numbers (they are deterministic, not random). Suppose ω^* is an outcome under which these sequences satisfy

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega^*) = 5.7 \tag{28}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} T_i(\omega^*) = 17.1 \tag{29}$$

where limits of real-valued sequences are defined in basic calculus. Then we can conclude:

$$\begin{aligned}\lim_{n\to\infty} Z_n(\omega^*) &\overset{(a)}{=} \lim_{n\to\infty} \frac{\sum_{i=1}^n X_i(\omega^*)}{\sum_{i=1}^n T_i(\omega^*)} \\ &\overset{(b)}{=} \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^n X_i(\omega^*)}{\frac{1}{n}\sum_{i=1}^n T_i(\omega^*)} \\ &\overset{(c)}{=} \frac{5.7}{17.1} \end{aligned} \tag{30}$$

where the equality (a) uses the definition of $Z_n(\omega^*)$; the equality (b) multiplies and divides by $1/n$; the equality (c) uses (28)-(29) and the basic calculus fact that the limit of a ratio is equal to the ratio of limits (provided the limits exist, are finite, and the limit in the denominator is nonzero). Therefore, if our particular ω^* satisfies *both* (28) *and* (29), then we can conclude

$$\lim_{n\to\infty} Z_n(\omega^*) = \frac{5.7}{17.1}$$

Of course, in a real problem, the outcome $\omega \in \Omega$ is chosen *randomly* by nature (according to the probability measure P). How likely is it that the particular ω chosen by nature satisfies *both* (28) and (29)? The LLN applied to the i.i.d. sequence $\{X_i\}_{i=1}^\infty$ ensures that our randomly selected ω satisfies (28) *with probability 1*. The LLN applied to the i.i.d. sequence $\{T_i\}_{i=1}^\infty$ ensures that our randomly selected ω satisfies (29) *with probability 1*. Hence, our randomly selected ω satisfies *both* (28) and (29) with probability 1. We formalize this argument in the following claim.

Claim: We have

$$Z_n \to \frac{5.7}{17.1} \quad \text{almost surely}$$

Proof: Define the following events A, B, C:[17]

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i(\omega) = 5.7\right\}$$
$$B = \left\{\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n T_i(\omega) = 17.1\right\}$$
$$C = \left\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = \frac{5.7}{17.1}\right\}$$

The LLN ensures $P[A] = 1$ and $P[B] = 1$. Thus, $P[A \cap B] = 1$. However, if $\omega \in A \cap B$, we know (by the fact that the limit of a ratio is the ratio of limits) that $\omega \in C$. Indeed, just repeat the argument (30) but change ω^* to ω. It follows that

$$A \cap B \subseteq C$$

Thus

$$\underbrace{P[A \cap B]}_{1} \leq P[C]$$

Since the left-hand-side of the above inequality is 1, and $P[C]$ cannot be larger than 1, it follows that $P[C] = 1$. □

We have made good use of the following probability fact:

$$(P[A] = 1 \text{ and } P[B] = 1) \implies (P[A \cap B] = 1)$$

The following theorem is a generalized version of this fact.

Theorem 3: Let $\{A_i\}_{i=1}^\infty$ be a countably infinite sequence of events.
a) If $P[A_i] = 0$ for all $i \in \{1, 2, 3, \ldots\}$ then $P[\cup_{i=1}^\infty A_i] = 0$.
b) If $P[A_i] = 1$ for all $i \in \{1, 2, 3, \ldots\}$ then $P[\cap_{i=1}^\infty A_i] = 1$.
Proof: To prove part (a), suppose $P[A_i] = 0$ for all $i \in \{1, 2, 3, \ldots\}$. The union bound implies

$$P[\cup_{i=1}^\infty A_i] \leq \sum_{i=1}^\infty P[A_i] = \sum_{i=1}^\infty 0 = 0$$

This proves part (a).

To prove part (b), suppose $P[A_i] = 1$ for all $i \in \{1, 2, 3, \ldots\}$. It suffices to prove that

$$P[(\cap_{i=1}^\infty A_i)^c] = 0$$

[17] A representation in (44) of Appendix C formally proves the sets A, B, C are events.

We have by De Morgan's law

$$(\cap_{i=1}^{\infty} A_i)^c = \cup_{i=1}^{\infty} A_i^c$$

Thus

$$P[(\cap_{i=1}^{\infty} A_i)^c] = P[\cup_{i=1}^{\infty} A_i^c] \leq \sum_{i=1}^{\infty} P[A_i^c] = 0$$

which uses the fact that $P[A_i^c] = 0$ for all $i \in \{1, 2, 3, \ldots\}$. □

D. Exercises related to i.i.d. dice rolls

Consider an infinite sequence of fair and independent dice rolls. Model the system with probability space $(\Omega, \mathcal{F}, P)$ where

$$\Omega = \{(r_1, r_2, r_3, ...) : r_i \in \{1, 2, 3, 4, 5, 6\} \quad \forall i \in \mathbb{N}\}$$

so each outcome $\omega \in \Omega$ has the form $\omega = (r_1, r_2, r_3, ...)$. For $i \in \mathbb{N}$ define

$$X_i(\omega) = X_i(r_1, r_2, r_3, ...) = r_i \quad \forall \omega \in \Omega$$

Assume $P : \mathcal{F} \to [0, 1]$ is such that $\{X_i\}_{i=1}^{\infty}$ are i.i.d. random variables and

$$P[X_i = r] = 1/6 \quad \forall r \in \{1, 2, 3, 4, 5, 6\}$$

Define $X = X_1$ and observe that $\mathbb{E}[X] = 3.5$. Define A as in (26) (using $\mu = 3.5$).

1) The law of large numbers says that $P[A]$ and $P[A^c]$ have what values?
2) Define $\omega_1 = (1, 1, 1, 1, 1, 1, 1, 1, \ldots)$ as the outcome where all rolls are 1. Is $\omega_1 \in A$?
3) Define $\omega_2 = (5, 6, 5, 6, 5, 6, 5, 6, \ldots)$ as the outcome where rolls alternate between 5 and 6. Is $\omega_2 \in A$?
4) Define $\omega_3 = (3, 4, 3, 4, 3, 4, 3, 4, \ldots)$ as the outcome where rolls alternate between 3 and 4. Is $\omega_3 \in A$?
5) Specify two new outcomes ω_4, ω_5 such that $\{\omega_4, \omega_5\} \subseteq A^c$.
6) Specify two new outcomes ω_6, ω_7 such that $\{\omega_6, \omega_7\} \subseteq A$.
7) Explain why there are infinitely many outcomes in A^c (in fact, uncountably many).
8) Explain why there are infinitely many outcomes in A (in fact, uncountably many).
9) Compute $P[\cap_{i=1}^{n} \{X_i = 1\}]$ for all $n \in \mathbb{N}$. Compute $P[\{\omega_1\}] = P[\cap_{i=1}^{\infty} \{X_i = 1\}]$.
10) Compute $P[\{\omega_3\}]$.
11) For $\omega \in \Omega$, compute $P[\{\omega\}]$.
12) Complete the following two sentences using words/phrases/numbers in the set:

    {dominate, 42, countably many, uncountably many, are negligible, exciting, law of large numbers, electoral college}

    Even though there are [ ] outcomes in A^c, there are also [ ] outcomes in A. The [ ] ensures the outcomes in A collectively [ ] in terms of probability.
13) For each $n \in \mathbb{N}$, use the Markov/Chebyshev inequality to upper bound $P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \notin (3.49, 3.51)\right]$.

E. Exercises with multiple choice

For the following exercises, assume

- $\{X_i\}_{i=1}^{\infty}$ are i.i.d. $Bern(1/2)$
- $\{Y_i\}_{i=1}^{\infty}$ are i.i.d. $N(0, 1)$
- $\{Z_i\}_{i=1}^{\infty}$ are i.i.d. $Geom(1/5)$
- $\{W_i\}_{i=1}^{\infty}$ are i.i.d. $Bern(p)$
- The sequences $\{X_i\}, \{Y_i\}, \{Z_i\}, \{W_i\}$ may depend on each other

1) The expectation of an indicator function is
   a) 0
   b) the variance of its inverse
   c) an outcome
   d) the probability the indicator function is 1

2) For insight into how closely $\frac{1}{100}\sum_{i=1}^{100} W_i$ approximates p we can
   a) discard half the samples
   b) compute $Var(\frac{1}{100}\sum_{i=1}^{100} W_i)$
   c) assume $W_i \sim N(0, 1)$.
   d) use the Slytherin theorem

3) The probability that both $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = 1/2$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} Y_i = 0$ is
   a) Unknown
   b) 1
   c) 0
   d) $1/4$

4) With prob 1 $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \cos(X_i) =$
   a) 0
   b) 0.2309
   c) -0.3480
   d) 0.7702

5) With prob 1 $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (Y_i + Z_i) =$
   a) 5
   b) 25
   c) 1/10
   d) 1

6) With prob 1 $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Z_i} =$
   a) 1/20
   b) 4
   c) 1/10
   d) Unknown

7) With prob 1 $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} (X_i+Y_i)}{\sum_{i=1}^{n} (1+Y_i^2)} =$
   a) 2
   b) 1/4
   c) 1/8
   d) 1/5

8) With prob 1 $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i =$
   a) 1/2
   b) 1/3
   c) 2
   d) Unknown

9) With prob 1 $\lim_{n\to\infty} \frac{1}{8n}\sum_{i=1}^{n} X_{2i} =$
   a) 1/8
   b) 1/32
   c) 1/2
   d) 1/16

10) With prob 1 $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} e^{X_i}}{\sum_{i=1}^{n} (1+Z_i)} =$
    a) 0.2099
    b) 0.3099
    c) 0.4099
    d) 0.5099

11) $\mathbb{E}\left[\frac{1}{100}\sum_{i=1}^{100} Y_i\right] =$
    a) 5
    b) 0
    c) 10
    d) 2.5

12) $Var(\frac{1}{100}\sum_{i=1}^{100} X_i) =$
    a) 1/400

b) 1/1600
c) 1/100
d) 2/350

13) $Var(\frac{1}{1000}\sum_{i=1}^{1000} Z_i) =$
a) 0.02
b) 0.15
c) 0.90
d) 1.80

14) Explain why $Var(\frac{1}{100}\sum_{i=1}^{100} X_i Z_i)$ is unknown.

F. Design your own examples

1) Fix $p \in (0,1)$. Give an example of a sequence $\{V_i\}_{i=1}^{\infty}$ of $Bern(p)$ random variables such that $\frac{1}{n}\sum_{i=1}^{n} V_i$ does *not* converge to p almost surely as $n \to \infty$. Explain why the example requires $p \in (0,1)$ instead of $p \in [0,1]$.

2) Fix $p \in [0,1]$. Give an example of a sequence $\{V_i\}_{i=1}^{1000}$ of $Bern(p)$ random variables such that
$$Var\left(\frac{1}{1000}\sum_{i=1}^{1000} V_i\right) = \frac{p(1-p)}{2}$$

3) Give an example of mutually independent random variables $\{X_i\}_{i=1}^{\infty}$ with $Var(X_i) = 1$ for all $i \in \{1,2,3,\ldots\}$ and with distinct means $\mathbb{E}[X_i] = \mu_i$ (so that $i \neq j \implies \mu_i \neq \mu_j$), such that $\frac{1}{n}\sum_{i=1}^{n} X_i \to 0$ with probability 1.
4) Give an example of a sequence of $Bern(1/2)$ variables $\{X_n\}_{n=1}^{\infty}$ such that with probability 1 the string "00111" appears an infinite number of times, but the string "010" never appears.
5) Give an example of a sequence of mutually independent Bernoulli random variables $\{X_n\}_{n=1}^{\infty}$ such that with probability 1 the string "00111" appears an infinite number of times, but "000" never appears.
6) Give an example of a sequence of mutually independent Bernoulli random variables $\{X_n\}_{n=1}^{\infty}$ such that with probability 1 the string "00111" appears an infinite number of times, the string "000" never appears, and $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} X_i$ does not exist.

G. Exercises related to simulation

Suppose that a randomly formed network with a certain structure is connected with some unknown probability p. To estimate p, we write a computer program that loops over iterations $i \in \{1,\ldots,n\}$. At iteration i, a new network i with the desired structure is randomly generated using multiple calls to a random number generator rand(). Then, a subroutine is run to determine whether or not that particular network i is connected. Let X_i be a Bernoulli random variable that is 1 if network i is connected, and 0 else. Define $\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimate of p based on these n samples.

1) $\{X_i\}_{i=1}^{n}$ are identically distributed because
a) They are independent
b) They are part of an infinite sequence
c) The same software is used to generate each network i
d) All of the networks $i \in \{1,...,n\}$ are connected

2) $\{X_i\}_{i=1}^{n}$ are mutually independent because
a) They are part of an infinite sequence
b) The same software is used to generate each network i
c) They have the same distribution
d) Each call to rand() generates a random variable that is independent of the past

3) Which theorem relates to using $\hat{P}_n$ as an estimate of p?
a) The law of large numbers
b) The law of total probability
c) The law of total expectation
d) De Morgan's law

4) Compute $\mathbb{E}\left[\hat{P}_n\right]$ and $Var(\hat{P}_n)$ for all $n \in \{1,2,3,...\}$.

5) Fix $n = 800$. Give a numerical upper bound on the mean square error of the estimate $\hat{P}_{800}$. That is, compute an upper bound on $\mathbb{E}\left[(\hat{P}_{800} - p)^2\right]$. Your bound should be a number (not a function of p).

6) Fix $n = 800$. Use the Chebyshev inequality (see Section V) to find a 95% confidence interval for p based on 800 samples. That is, find a numerical value $\epsilon > 0$ (that does not depend on p) so that $P[p \in [\hat{P}_{800} - \epsilon, \hat{P}_{800} + \epsilon]] \geq 0.95$.

7) Fix $n = 800$. Use the Hoeffding inequality (see Section V) to find a 95% confidence interval for p based on 800 samples. That is, find a numerical value $\epsilon > 0$ (that does not depend on p) so that $P[p \in [\hat{P}_{800} - \epsilon, \hat{P}_{800} + \epsilon]] \geq 0.95$. How does the size of the interval compare to the interval obtained by the Chebyshev inequality?

8) Using the Hoeffding inequality (see Section V), how many samples n are needed to obtain a 95% confidence interval $[\hat{P}_n - 0.01, \hat{P}_n + 0.01]$ for p?

H. Additional exercises on convergence

1) Define $\Omega = \{\texttt{apple}, \texttt{banana}, \texttt{carrot}\}$ and $\mathcal{F} = Pow(\Omega)$. Assume $P[\{\texttt{apple}\}] = P[\{\texttt{banana}\}] = 1/2$ and $P[\{\texttt{carrot}\}] = 0$. For $n \in \{1, 2, 3, \ldots\}$ define $X_n : \Omega \to \mathbb{R}$ by $X_n(\texttt{apple}) = 1/n$, $X_n(\texttt{banana}) = 2^{-n}$, $X_n(\texttt{carrot}) = n$. Show $X_n \to 0$ almost surely (but not surely).
2) Let $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}\left[X^2\right] = \infty$. Define $Y_n = X/n$ for $n \in \{1, 2, 3, \ldots\}$. Show that $Y_n \to 0$ surely (and hence almost surely and in probability), but not in mean square.
3) Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$ be a random variable. Suppose $X \sim Geom(p)$ for some $p \in (0, 1]$. Define $Y_n = 2^n 1_{\{X > n\}}$ for $n \in \{1, 2, 3, ...\}$. Show that $Y_n \to 0$ surely. Compute $\mathbb{E}[Y_n]$ and $\mathbb{E}\left[Y_n^2\right]$ for all n. For which values of p does $Y_n \to 0$ in mean square?
4) Fix $(\Omega, \mathcal{F}, P)$. For simplicity, assume $\{\omega\} \in \mathcal{F}$ for all $\omega \in \Omega$. Show that in the special case when Ω is a finite or countably infinite set, then convergence in probability is equivalent to convergence almost surely. [Hint: It suffices to consider random variables Y_n that converge to 0 in probability, then show they also converge to 0 almost surely.]
5) Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. $Bern(1/2)$. Define $Y_n = X_n X_{n+1}$ for $n \in \{1, 2, 3, ...\}$. Show there is a constant $c \in \mathbb{R}$ such that $\frac{1}{n}\sum_{i=1}^{n} Y_i \to c$ with probability 1.

V. Sequences of random variables

Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of random variables, each with finite mean and variance. It holds that

$$i.i.d. \implies \text{pairwise independent} \implies \text{pairwise uncorrelated}$$

Suppose $\{X_i\}_{i=1}^{\infty}$ are pairwise uncorrelated, possibly having different distributions, but all having the same mean and variance:

$$\mathbb{E}[X_i] = \mu \quad \forall i \in \{1, 2, 3, ...\}$$
$$Var(X_i) = \sigma^2 \quad \forall i \in \{1, 2, 3, ...\}$$

where $\mu \in \mathbb{R}$ and $0 < \sigma^2 < \infty$. For each positive integer n define

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$G_n = \frac{1}{\sqrt{n\sigma^2}}\sum_{i=1}^{n}(X_i - \mu)$$

Linearity of expectation and the variance properties for pairwise uncorrelated random variables can be used to show

$$\mathbb{E}[M_n] = \mu \quad \forall n \in \{1, 2, 3, ...\}$$
$$Var(M_n) = \frac{\sigma^2}{n} \quad \forall n \in \{1, 2, 3, ...\}$$

which motivates the *law of large numbers*, and

$$\mathbb{E}[G_n] = 0 \quad \forall n \in \{1, 2, 3, ...\}$$
$$Var(G_n) = 1 \quad \forall n \in \{1, 2, 3, ...\}$$

which motivates the *central limit theorem*. In particular, if we additionally assume $\{X_i\}_{i=1}^{\infty}$ are i.i.d. then the *central limit theorem* (CLT) implies that the distribution of G_n converges to the $N(0, 1)$ distribution, being the distribution of a Gaussian random variable with mean 0 and variance 1:

$$\lim_{n\to\infty} P[G_n \leq x] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad \forall x \in \mathbb{R}$$

This type of convergence is called *convergence in distribution*. The *law of large numbers* (LLN) implies a stronger form of convergence: If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. with finite mean μ then $M_n \to \mu$ with probability 1.[18]

Since $\mathbb{E}[M_n] = \mu$ and $Var(M_n)$ is known, we can fix $\epsilon > 0$ and use the *Chebyshev inequality* (20) to conclude

$$P[|M_n - \mu| \geq \epsilon] \leq \frac{Var(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \quad \forall n \in \{1, 2, 3, ...\} \tag{31}$$

A. Hoeffding inequality

A related *Hoeffding inequality* is useful when the random variables are bounded. Here we state the result without proof: Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of mutually independent random variables (they do not need to be identically distributed). Define $\mu_i = \mathbb{E}[X_i]$ for $i \in \{1, 2, 3, \ldots\}$. If there are real numbers a_i, b_i such that[19]

$$a_i \leq X_i(\omega) \leq b_i \quad \forall i \in \{1, 2, 3, \ldots\}, \quad \forall \omega \in \Omega$$

then

- Single-sided Hoeffding inequality: For all $\epsilon > 0$ we have

$$P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \forall n \in \{1, 2, 3, \ldots\}$$

- Double-sided Hoeffding inequality: For all $\epsilon > 0$ we have

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \forall n \in \{1, 2, 3, \ldots\} \tag{32}$$

In the special case when $\{X_i\}_{i=1}^{\infty}$ are i.i.d. with mean μ we have $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) = M_n - \mu$. If μ is unknown, but if there are known a, b values that satisfy $a \leq X_i \leq b$ surely for all i, the double-sided Hoeffding inequality can be used to find a 95% confidence interval $[M_n - \epsilon, M_n + \epsilon]$ based on n i.i.d. samples. That is, we can find a radius $\epsilon > 0$ such that

$$P[\mu \in [M_n - \epsilon, M_n + \epsilon]] \geq 0.95 \tag{33}$$

simply by finding an ϵ that makes the right-hand-side of (32) equal to 0.05. Also, the right-hand-side of (32) reduces to $2\exp(\frac{-2\epsilon^2 n}{(b-a)^2})$, which decreases exponentially fast with n (as compared to the $O(1/n)$ decrease of Chebyshev in (31)). Thus, the Hoeffding inequality is typically much tighter than the Chebyshev inequality. Of course, the assumption that all random variables X_i are in the bounded interval $[a, b]$ is more detailed than the assumption that $Var(X_i) < \infty$.

An important special case is when $\{X_i\}_{i=1}^{\infty}$ are i.i.d. and $Bern(p)$ with some unknown $p \in [0, 1]$ (see also Section IV-G). In this case we can use $a = 0$, $b = 1$, and the right-hand-side of (32) becomes $2e^{-2\epsilon^2 n}$. In this special case we obtain the following radii ϵ for 95% confidence intervals $[M_n - \epsilon, M_n + \epsilon]$ for the unknown p:

$$\begin{aligned}
n = 100 &\implies \epsilon = 0.135810 \\
n = 200 &\implies \epsilon = 0.096032 \\
n = 300 &\implies \epsilon = 0.078410 \\
n = 10^3 &\implies \epsilon = 0.042947 \\
n = 10^4 &\implies \epsilon = 0.013581 \\
n = 10^5 &\implies \epsilon = 0.004295 \\
n = 10^6 &\implies \epsilon = 0.0013581
\end{aligned}$$

It should be emphasized that these radii ϵ are computed *before observing any of the random samples*. The width of the confidence interval is 2ϵ, although the left-endpoint of the confidence interval is the random variable $M_n - \epsilon$.

The inequality (33) states that, even before observing the M_n random variable, we know the probability that the random interval $[M_n - \epsilon, M_n + \epsilon]$ includes the (unknown) value μ is greater than or equal to 0.95. In particular, inequality (33) holds *before looking at the value* M_n. One may wonder what the *conditional* probability is, given we observe M_n:

$$P[\mu \in [M_n - \epsilon, M_n + \epsilon] | M_n] = ??$$

[18] While the standard statement of the LLN requires $\{X_i\}_{i=1}^{\infty}$ to be i.i.d. and to have finite mean μ (with possibly infinite variance), it can be shown that if $\{X_i\}_{i=1}^{\infty}$ are pairwise uncorrelated, have possibly different distributions with the same mean μ, and have bounded variances so that $Var(X_i) \leq \sigma_{max}^2$ for all $i \in \{1, 2, 3, \ldots\}$ for some $\sigma_{max}^2 < \infty$, then we can again conclude $M_n \to \mu$ with probability 1.

[19] The assumption that X_i is surely in $[a, b]$ can be replaced by the assumption that X_i is almost surely in $[a, b]$, with no change in any probability result.

The correct answer to this question is not particularly helpful: It depends on M_n, and it is always either 0 or 1 depending on the value M_n. That is because there is no randomness to the parameter μ: Either μ is in the given interval or it is not. For example, if we happen to observe $M_n = 0.87$ then:[20]

$$P[\mu \in [M_n - \epsilon, M_n + \epsilon] | M_n = 0.87] = P[\mu \in [0.87 - \epsilon, 0.87 + \epsilon] | M_n = 0.87] = \begin{cases} 1 & \text{if } \mu \in [0.87 - \epsilon, 0.87 + \epsilon] \\ 0 & \text{else} \end{cases}$$

and so if μ is unknown, the answer exists but cannot be evaluated. It is more useful to consider the unconditional probability in (33), which can be obtained from bounds such as the Chebyshev and Hoeffding inequalities.

B. Relation to Gaussian distributions in statistics

In the case when $\{X_i\}_{i=1}^{\infty}$ are i.i.d. $N(\mu, \sigma^2)$ with unknown μ and σ^2 (with $\mu \in \mathbb{R}$ and $0 < \sigma^2 < \infty$) a remarkable result in the theory of statistics holds: A confidence interval I_n (with random endpoints and a random width that is roughly proportional to $1/\sqrt{n}$) can be constructed from n samples, using the sample mean and sample variance, such that $P[\mu \in I_n]$ can be *exactly* computed using a student's t-distribution [1]. Specifically, for any integer $n \geq 2$ and any real number $t > 0$ we can define the interval

$$I_n = \left[M_n - t\frac{\sqrt{V_n}}{\sqrt{n}}, M_n + t\frac{\sqrt{V_n}}{\sqrt{n}} \right] \tag{34}$$

where random variables M_n and V_n are the sample mean and sample variance

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - M_n)^2$$

where dividing by $n-1$ in the sample variance, rather than n, ensures the result is unbiased. Defining the random variable T_n by[21]

$$T_n = \frac{\sqrt{n}}{\sqrt{V_n}}(M_n - \mu)$$

yields the remarkable result that T_n has a student's t-distribution with $n-1$ degrees of freedom, so $P[|T_n| > t]$ can be calculated for all $t > 0$. If we choose t such that $P[|T_n| > t] = 0.05$ then

$$0.95 = P[|T_n| \leq t] = P\left[\frac{\sqrt{n}}{\sqrt{V_n}} |M_n - \mu| \leq t \right] = P\left[|M_n - \mu| \leq t\frac{\sqrt{V_n}}{\sqrt{n}} \right] = P[\mu \in I_n]$$

The value $t > 0$ that yields $P[|T_n| > t] = 0.05$ is often called $t_{0.025,n-1}$ and is found from a lookup table to obtain the (one-sided) probability $P[T_n > t] = 0.025$. Symmetry of the PDF for a student's-t random variable with $n-1$ degrees of freedom implies $P[|T_n| > t] = 2P[T_n > t]$. The student's-$t$ CDF is continuous so $P[T_n = x] = 0$ for all $x \in \mathbb{R}$.

When the distribution of each X_i is not $N(\mu, \sigma^2)$, but when an upper bound on the variance is known, the Chebyshev inequality (31) can be used. The Chebyshev inequality is generally considered to be loose. For a tighter result when the random variables X_i are surely in some bounded interval $[a, b]$, the Hoeffding inequality (32) can be used. Observe that (31) and (32) are solid mathematical bounds that hold for any sample size $n \in \{1, 2, 3, \ldots\}$.

When the number of (non-Gaussian) samples n is large, an alternative statistical practice is to obtain *approximate bounds* via a conventional method of *batch means* that groups the i.i.d. samples into batches of sufficiently large size for the central limit theorem to suggest each batch mean is *approximately* Gaussian [1]. This can give *approximate* confidence intervals that are sometimes tighter than the (rigorous) confidence intervals obtained by Chebyshev or Hoeffding. The approximate confidence intervals are of the type (34) with the exception that the sample mean and sample variance are computed from the batch mean samples, not the raw samples. This method is particularly useful when the random variables X_i are not bounded, so that Hoeffding cannot be applied. In practice, the method of batch means often uses batches of size 15-20 samples each, although the size required for a good approximation to a Gaussian actually depends on the distribution of X_i.

[20] A formal treatment of conditional probability given a random variable M_n recognizes that there can be many *versions* of $\mathbb{E}\left[1_{\{\mu \in [M_n - \epsilon, M_n + \epsilon]\}} | M_n\right]$, every version has the form $g(M_n)$ for some measurable function $g : \mathbb{R} \to \mathbb{R}$, and any two versions $g_1(M_n)$ and $g_2(M_n)$ satisfy $P[g_1(M_n) = g_2(M_n)] = 1$. Our consideration of the case $M_n = 0.87$ is consistent with the version given by $g(M_n)$ for $g(x) = 1_{\{\mu \in [x - \epsilon, x + \epsilon]\}}$ for all $x \in \mathbb{R}$.

[21] Formally we define $T_n = 0$ if the (probability zero) event $V_n = 0$ holds (this can only hold if $X_i = X_1$ for all $i \in \{1, ..., n\}$).

C. Central limit theorem (CLT)

If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. random variables with finite mean $\mu \in \mathbb{R}$ and finite variance $\sigma^2 > 0$ then the *central limit theorem* (CLT) states:

$$\lim_{n\to\infty} P[G_n \leq x] = P[G \leq x] \quad \forall x \in \mathbb{R}$$

where

$$G_n = \frac{1}{\sqrt{n\sigma^2}} \sum_{i=1}^{n} (X_i - \mu) \quad \forall n \in \{1, 2, 3, \ldots\}$$

and G is any random variable with the $N(0, 1)$ distribution:

$$P[G \leq x] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad \forall x \in \mathbb{R}$$

In particular, the distribution of G_n converges to the distribution of G. For a proof, see, for example, [3].

Significant insight into the CLT can be gained from the following two facts:

1) Using basic properties of expectation and variance, it holds that $\mathbb{E}[G_n] = 0$ and $Var(G_n) = 1$ for all $n \in \{1, 2, 3, \ldots\}$.
2) If $\{X_i\}_{i=1}^{\infty}$ are i.i.d. $N(\mu, \sigma^2)$ then, using the fact that a linear combination of mutually independent Gaussian random variables is again Gaussian, it holds that $G_n \sim N(0, 1)$ for all $n \in \{1, 2, 3, \ldots\}$. Thus

$$P[G_n \leq x] = P[G \leq x] \quad \forall x \in \mathbb{R}, \forall n \in \{1, 2, 3, \ldots\}$$

meaning that the limiting behavior ensured by the CLT is trivially true in this special case when the i.i.d. random variables $\{X_i\}_{i=1}^{\infty}$ are already Gaussian.

Here are some common misunderstandings to be aware of:

- Caveat 1: There is no random variable X for which $G_n \to X$ in probability. It is only the *distribution* of G_n that converges to the $N(0, 1)$ distribution (this is called *convergence in distribution*). It can be shown that for any positive integer n, the random variable G_n is *almost independent* of G_k when k is much larger than n:

$$\lim_{k\to\infty} P[G_k \leq x | G_n] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \quad \forall x \in \mathbb{R}$$

Further, with probability 1, there will be an infinite number of indices n for which $G_n \geq 1$, and an infinite number of indices n for which $G_n \leq -1$. Thus, G_n almost surely oscillates without converging to anything.

- Caveat 2: The CLT considers $P[G_n \leq x]$ for a fixed tail threshold $x \in \mathbb{R}$. One should be skeptical when the CLT is invoked to study $P[G_n \leq x_n]$ when x_n itself changes with n, particularly when either $x_n \to \infty$ or $x_n \to -\infty$. For example, suppose $\{X_i\}_{i=1}^{\infty}$ are i.i.d. Unif$[0, 1]$. Then $\mathbb{E}[X_i] = 1/2$ and $Var(X_i) = 1/12$ for all i and

$$0 = P[\textstyle\sum_{i=1}^{n} X_i \leq 0] = P[\sum_{i=1}^{n} (X_i - 1/2) \leq -\frac{n}{2}] \quad \forall n \in \{1, 2, 3, \ldots\}$$

Dividing by $\sqrt{n/12}$ gives

$$P[G_n \leq \tfrac{-\sqrt{12n}}{2}] = 0 \quad \forall n \in \{1, 2, 3, \ldots\}$$

Therefore, even for large n, *we cannot accurately approximate* $P[G_n \leq \frac{-\sqrt{12n}}{2}]$ *by* $P[G \leq \frac{-\sqrt{12n}}{2}]$. While both converge to 0 as $n \to \infty$, one is *always zero* while the other converges to zero at an asymptotic rate determined by the Gaussian tail. The two asymptotic rates of converging to zero have nothing in common.

VI. Additional topics

Other important topics include covariance; convolution of PDFs; joint PDFs and PDF transformations; characteristic functions $\mathbb{E}\left[e^{itX}\right]$; minimizing mean square error; MAP detection; ML detection; properties of Gaussian random variables; central limit theorem approximations; discrete time Markov chains (see [1][2]).

APPENDIX A – EXPECTATION

Fix a probability space $(\Omega, \mathcal{F}, P)$.

A. Indicator random variable

Let A be an event. Define the binary-valued random variable 1_A by

$$1_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{else} \end{cases}$$

In particular, $1_A(\omega) = 1$ if and only if $\omega \in A$. The random variable 1_A is called an *indicator random variable* or *indicator function*.

B. Expectation

If X is a random variable, the *expectation of* X, written $\mathbb{E}[X]$, is a value that satisfies

$$\mathbb{E}[X] \in \mathbb{R} \cup \{\infty\} \cup \{-\infty\} \cup \{DNE\}$$

where DNE represents *does not exist*. We say the expectation is finite if $\mathbb{E}[X] \in \mathbb{R}$. The expectation can be defined using measure theory and the Lebesgue integral. The expectation $\mathbb{E}[X]$ depends only on the CDF of X, so all random variables with the same CDF have the same expectation. It can be shown that if X is a nonnegative random variable, meaning that $X(\omega) \geq 0$ for all $\omega \in \Omega$, then the expectation always exists and $\mathbb{E}[X] \in [0, \infty) \cup \{\infty\}$. It can further be shown that if X is *any* random variable (possibly one that can take negative values), then $\mathbb{E}[X] \in \mathbb{R}$ if and only if $\mathbb{E}[|X|] < \infty$.

Let a, b be real numbers, let A be an event, and let X, Y be random variables that have finite expectations. The expectation has the following important properties:

1) Expectation of indicator:
$$\mathbb{E}[1_A] = P[A]$$

2) Linearity of expectation:
$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

In particular, linearity of expectation says that if X and Y have finite expectations then $aX + bY$ also has a finite expectation, and this expectation is given by the above formula. By induction, linearity of expectation can be used to prove that if n is a positive integer, $a_1, \ldots, a_n$ are real numbers, and $X_1, \ldots, X_n$ are random variables with finite expectations, then

$$\mathbb{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i]$$

This can be used, together with the indicator function property, to prove a useful formula for random variables that can take at most finitely many values. Fix n as a positive integer and let $\{x_1, \ldots, x_n\}$ be a set of n distinct real numbers. Let $X : \Omega \to \{x_1, ..., x_n\}$ be a random variable. Observe that for each $i \in \{1, ..., n\}$ we have $\{X = x_i\}$ is an event and

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

Taking expectations of the above equality (and using linearity of expectation and the indicator function property) yields

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i \mathbb{E}\left[1_{\{X=x_i\}}\right] = \sum_{i=1}^{n} x_i P[X = x_i]$$

This formula can be generalized in certain special cases when $\mathbb{E}[|X|] < \infty$:

- If X is a *discrete random variable*, meaning $X : \Omega \to D$ where $D \subseteq \mathbb{R}$ is a nonempty set that is either finite or countably infinite, and if $\mathbb{E}[|X|] < \infty$, it can be shown that
$$\mathbb{E}[X] = \sum_{x \in D} x P[X = x]$$
- If X is a *continuous random variable* with a valid *probability density function* (PDF), and if $\mathbb{E}[|X|] < \infty$, then it can be shown
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

It can be shown that if X is any nonnegative random variable then:

$$\mathbb{E}[X] = \int_0^\infty P[X > x]dx$$

with the understanding that $\mathbb{E}[X] = \infty$ if and only if the right-hand-side of the above equality is ∞. A formal derivation of this uses the Lebesgue integral and the *Fubini-Tonelli theorem* of measure theory:

$$\left(X = \int_0^\infty 1_{\{X>x\}}dx\right) \implies \left(\mathbb{E}[X] = \int_0^\infty \mathbb{E}\left[1_{\{X>x\}}\right]dx\right)$$

C. Variance

If X is a random variable with finite expectation $\mathbb{E}[X]$ we define the *variance of* X, written $Var(X)$, by

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

It can be shown that

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

The value $\mathbb{E}\left[X^2\right]$ is called the *second moment* of X. It can be shown that a random variable X has a finite second moment if and only if $\mathbb{E}[X]$ and $Var(X)$ are both finite. If X, Y are two random variables with finite second moments then

$$Var(X+Y) = Var(X) + Var(Y) - 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

We say that X and Y are *uncorrelated* if

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0$$

It holds that X and Y are uncorrelated if and only if

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

It can be shown that if X and Y are independent with finite expectations, then they must be uncorrelated.

If X is a random variable with finite second moment, and if $a, b \in \mathbb{R}$, then

$$Var(aX + b) = a^2 Var(X)$$

If $\{X_1, ..., X_n\}$ have finite second moments and are *pairwise uncorrelated* (meaning $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$ for all $i \neq j$) then

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i)$$

D. Law of total expectation

If X is a random variable with $\mathbb{E}[|X|] < \infty$ and if B is an event with $P[B] > 0$ then we define the *conditional expectation of* X *given* B by

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X1_B]}{P[B]}$$

It can be shown that this is the same as the expectation of a random variable Y that has distribution equal to the conditional distribution of X given B, that is

$$(P[Y \leq y] = P[X \leq y|B] \quad \forall y \in \mathbb{R}) \implies (\mathbb{E}[Y] = \mathbb{E}[X|B])$$

If X is a random variable with $\mathbb{E}[|X|] < \infty$ and if $\{B_i\}_{i=1}^\infty$ are events that partition the sample space then

$$X = \sum_{i=1}^\infty X1_{B_i}$$

and it can be shown that:[22]

$$\boxed{\mathbb{E}[X] = \sum_{i=1}^\infty \mathbb{E}[X1_{B_i}] = \sum_{i:P[B_i]>0} \mathbb{E}[X|B_i]P[B_i]}$$

This is called the *law of total expectation*.[23]

[22]The thing to be careful of here is whether or not we can push the expectation through the *infinite sum*. The Fubini-Tonelli theorem of measure theory ensures that $\mathbb{E}\left[\sum_{i=1}^\infty X_i\right] = \sum_{i=1}^\infty \mathbb{E}[X_i]$ whenever either X_i are nonnegative random variables for all i, or when $\sum_{i=1}^\infty \mathbb{E}[|X_i|] < \infty$. In the formula above we have $|X| = \sum_{i=1}^\infty |X1_{B_i}|$ and so $\mathbb{E}[|X|] < \infty$ implies $\sum_{i=1}^\infty \mathbb{E}\left[|X1_{B_i}|\right] < \infty$, which allows pushing the expectation through the infinite sum.

[23]The equality implicitly uses $P[B_i] = 0 \implies \mathbb{E}\left[X1_{B_i}\right] = 0$. This holds because two random variables with the same CDF must have the same expectation: If a random variable is almost surely zero then its CDF is the same as that of the always-zero random variable 1_ϕ, so its expectation must be 0.

E. Iterated expectations

A result that is closely related to the law of total expectation is the *law of iterated expectations*, which holds for any random variables X, Y with $\mathbb{E}[|X|] < \infty$. The law of iterated expectations is $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. When Y has a PDF then this can be written in integral form:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] f_Y(y) dy$$

The above formula is also called the *law of total expectation in integral form*. The corresponding *law of total probability in integral form* for an event A can be obtained by viewing $P[A] = \mathbb{E}[1_A]$:

$$P[A] = \int_{-\infty}^{\infty} P[A|Y=y] f_Y(y) dy$$

A useful scenario for the $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ notation is when we have a sum of a random number of random variables. Fix $(\Omega, \mathcal{F}, P)$. Let $Y : \Omega \to \mathbb{R}$ be a random variable that takes values in the set of nonnegative integers, so $Y \in \{0, 1, 2, ...\}$ surely. For each $i \in \{1, 2, 3, \ldots\}$ let $V_i : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}[V_i] = 31$. Define the random variable $X : \Omega \to \mathbb{R}$ by

$$X = \sum_{i=1}^{Y} V_i \tag{35}$$

with the understanding that $X = 0$ if $Y = 0$. Then X is the sum of a random number of random variables. We have

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{Y} V_i | Y\right]\right] = \mathbb{E}\left[\sum_{i=1}^{Y} \mathbb{E}[V_i|Y]\right] \tag{36}$$

We cannot further simplify the expression on the right-hand-side of (36) without more assumptions that allow us to compute $\mathbb{E}[V_i|Y]$. Suppose that for all $i \in \{1, 2, 3, \ldots\}$ we have the additional information that V_i is independent of Y. In this special case we have

$$\mathbb{E}[V_i|Y] = \mathbb{E}[V_i] = 31 \quad \forall i \in \{1, 2, 3, ...\}$$

and so

$$\sum_{i=1}^{Y} \mathbb{E}[V_i|Y] = \sum_{i=1}^{Y} 31 = 31Y$$

and so from (36)

$$\mathbb{E}[X] = 31\mathbb{E}[Y]$$

A reader may naively guess $\mathbb{E}[X] = 31\mathbb{E}[Y]$ by looking at the definition of X in (35). However, the guess may be wrong without the additional information that V_i is independent of Y for all i. For example, suppose $Y \in \{1, 2\}$ surely and $P[Y = 1] = P[Y = 2] = 1/2$, while for each $i \in \{1, 2, 3, ...\}$ we have

$$V_i = V_1 = \begin{cases} 0 & \text{if } Y = 1 \\ 62 & \text{if } Y = 2 \end{cases}$$

Here $\mathbb{E}[V_i] = 31$ for all $i \in \{1, 2, 3, ...\}$, but there is a clear dependence between Y and V_i. We have $\mathbb{E}[Y] = 1.5$ so

$$31\mathbb{E}[Y] = 46.5$$

However

$$\mathbb{E}[V_i|Y] = 62 \cdot 1_{\{Y=2\}} \quad \forall i \in \{1, 2, 3, ...\}$$

where "$\cdot$" indicates multiplication of 62 and $1_{\{Y=2\}}$. So

$$\sum_{i=1}^{Y} \mathbb{E}[V_i|Y] = \sum_{i=1}^{Y} 62 \cdot 1_{\{Y=2\}} = 124 \cdot 1_{\{Y=2\}}$$

and so from (36)

$$\mathbb{E}[X] = 124\mathbb{E}\left[1_{\{Y=2\}}\right] = 62 \neq 31\mathbb{E}[Y]$$

APPENDIX B – DETAILS ON ALMOST SURE CONVERGENCE

This appendix contains advanced material that may be of interest. Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of random variables. Let Y be another random variable.

Theorem 4: $Y_n \to Y$ almost surely if and only if for all $\epsilon > 0$ we have

$$\lim_{n\to\infty} P[\cup_{i=n}^{\infty}\{|Y_i - Y| \geq \epsilon\}] = 0 \tag{37}$$

Proof: See Appendix C. □

A. Comparison to convergence in probability

Theorem 4 helps to understand the relationship between *convergence almost surely* and *convergence in probability*. We say that $Y_n \to Y$ *in probability* if for all $\epsilon > 0$ we have

$$\lim_{n\to\infty} P[|Y_n - Y| \geq \epsilon] = 0$$

Comparing this condition to (37), it is apparent that (37) is more stringent because for all $n \in \{1, 2, 3, \ldots\}$ and all $\epsilon > 0$ we have

$$\{|Y_n - Y| \geq \epsilon\} \subseteq \cup_{i=n}^{\infty}\{|Y_i - Y| \geq \epsilon\}$$

and so

$$P[|Y_n - Y| \geq \epsilon] \leq P[\cup_{i=n}^{\infty}\{|Y_i - Y| \geq \epsilon\}]$$

Taking a limit of both sides as $n \to \infty$ proves that

$$\text{Convergence almost surely} \implies \text{Convergence in probability}$$

B. Sufficient condition for almost sure convergence

Let $\{Y_i\}_{i=1}^{\infty}$ be a sequence of random variables. Fix $c \in \mathbb{R}$.

Lemma 1: If there is an $\alpha > 0$ such that

$$\sum_{i=1}^{\infty} \mathbb{E}\left[|Y_i - c|^{\alpha}\right] < \infty$$

then $Y_n \to c$ almost surely.

Proof: Fix $\epsilon > 0$. By Theorem 4 it suffices to show

$$\lim_{n\to\infty} P[\cup_{i=n}^{\infty}\{|Y_i - c| \geq \epsilon\}] = 0$$

We have by the union bound

$$\begin{aligned} P[\cup_{i=n}^{\infty}\{|Y_i - c| \geq \epsilon\}] &\leq \sum_{i=n}^{\infty} P[|Y_i - c| \geq \epsilon] \\ &= \sum_{i=n}^{\infty} P[|Y_i - c|^{\alpha} \geq \epsilon^{\alpha}] \\ &\leq \sum_{i=n}^{\infty} \frac{\mathbb{E}\left[|Y_i - c|^{\alpha}\right]}{\epsilon^{\alpha}} \end{aligned}$$

where the final inequality holds by the Markov inequality. Taking a limit as $n \to \infty$ gives

$$\lim_{n\to\infty} P[\cup_{i=n}^{\infty}\{|Y_i - c| \geq \epsilon\}] \leq \frac{1}{\epsilon^{\alpha}}\left(\lim_{n\to\infty}\sum_{i=n}^{\infty}\mathbb{E}\left[|Y_i - c|^{\alpha}\right]\right) \tag{38}$$

Since $\sum_{i=1}^{\infty}\mathbb{E}\left[|Y_i - c|^{\alpha}\right] < \infty$, the limit of its tail sum is zero:

$$\lim_{n\to\infty}\sum_{i=n}^{\infty}\mathbb{E}\left[|Y_i - c|^{\alpha}\right] = 0$$

Substituting this limit into (38) gives the result. □

C. Proof of law of large numbers for finite variance case

Let $\{X_i\}_{i=1}^{\infty}$ be i.i.d. random variables with finite mean μ and variance σ^2. For each $n \in \{1, 2, 3, \ldots\}$ define

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

It follows by linearity of expectation and basic properties of the variance for a linear combination of independent random variables that

$$\mathbb{E}[M_n] = \mu \quad \forall n \in \{1, 2, 3, \ldots\}$$
$$Var(M_n) = \mathbb{E}\left[(M_n - \mu)^2\right] = \frac{\sigma^2}{n} \quad \forall n \in \{1, 2, 3, \ldots\}$$

We first show almost sure convergence to μ when we sample only over indices that are perfect squares, that is, we consider $M_1, M_4, M_9, M_{16}, M_{25}$ and so on.

Lemma 2: $M_{n^2} \to \mu$ almost surely.

Proof: We use Lemma 1 with $\alpha = 2$:

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{E}\left[(M_{n^2} - \mu)^2\right] &= \sum_{n=1}^{\infty} Var(M_{n^2}) \\ &= \sum_{n=1}^{\infty} \frac{\sigma^2}{n^2} \\ &< \infty\end{aligned}$$

□

We now use the fact that we have almost sure convergence over perfect squares to get almost sure convergence.

Lemma 3: $M_n \to \mu$ almost surely.

Proof: First suppose that, in addition to $\{X_i\}_{i=1}^{\infty}$ being i.i.d. with finite mean and variance, we have that all X_i are surely nonnegative (so $X_i(\omega) \geq 0$ for all $\omega \in \Omega$ and all $i \in \{1, 2, 3, \ldots\}$). We prove the result in this special nonnegative case, then show that this special case can be used to prove the general case.

For each positive integer n, define k_n as the largest integer such that $k_n^2 \leq n$. Then

$$k_n^2 \leq n < (k_n + 1)^2 = k_n^2 + 2k_n + 1$$

It is clear that $\lim_{n\to\infty} k_n = \infty$. It is not difficult to show that

$$\lim_{n\to\infty} \frac{k_n^2}{n} = \lim_{n\to\infty} \frac{(k_n+1)^2}{n} = 1 \tag{39}$$

Since $X_i \geq 0$ surely for all i, we have for each positive integer n that

$$\sum_{i=1}^{k_n^2} X_i \leq \sum_{i=1}^{n} X_i \leq \sum_{i=1}^{(k_n+1)^2} X_i$$

Dividing both sides by n gives

$$\left(\frac{k_n^2}{n}\right)\left(\frac{1}{k_n^2}\sum_{i=1}^{k_n^2} X_i\right) \leq M_n \leq \left(\frac{(k_n+1)^2}{n}\right)\left(\frac{1}{(k_n+1)^2}\sum_{i=1}^{(k_n+1)^2} X_i\right)$$

That is

$$\left(\frac{k_n^2}{n}\right) M_{k_n^2} \leq M_n \leq \left(\frac{(k_n+1)^2}{n}\right) M_{(k_n+1)^2} \tag{40}$$

We already know from Lemma 2 that

$$M_{k_n^2} \to \mu \quad \text{almost surely}$$
$$M_{(k_n+1)^2} \to \mu \quad \text{almost surely}$$

Taking a limit of (40) and using these results and (39) gives

$$\mu \leq \lim_{n\to\infty} M_n \leq \mu \quad \text{almost surely}$$

This proves the desired result for the special case when X_i are surely nonnegative.

For the general case when X_i can be negative, break X_i into its positive and negative parts:

$$X_i = X_i^+ - X_i^- \quad \forall i \in \{1, 2, 3, \ldots\}$$

where $X_i^+ = \max\{X_i, 0\}$ and $X_i^- = \max\{-X_i, 0\}$. Note that since $\{X_i\}_{i=1}^\infty$ are i.i.d., it holds that the nonnegative random variables $\{X_i^+\}_{i=1}^\infty$ are i.i.d., as are the nonnegative random variables $\{X_i^-\}_{i=1}^\infty$. It also holds that X_i^+ has a finite mean and variance, as does X_i^-. Define $X = X_1$ and note that

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

Our proven result for nonnegative i.i.d. random variables with finite mean and variance implies:

$$\frac{1}{n}\sum_{i=1}^n X_i^+ \to \mathbb{E}\left[X^+\right] \quad \text{almost surely}$$
$$\frac{1}{n}\sum_{i=1}^n X_i^- \to \mathbb{E}\left[X^-\right] \quad \text{almost surely}$$

It follows that we almost surely have:

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n X_i &= \left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n X_i^+\right) - \left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n X_i^-\right)\\ &= \mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right]\\ &= \mathbb{E}[X]\end{aligned}$$

□

D. Convergence in probability does not imply convergence almost surely

Fix $(\Omega, \mathcal{F}, P)$. Let $\{Y_i\}_{i=1}^\infty$ be mutually independent random variables. For each $i \in \{1, 2, 3, ...\}$, assume $Y_i \sim Bern(1/i)$. In particular, $P[Y_i = 1] = 1/i$ and $P[Y_i = 0] = 1 - 1/i$ for all $i \in \{1, 2, 3, \ldots\}$. It follows that $Y_i \to 0$ in probability.

Now define A as the set of all $\omega \in \Omega$ for which $Y_i(\omega) = 1$ for an infinite number of indices $i \in \{1, 2, 3, ...\}$. For any $\omega \in A$, it is clear that the sequence of real numbers $\{Y_i(\omega)\}_{i=1}^\infty$ does not converge to 0. It can be shown that $A \in \mathcal{F}$ (see next paragraph). Since $\{Y_i\}_{i=1}^\infty$ are mutually independent and

$$\sum_{i=1}^\infty P[Y_i = 1] = \sum_{i=1}^\infty 1/i = \infty$$

the *Borel-Cantelli lemma* implies that $P[A] = 1$ (see Section VI-E). Therefore, even though $Y_i \to 0$ in probability, Y_i does not converge to 0 almost surely. Quite the opposite: Almost surely, Y_i does *not* converge to 0.

We can show $A \in \mathcal{F}$ as follows: Since for any positive integer n we know that $\sum_{i=1}^n Y_i$ is a random variable, we have

$$\left\{\sum_{i=1}^n Y_i \geq m\right\} \in \mathcal{F} \quad \forall n, m \in \{1, 2, 3, ...\}$$

Since the countable union of events is also an event, we have for all positive integers m:

$$\cup_{n=1}^\infty \left\{\sum_{i=1}^n Y_i \geq m\right\} \in \mathcal{F}$$

Since the countable intersection of events is also an event, we have

$$\cap_{m=1}^\infty \cup_{n=1}^\infty \left\{\sum_{i=1}^n Y_i \geq m\right\} \in \mathcal{F}$$

Now observe

$$A = \left\{\sum_{i=1}^\infty Y_i = \infty\right\} = \cap_{m=1}^\infty \cup_{n=1}^\infty \left\{\sum_{i=1}^n Y_i \geq m\right\}$$

Therefore, $A \in \mathcal{F}$.

E. Predicting the future with Borel-Cantelli

Fix a probability space $(\Omega, \mathcal{F}, P)$. Let $\{A_i\}_{i=1}^{\infty}$ be a sequence of (possibly dependent) events. We can imagine observing the probability experiment being carried out over a sequence of time steps $i \in \{1, 2, 3, \ldots\}$. At each step i we witness whether or not event A_i occurs. We want to know if *infinitely many* of the events will occur, or if only *finitely many* will occur. For example, we might have an infinite sequence of (possibly dependent and biased) coin flips, and we define A_i as the event that flip i is HEADS. Will there be infinitely may HEADS, or only finitely many? We first suspect that it is impossible to know because nobody can predict the future. Or can they? Enter Borel-Cantelli.

First recall that the outcome $\omega \in \Omega$ determines the occurrence or non-occurrence of each of the events A_1, A_2, A_3, and so on. In particular, if we imagine an experiment where nature randomly chooses an outcome $\omega \in \Omega$ according to the probability measure P, we say that *event A_i occurs* if and only if $\omega \in A_i$. Equivalently, we say event A_i is *true* if and only if that event occurs. We first define the subset of all outcomes of the sample space that are in infinitely many of the sets A_i. Every outcome in this subset ensures the events occur *infinitely often* (i.o.). The complement of this is the set of all outcomes that are in only finitely many of the sets A_i. This corresponds to the events occurring only *finitely often* (f.o.). In particular, define

$$\{A_i \quad \text{i.o.}\} = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i$$
$$\{A_i \quad \text{f.o.}\} = \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i^c$$

The set $\{A_i \quad \text{i.o.}\} \subseteq \Omega$ includes an outcome $\omega \in \Omega$ if and only if for all positive integers n we have $\omega \in \cup_{i=n}^{\infty} A_i$. This holds if and only if $\omega \in A_i$ for infinitely many indices $i \in \{1, 2, 3, \ldots\}$. With these definitions it holds by De Morgan's law:

$$\{A_i \quad \text{f.o.}\} = \{A_i \quad \text{i.o.}\}^c$$

The set $\{A_i \quad \text{i.o.}\}$ is defined by countable unions and intersections of events, so it is itself an event. Indeed, for each positive integer n, the set $\cup_{i=n}^{\infty} A_i$ is a countable union of events and so is itself an event. Thus, any countable intersection of these events is again an event. Since $\{A_i \quad \text{i.o.}\}$ is an event, its complement $\{A_i \quad \text{f.o.}\}$ is also an event and

$$P[A_i \quad \text{i.o.}] + P[A_i \quad \text{f.o.}] = 1$$

Fact: If $\{x_i\}_{i=1}^{\infty}$ is a sequence of nonnegative real numbers then $\sum_{i=1}^{\infty} x_i$ always exists (possibly being ∞) and

$$\sum_{i=1}^{\infty} x_i < \infty \iff \lim_{n\to\infty} \sum_{i=n}^{\infty} x_i = 0 \tag{41}$$

Lemma 4: (Borel-Cantelli) Fix $(\Omega, \mathcal{F}, P)$. Let $\{A_i\}_{i=1}^{\infty}$ be a sequence of events.
a) If $\sum_{i=1}^{\infty} P[A_i] < 1$ then $P[A_i \quad \text{i.o.}] = 0$.
b) If $\sum_{i=1}^{\infty} P[A_i] = \infty$, and if $\{A_i\}_{i=1}^{\infty}$ are mutually independent, then $P[A_i \quad \text{i.o.}] = 1$.
Proof: To prove (a), suppose $\sum_{i=1}^{\infty} P[A_i] < \infty$. By (41) it holds that

$$\lim_{n\to\infty} \sum_{i=n}^{\infty} P[A_i] = 0 \tag{42}$$

Now fix n as a positive integer. By definition of $\{A_i \quad \text{i.o.}\}$ we have

$$\{A_i \quad \text{i.o.}\} \subseteq \cup_{i=n}^{\infty} A_i$$

and so the subset bound implies

$$P[A_i \quad \text{i.o.}] \leq P[\cup_{i=n}^{\infty} A_i] \leq \sum_{i=n}^{\infty} P[A_i]$$

where the final inequality is the union bound. Taking $n \to \infty$ and using (42) gives

$$P[A_i \quad \text{i.o.}] \leq 0$$

which concludes the proof of part (a).

To prove part (b), assume mutual independence and $\sum_{i=1}^{\infty} P[A_i] = \infty$. For each $n \in \{1, 2, 3, \ldots\}$ we have

$$P[\cap_{i=n}^{\infty} A_i^c] = \prod_{i=n}^{\infty} P[A_i^c] = \prod_{i=n}^{\infty} (1 - P[A_i]) = 0 \tag{43}$$

where the final equality can be established by considering the case $P[A_i] = 1$ for some $i \geq n$ (so the result is trivial) and its opposite case $P[A_i] < 1$ for all $i \geq n$ (take $\log(\cdot)$ and use $\log(1-x) \leq -x$ for all $x < 1$). By definition of $\{A_i \quad \text{f.o.}\}$

$$P[A_i \quad \text{f.o.}] = P[\cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i^c] \leq \sum_{n=1}^{\infty} P[\cap_{i=n}^{\infty} A_i^c] = \sum_{n=1}^{\infty} 0 = 0$$

where the inequality holds by the union bound; the subsequent equality holds by (43). □

As an example, if $P[A_i] = 1/i^2$ for $i \in \{1, 2, 3, ...\}$ then (almost surely) only finitely many of the events will occur. On the other hand if $P[A_i] = 1/i$ *and if the events are mutually independent* then (almost surely) infinitely many will occur. If $P[A_i] = 1/i$ but the events are not necessarily mutually independent then the Borel-Cantelli lemma is inconclusive. Also, in this inconclusive case, examples can be constructed to make $P[A_i \quad \text{i.o.}]$ any desired value in $[0, 1]$.

APPENDIX C – PROOF OF THEOREM 4

This appendix contains advanced material that may be of interest. Note that $Z_n \to Z$ almost surely if and only if $(Z_n - Z) \to 0$ almost surely. In particular, it suffices to consider convergence to the constant 0.

Let $\mathbb{N} = \{1, 2, 3, ...\}$. Let $\{Y_n\}_{n=1}^{\infty}$ be random variables. Define

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = 0\right\}$$

Observe that

$$\boxed{(\omega \in A) \iff \text{"For all } k \in \mathbb{N}\text{, there is an } n \in \mathbb{N} \text{ such that } |Y_i(\omega)| < 1/k \text{ for all } i \geq n."}$$

Hence

$$\boxed{A = \cap_{k\in\mathbb{N}} \cup_{n\in\mathbb{N}} \cap_{i=n}^{\infty} \{|Y_i| < 1/k\}} \tag{44}$$

where

- $\cap_{k\in\mathbb{N}}$ means "For all $k \in \mathbb{N}$"
- $\cup_{n\in\mathbb{N}}$ means "there is an $n \in \mathbb{N}$"
- $\cap_{i=n}^{\infty}\{|Y_i| \leq 1/k\}$ means "$|Y_i| < 1/k$ for all $i \geq n$."

The representation (44) formally proves the set A is an event, that is, $A \in \mathcal{F}$. That is because $|Y_i|$ is a random variable (it is a continuous function of Y_i) and so $\{|Y_i| < 1/k\} \in \mathcal{F}$ for all $i, k \in \mathbb{N}$. Therefore $\cap_{i=n}^{\infty}\{|Y_i| < 1/k\}$ is a countable intersection of events and so $\cap_{i=n}^{\infty}\{|Y_i| < 1/k\} \in \mathcal{F}$ for all $n, k \in \mathbb{N}$. Since the countable union of events in $\mathcal{F}$ is also in $\mathcal{F}$, we obtain $\cup_{n\in\mathbb{N}} \cap_{i=n}^{\infty} \{|Y_i| < 1/k\} \in \mathcal{F}$ for all $k \in \mathbb{N}$, and the countable intersection of these events over all $k \in \mathbb{N}$ is also in $\mathcal{F}$.

Clearly $P[A] = 1$ if and only if $P[A^c] = 0$. By De Morgan's law:

$$A^c = \cup_{k\in\mathbb{N}} \cap_{n\in\mathbb{N}} \cup_{i=n}^{\infty}\{|Y_i| \geq 1/k\} \tag{45}$$

For $k, n \in \mathbb{N}$ define

$$B_{n,k} = \cup_{i=n}^{\infty}\{|Y_i| \geq 1/k\}$$

It holds that for each fixed $k \in \mathbb{N}$ we have

$$B_{n,k} \searrow \cap_{n=1}^{\infty} B_{n,k}$$

and by continuity of probability

$$\lim_{n\to\infty} P[B_{n,k}] = P[\cap_{n=1}^{\infty} B_{n,k}] \quad \forall k \in \mathbb{N} \tag{46}$$

We want to show the following statement (which is equivalent to the statement of Theorem 4):

$$\left(\lim_{n\to\infty} P[B_{n,k}] = 0 \quad \forall k \in \mathbb{N}\right) \iff P[A^c] = 0 \tag{47}$$

To prove the forward direction of (47), suppose for all $k \in \mathbb{N}$ we have $\lim_{n\to\infty} P[B_{n,k}] = 0$. Fix $k \in \mathbb{N}$. By (46) we have

$$P[\cap_{n=1}^{\infty} B_{n,k}] = 0$$

This holds for all $k \in \mathbb{N}$ and so by Theorem 3

$$P[\cup_{k=1}^{\infty} \cap_{n=1}^{\infty} B_{n,k}] = 0$$

That is, $P[A^c] = 0$. This proves the forward direction of (47).

To prove the reverse direction, suppose $P[A^c] = 0$. Fix $k \in \mathbb{N}$. By (45) we have

$$A^c \supseteq \cap_{n\in\mathbb{N}} \cup_{i=n}^{\infty} \{|Y_i| \geq 1/k\}$$

Since $P[A^c] = 0$, it follows that

$$P[\cap_{n\in\mathbb{N}} \cup_{i=n}^{\infty} \{|Y_i| \geq 1/k\}] = 0$$

That is

$$P[\cap_{n\in\mathbb{N}} B_{n,k}] = 0$$

which by (46) gives

$$\lim_{n\to\infty} P[B_{n,k}] = 0$$

This holds for all $k \in \mathbb{N}$, which proves the reverse direction of (47).

APPENDIX D – MEASURABLE SETS AND FUNCTIONS

This appendix contains advanced material that may be of interest. Let Ω be a nonempty set. It can be shown that the intersection of an arbitrary (possibly uncountably infinite) number of sigma algebras on Ω is again a sigma algebra on Ω. Let $\mathcal{C}$ be a collection of subsets of Ω. The *sigma algebra on Ω generated by sets in the collection $\mathcal{C}$*, denoted $\sigma(\mathcal{C})$, is defined as the intersection of all sigma algebras on Ω that contain all sets in $\mathcal{C}$. There is at least one sigma algebra on Ω that contains all sets in $\mathcal{C}$, namely, $Pow(\Omega)$. We can view $\sigma(\mathcal{C})$ as the "smallest" sigma algebra on Ω that contains all sets in $\mathcal{C}$. Note that $\mathcal{C} \subseteq \sigma(\mathcal{C}) \subseteq Pow(\Omega)$. Also, $\sigma(\mathcal{C}) = \mathcal{C}$ if and only if $\mathcal{C}$ is already a sigma algebra on Ω. If $\mathcal{C}_1$ and $\mathcal{C}_2$ are two collections of subsets of Ω then $\mathcal{C}_1 \subseteq \mathcal{C}_2 \implies \sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2)$.

A. The sigma algebra $\mathcal{B}(\mathbb{R})$

With $\mathbb{R}$ being the set of real numbers, define $\mathcal{B}(\mathbb{R})$, called the *standard Borel sigma algebra on* $\mathbb{R}$, as the sigma algebra generated by all intervals of $\mathbb{R}$ of the form $(-\infty, x]$ for $x \in \mathbb{R}$. It can be shown that $\mathcal{B}(\mathbb{R})$ includes all finite or countably infinite subsets of $\mathbb{R}$, all intervals of $\mathbb{R}$, all open subsets of $\mathbb{R}$, all closed subsets of $\mathbb{R}$, and more.[24] Sets in $\mathcal{B}(\mathbb{R})$ are called *Borel measurable subsets of* $\mathbb{R}$ and can be viewed as those subsets of $\mathbb{R}$ which have a well defined *length*.[25] The *Borel measure* $\mu : \mathcal{B}(\mathbb{R}) \to [0, \infty) \cup \{\infty\}$ maps every set $A \in \mathcal{B}(\mathbb{R})$ to its length $\mu(A)$. The length can be 0 or ∞. For example

$$\begin{aligned}
&\mu([-1.5, 1.5]) = \mu((-1.5, 1.5)) = 3\\
&\mu([0, 1] \cup [2, 3.5]) = 2.5\\
&\mu([0, \infty)) = \infty\\
&\mu(\{0.3, 0.9, 0.987\}) = 0\\
&\mu(\mathbb{N}) = \mu(\mathbb{Q}) = 0
\end{aligned}$$

B. The sigma algebra $\mathcal{B}(\mathbb{R}^n)$

If n is a positive integer we define $\mathcal{B}(\mathbb{R}^n)$ as the sigma algebra generated by all subsets of $\mathbb{R}^n$ of the form

$$(-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n] \tag{48}$$

for some $(x_1, ..., x_n) \in \mathbb{R}^n$. It can be shown that $\mathcal{B}(\mathbb{R}^n)$ is equal to the sigma algebra generated by all open subsets of $\mathbb{R}^n$ (see footnote 24 for a proof with $n = 1$). As before, $\mathcal{B}(\mathbb{R}^n)$ contains all finite or countably infinite subsets of $\mathbb{R}^n$, all open or closed subsets of $\mathbb{R}^n$, and more. Sets in $\mathcal{B}(\mathbb{R}^n)$ can be viewed as those which have a well defined multidimensional volume. The Borel measure $\mu : \mathcal{B}(\mathbb{R}^n) \to [0, \infty) \cup \{\infty\}$ maps every set $A \in \mathcal{B}(\mathbb{R}^n)$ to its volume $\mu(A)$.

C. Measurable functions

A *measurable space* is a pair $(\Omega, \mathcal{F})$ such that Ω is a nonempty set and $\mathcal{F}$ is a sigma algebra on Ω. Let $(\Omega, \mathcal{F})$ and $(V, \mathcal{G})$ be two measurable spaces. A function $f : \Omega \to V$ is said to be *measurable with respect to input space* $(\Omega, \mathcal{F})$ *and output space* $(V, \mathcal{G})$ if

$$f^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{G}$$

where $f^{-1}(B)$ denotes the inverse image and is defined

$$f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\}$$

For ease of notation, a function $f : \Omega \to V$ is simply said to be *measurable* when the input and output measurable spaces are clear. It is easy to see that all constant functions $f : \Omega \to V$ are measurable. That is, if there is a $v \in V$ such that $f(\omega) = v$ for all $\omega \in \Omega$, then f is measurable.

If the input or output set is $\mathbb{R}^n$ for some positive integer n, then, unless otherwise stated, the corresponding measurable space is assumed to be $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. If $(\Omega, \mathcal{F})$ is a measurable space then a measurable function $f : \Omega \to \mathbb{R}^n$ is often called a *Borel measurable function* to emphasize the output space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

[24] Define $\mathcal{C}_1$ as the collection of all intervals $(-\infty, x]$ for $x \in \mathbb{R}$, so $\sigma(\mathcal{C}_1) = \mathcal{B}(\mathbb{R})$. Define $\mathcal{C}_2$ as the collection of all open subsets of $\mathbb{R}$. We want to show $\sigma(\mathcal{C}_1) = \sigma(\mathcal{C}_2)$. First show $\mathcal{C}_1 \subseteq \sigma(\mathcal{C}_2)$ to conclude $\sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2)$. For the opposite inclusion, define $\mathcal{C}_3$ as the set of all open intervals of the form $(x - \epsilon, x + \epsilon)$ for some rational number x and some rational $\epsilon > 0$. Observe that $\mathcal{C}_3$ is a countable set ($\mathbb{Q}^2$ is countable) and show $\mathcal{C}_2 \subseteq \sigma(\mathcal{C}_3) \subseteq \sigma(\mathcal{C}_1)$.

[25] See related definition of *Lebesgue measurable sets* at the end of this section.

D. Countable cutting and pasting

Let $(\Omega, \mathcal{F})$ and $(V, \mathcal{G})$ be two measurable spaces. Let $\{A_i\}_{i=1}^{\infty}$ be sets in $\mathcal{F}$ that form a countable partition of Ω, so that

- $A_i \in \mathcal{F} \quad \forall i \in \{1, 2, 3, \ldots\}$
- $A_i \cap A_j = \phi$ for $i \neq j$
- $\cup_{i=1}^{\infty} A_i = \Omega$

For each $i \in \{1, 2, 3, \ldots\}$, let $f_i : \Omega \to V$ be a measurable function. Define the function $g : \Omega \to V$ by

$$g(\omega) = \begin{cases} f_1(\omega) & \text{if } \omega \in A_1 \\ f_2(\omega) & \text{if } \omega \in A_2 \\ f_3(\omega) & \text{if } \omega \in A_3 \\ \cdots \end{cases}$$

so that $g(\omega) = f_i(\omega)$ whenever $\omega \in A_i$.

Lemma 5: The function $g : \Omega \to V$ defined above is measurable.

Proof: Fix $B \in \mathcal{G}$. We want to show $g^{-1}(B) \in \mathcal{F}$. We have

$$\begin{aligned} g^{-1}(B) &= \cup_{i=1}^{\infty}(A_i \cap g^{-1}(B)) \\ &= \cup_{i=1}^{\infty}(A_i \cap f_i^{-1}(B)) \end{aligned} \tag{49}$$

For each i, $f_i : \Omega \to V$ is measurable and so $f_i^{-1}(B) \in \mathcal{F}$. Therefore, the right-hand-side of (49) is a countable union of sets in the sigma algebra $\mathcal{F}$ and so it is in $\mathcal{F}$. □

We allow $A_i = \phi$ for some indices i, so a finite partition of Ω of the form $\{A_i\}_{i=1}^{n}$ can be viewed in the framework of the above lemma using $A_i = \phi$ for $i \geq n + 1$.

E. Compositions of measurable functions

Lemma 6: Let $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$, $(\Omega_3, \mathcal{F}_3)$ be three measurable spaces. Let $f : \Omega_1 \to \Omega_2$ and $g : \Omega_2 \to \Omega_3$ be measurable functions. Then the function $h : \Omega_1 \to \Omega_3$ defined by $h(\omega) = g(f(\omega))$ for all $\omega \in \Omega_1$ is measurable.

Proof: Fix $B \in \mathcal{F}_3$. It holds that

$$h^{-1}(B) = f^{-1}(g^{-1}(B))$$

We know $g^{-1}(B) \in \mathcal{F}_2$ by measurability of g. Thus, by measurability of f we have $f^{-1}(g^{-1}(B)) \in \mathcal{F}_1$. □

F. Checking inverse images over an arbitrary collection of sets

Lemma 7: Let $(\Omega, \mathcal{F})$ be a measurable space. Let V be a nonempty set and let $f : \Omega \to V$ be any function. Let $\mathcal{C}$ be any nonempty collection of subsets of V. Then

$$\left(f^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{C}\right) \implies \left(f^{-1}(B) \in \mathcal{F} \quad \forall B \in \sigma(\mathcal{C})\right)$$

Proof: Define $\mathcal{D}$ as the set of all $B \subseteq V$ for which $f^{-1}(B) \in \mathcal{F}$. It can be shown that

- $\mathcal{C} \subseteq \mathcal{D}$
- $\mathcal{D}$ is a sigma algebra on V

Since $\mathcal{C} \subseteq \mathcal{D}$ we obtain

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

where the final equality holds because $\mathcal{D}$ is already a sigma algebra on V. □

Lemma 7 means we can test measurability of $f : \Omega \to V$ with respect to input and output spaces $(\Omega, \mathcal{F})$ and $(V, \mathcal{G})$ by only testing whether or not $f^{-1}(B) \in \mathcal{F}$ for those sets $B \in \mathcal{G}$ within some collection $\mathcal{C}$ such that $\sigma(\mathcal{C}) = \mathcal{G}$. This is useful for understanding the definition of a random variable given in (11). Suppose $X : \Omega \to \mathbb{R}$ is a random variable. By definition, X satisfies the requirement (11), which is equivalent to

$$X^{-1}((-\infty, x]) \in \mathcal{F} \quad \forall x \in \mathbb{R} \tag{50}$$

Since the intervals $(-\infty, x]$ generate $\mathcal{B}(\mathbb{R})$, Lemma 7 implies that

$$X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{B}(\mathbb{R}) \tag{51}$$

Thus, (50) implies (51). On the other hand, it is clear that (51) implies (50). Thus, a function $X : \Omega \to \mathbb{R}$ is a random variable if and only if it is a Borel measurable function.

G. Borel measurable functions of many variables

Lemma 8: Let $(\Omega, \mathcal{F})$ be a measurable space. Fix $n, k \in \mathbb{N}$. Let $X_1 : \Omega \to \mathbb{R}, \ldots, X_n : \Omega \to \mathbb{R}$ be Borel measurable functions (random variables). Let $f : \mathbb{R}^n \to \mathbb{R}^k$ be Borel measurable. Then

a) The function $h : \Omega \to \mathbb{R}^n$ given by $h(\omega) = (X_1(\omega), \ldots, X_n(\omega))$ is Borel measurable.

b) The function $g : \Omega \to \mathbb{R}^k$ given by $g(\omega) = f(X_1(\omega), \ldots, X_n(\omega))$ for all $\omega \in \Omega$ is Borel measurable.

Proof: To prove (a), let $\mathcal{C}$ be the collection of all subsets of $\mathbb{R}^n$ of the form (48). Let $B = (-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_n]$ be a set in $\mathcal{C}$. Then

$$\begin{aligned} h^{-1}(B) &= \{\omega \in \Omega : X_1(\omega) \in (-\infty, x_1], \ldots, X_n(\omega) \in (-\infty, x_n]\} \\ &= \cap_{i=1}^n X_i^{-1}((-\infty, x_i]) \end{aligned} \tag{52}$$

Since $X_i^{-1}((-\infty, x_i]) \in \mathcal{F}$ for all i, (52) is a finite intersection of sets in the sigma algebra $\mathcal{F}$. Hence $h^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{C}$, so Lemma 7 implies $h^{-1}(B) \in \mathcal{F}$ for all $B \in \sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R}^n)$. This proves part (a).

Since g is the composition $g(\omega) = f(h(\omega))$ for all $\omega \in \Omega$, part (b) follows immediately from part (a) and Lemma 6. □

H. Continuous functions and USC functions are measurable

Lemma 9: Fix $n, k \in \mathbb{N}$. If $f : \mathbb{R}^n \to \mathbb{R}^k$ is continuous, then it is Borel measurable.

Proof: The open subsets of $\mathbb{R}^k$ generate $\mathcal{B}(\mathbb{R}^k)$, so Lemma 7 ensures it is enough to show $f^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$ for all open sets $B \subseteq \mathbb{R}^k$. Let $B \subseteq \mathbb{R}^k$ be an open set. Since f is continuous, $f^{-1}(B)$ is an open subset of $\mathbb{R}^n$ and hence is in $\mathcal{B}(\mathbb{R}^n)$. □

By Lemmas 8 and 9, it holds that if $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function and $X_1, \ldots, X_n$ are random variables on some probability space, then $f(X_1, \ldots, X_n)$ is a random variable on that space. Two generalizations of Lemma 9 are below.

Lemma 10: Fix $n, k \in \mathbb{N}$. If $f : \mathbb{R}^n \to \mathbb{R}^k$ has at most countably many points of discontinuity, then it is Borel measurable.

Proof: Let B be an open subset of $\mathbb{R}^k$. As before, it is enough to show $f^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$. Let $D \subseteq \mathbb{R}^n$ be the countable set of all points $x \in \mathbb{R}^n$ at which f is not continuous. Then

$$f^{-1}(B) = \big(D \cap f^{-1}(B)\big) \cup \big(D^c \cap f^{-1}(B)\big) \tag{53}$$

We know $D \cap f^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$ because it is a countable subset of $\mathbb{R}^n$. By (53), it suffices to show $D^c \cap f^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$. Fix $x \in D^c \cap f^{-1}(B)$. Then f is continuous at x and $f(x) \in B$. Since B is an open set that contains the point $f(x)$, B also contains an open ball of positive radius centered at $f(x)$; call that open ball $S(f(x))$. Since f is continuous at x, there is an open ball of positive radius centered at x, call it $R(x)$, such that

$$y \in R(x) \implies f(y) \in S(f(x)) \subseteq B$$

It follows that $x \in R(x) \subseteq f^{-1}(B)$. This holds for all $x \in D^c \cap f^{-1}(B)$ and so

$$D^c \cap f^{-1}(B) \subseteq \cup_{x \in D^c \cap f^{-1}(B)} R(x) \subseteq f^{-1}(B)$$

Thus

$$D^c \cap f^{-1}(B) = D^c \cap \big(\cup_{x \in D^c \cap f^{-1}(B)} R(x)\big)$$

We know $D^c \in \mathcal{B}(\mathbb{R}^n)$ because D is a countable subset of $\mathbb{R}^n$ and hence D is Borel measurable. The set $\cup_{x \in D^c \cap f^{-1}(B)} R(x)$ is the union of an arbitrary number of open subsets of $\mathbb{R}^n$, so it is again an open subset of $\mathbb{R}^n$ and hence is in $\mathcal{B}(\mathbb{R}^n)$. □

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *upper semicontinuous (USC)* if for all $x \in \mathbb{R}^n$ we have

$$\limsup_{y \to x} f(y) \leq f(x)$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *lower semicontinuous* (LSC) if $-f$ is USC. It follows that $f : \mathbb{R}^n \to \mathbb{R}$ is continuous if and only if it is both USC and LSC. USC functions are allowed to have discontinuities provided that the function value $f(x)$ at any point of discontinuity x is greater than or equal to the limiting function value taken over any sequence of points that approach x. The next result shows that all USC functions are Borel measurable. It immediately follows that all LSC functions are also Borel measurable.

Lemma 11: If $f : \mathbb{R}^n \to \mathbb{R}$ is USC then it is Borel measurable.

Proof: Fix $z \in \mathbb{R}$ and define $A = \{x \in \mathbb{R}^n : f(x) < z\}$. We want to show A is a Borel measurable set. Since all open subsets of $\mathbb{R}^n$ are Borel measurable, it suffices to show A is open. Fix $x \in A$. Then $f(x) < z$ and the USC property ensures $f(y) < z$ for all points y that are sufficiently close to x. That is, there is a real number $\epsilon > 0$ such that

$$||y - x|| < \epsilon \implies y \in A$$

and so A is open. □

I. Nondecreasing functions

Let $h : \mathbb{R} \to \mathbb{R}$ be a nondecreasing function, meaning that for all $x, y \in \mathbb{R}$

$$x \leq y \implies h(x) \leq h(y)$$

It can be shown that h is Borel measurable (observe that for each $t \in \mathbb{R}$, the set $\{x \in \mathbb{R} : h(x) \leq t\}$ is an interval).

Now fix n as a positive integer. A function $h : \mathbb{R}^n \to \mathbb{R}$ is *entrywise nondecreasing* if for all $x, y \in \mathbb{R}^n$ we have

$$x \leq y \implies h(x) \leq h(y)$$

where the inequality $x \leq y$ means $x_i \leq y_i$ for $i \in \{1, \ldots, n\}$. It is *not true* that all entrywise nondecreasing functions $h : \mathbb{R}^2 \to \mathbb{R}$ are Borel measurable. A counter-example is this: Define $L = \{(x, y) \in \mathbb{R}^2 : x + y = 0\}$. Let A be a subset of L that is nonBorel. Under the axiom of choice, such a set A must exist.[26] Define $h : \mathbb{R}^2 \to \mathbb{R}$ by

$$h(x, y) = \begin{cases} 0 & \text{if } x + y < 0 \\ 1/2 & \text{if } (x, y) \in A \\ 1 & \text{else} \end{cases}$$

The set $\{(x, y) \in \mathbb{R}^2 : h(x, y) = 1/2\}$ is not Borel measurable because this set is equal to A. So h is not a Borel measurable function.

However, it can be shown that all functions $h : \mathbb{R}^n \to \mathbb{R}$ that are both entrywise nondecreasing and right-continuous are Borel measurable. In particular, all joint CDF functions are Borel measurable. This can be proven by observing for all $t \in \mathbb{R}$

$$\{x \in \mathbb{R}^n : h(x) < t\} = \cup_{\{q \in \mathbb{Q}^n : h(q) < t\}} \{x \in \mathbb{R}^n : x \leq q\}$$

which is a countable union of sets in $\mathcal{B}(\mathbb{R}^n)$.

J. Random elements

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $(V, \mathcal{G})$ be another measurable space. A *random element* is a measurable function $X : \Omega \to V$. That is, a function $X : \Omega \to V$ is a random element if and only if

$$X^{-1}(B) \in \mathcal{F} \quad \forall B \in \mathcal{G}$$

It follows that if $X : \Omega \to V$ is a random element then $P[X \in B]$ exists for all $B \in \mathcal{G}$. The collection of probabilities $P[X \in B]$ for all $B \in \mathcal{G}$ is called the *probability distribution of* X.

If $(V_1, \mathcal{G}_1)$ and $(V_2, \mathcal{G}_2)$ are two measurable spaces, then random elements $X : \Omega \to V_1$ and $Y : \Omega \to V_2$ are defined as *independent* if

$$P[X \in B_1, Y \in B_2] = P[X \in B_1]P[Y \in B_2] \quad \forall B_1 \in \mathcal{G}_1, B_2 \in \mathcal{G}_2 \tag{54}$$

The definition of mutual independence for a finite or countably infinite collection of random elements is similar to the corresponding definitions for random variables.

A random variable is just a random element with output space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Lemma 8a implies that a random element $X : \Omega \to \mathbb{R}^n$ can equivalently be viewed as a n-tuple of random variables: $X = (X_1, \ldots, X_n)$. This is also called a *random vector*. It can be shown that the definition of independence in (54) holds for two random vectors $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^m$ if and only if

$$P[X \leq x, Y \leq y] = P[X \leq x]P[Y \leq y] \quad \forall x \in \mathbb{R}^n, y \in \mathbb{R}^m$$

where the vector inequality $X \leq x$ is taken entrywise. Specifically, with $X = (X_1, \ldots, X_n)$ and $x = (x_1, \ldots, x_n)$, the inequality $X \leq x$ means

$$X_1 \leq x_1, \ X_2 \leq x_2, \ \ldots, \ X_n \leq x_n$$

K. Random elements with infinite values

The random element structure can be used to formally define an *extended random variable* that can take infinite values as a measurable function $X : \Omega \to \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ with

$$V = \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$$
$$\mathcal{G} = \sigma(\mathcal{B}(\mathbb{R}) \cup \{\infty\} \cup \{-\infty\})$$

The structure can be extended to formally allow a *Does Not Exist (DNE)* option via

$$V = \mathbb{R} \cup \{\infty\} \cup \{-\infty\} \cup \{DNE\}$$
$$\mathcal{G} = \sigma(\mathcal{B}(\mathbb{R}) \cup \{\infty\} \cup \{-\infty\} \cup \{DNE\})$$

The random element structure allows the set V to be composed of any type of object (such as trees, graphs, sets, functions, and so on), provided that there is a corresponding sigma algebra $\mathcal{G}$ on V.

[26] One way to argue that a nonBorel set $A \subseteq L$ exists is to observe that, under the axiom of choice, the set of all subsets of L has cardinality $|Pow(\mathbb{R})|$ while the set of all Borel subsets of L has cardinality $|\mathbb{R}|$ [6].

L. Lebesgue measurable sets

Let A be a Borel measurable subset of $\mathbb{R}$ with $\mu(A) = 0$. If B is a set that satisfies $B \subseteq A$, *and if B is also Borel measurable*, then we must have $\mu(B) = 0$. On the other hand, if B is not Borel measurable then $\mu(B)$ does not exist. Are all subsets of A Borel measurable? Not necessarily. Under the axiom of choice, the sigma algebra $\mathcal{B}(\mathbb{R})$ is known to be *incomplete*, meaning there are sets $A \in \mathcal{B}(\mathbb{R})$ with $\mu(A) = 0$, but such that not all subsets of A are in $\mathcal{B}(\mathbb{R})$. For example, the *Cantor set* is a famous subset of $[0, 1]$ that is Borel measurable, has measure 0, but has the same cardinality as $\mathbb{R}$. Therefore, the set of all subsets of the Cantor set must then have strictly larger cardinality than $\mathbb{R}$. On the other hand, the axiom of choice implies that $\mathcal{B}(\mathbb{R})$ has the same cardinality as $\mathbb{R}$. We conclude that there are subsets of the Cantor set that are not in $\mathcal{B}(\mathbb{R})$.

The incompleteness of $\mathcal{B}(\mathbb{R})$ is not a cause of major concern. Incompleteness typically does not impact practical probability problems. However, some people prefer to work with the *completion* of $\mathcal{B}(\mathbb{R})$, which is the sigma algebra generated by all sets that are already in $\mathcal{B}(\mathbb{R})$ and all sets $\tilde{A}$ for which $\tilde{A} \subseteq A$ for some set $A \in \mathcal{B}(\mathbb{R})$ for which $\mu(A) = 0$. This is denoted $\mathcal{L}(\mathbb{R})$ and is called the *Lebesgue sigma algebra on* $\mathbb{R}$. A set in $\mathcal{L}(\mathbb{R})$ is called a *Lebesgue measurable subset of* $\mathbb{R}$. It can be shown that $A \in \mathcal{L}(\mathbb{R})$ if and only if A can be written in a decomposed form $A = A_1 \cup A_2$ where A_1, A_2 are disjoint, $A_1 \in \mathcal{B}(\mathbb{R})$, and $A_2 \subseteq A_3$ for some set $A_3 \in \mathcal{B}(\mathbb{R})$ such that $\mu(A_3) = 0$. Further, any two such decompositions $A = A_1 \cup A_2$ and $A = \tilde{A}_1 \cup \tilde{A}_2$ must satisfy $\mu(A_1) = \mu(\tilde{A}_1)$. The measure $\lambda : \mathcal{L}(\mathbb{R}) \to [0, \infty) \cup \{\infty\}$ maps every set $A \in \mathcal{L}(\mathbb{R})$ to its value $\mu(A_1)$ under a decomposition $A = A_1 \cup A_2$. Therefore, $\lambda(A) = \mu(A)$ for all $A \in \mathcal{B}(\mathbb{R})$. Thus, λ allows the concept of *length* to be extended beyond the original set $\mathcal{B}(\mathbb{R})$. One might hope that $\mathcal{L}(\mathbb{R})$ contains all subsets of $\mathbb{R}$, but this is not true. The axiom of choice can be used to prove there are subsets of $\mathbb{R}$ that are neither Borel measurable nor Lebesgue measurable. These sets are so complicated that they cannot be assigned a length (see, for example, chapter 3 in [3]).

A similar story holds in the multidimensional case. Let n be a positive integer. Under the axiom of choice, the sigma algebra $\mathcal{B}(\mathbb{R}^n)$ is incomplete. The completion is $\mathcal{L}(\mathbb{R}^n)$, which is the collection of Lebesgue measurable subsets of $\mathbb{R}^n$. A set $A \subseteq \mathbb{R}^n$ is in $\mathcal{L}(\mathbb{R}^n)$ if and only if $A = A_1 \cup A_2$ for some disjoint sets A_1, A_2 that satisfy $A_1 \in \mathcal{B}(\mathbb{R}^n)$ and $A_2 \subseteq A_3$ for some set $A_3 \in \mathcal{B}(\mathbb{R}^n)$ such that $\mu(A_3) = 0$. This allows the concept of multidimensional volume to be extended beyond the original set $\mathcal{B}(\mathbb{R}^n)$. As before, some subsets of $\mathbb{R}^n$ are so complicated that they cannot be assigned a volume and so are not in $\mathcal{L}(\mathbb{R}^n)$.

M. Completing a probability space

A related topic is the concept of an *incomplete probability space* $(\Omega, \mathcal{F}, P)$, being a space that contains at least one $A \in \mathcal{F}$ for which $P[A] = 0$, but that has a subset $B \subseteq A$ such that $B \notin \mathcal{F}$. It can be shown that all such spaces can be *completed* to a space $(\Omega, \tilde{\mathcal{F}}, \tilde{P})$ by adding some sets to $\mathcal{F}$, and the completion does not significantly change anything (see, for example, [7]). In particular, every set $A \in \tilde{\mathcal{F}}$ has the form $A = A_1 \cup A_2$ for some disjoint sets A_1, A_2 that satisfy $A_1 \in \mathcal{F}$ and $A_2 \subseteq A_3$ for some $A_3 \in \mathcal{F}$ with $P[A_3] = 0$. Under any such a decomposition $A = A_1 \cup A_2$ we have $\tilde{P}[A] = P[A_1]$. In this completed space, any subset of an event with probability zero is itself and event (and also has probability zero).

An example is $\Omega = \{a, b, c, d\}$ and sigma algebra

$$\mathcal{F} = \{\phi, \Omega, \{a, b\}, \{c, d\}\}$$

Let $P : \mathcal{F} \to [0, 1]$ be the probability measure with $P[\{a, b\}] = 1$ and $P[\{c, d\}] = 0$. The space $(\Omega, \mathcal{F}, P)$ is incomplete because $P[\{c, d\}] = 0$ and $\{c\} \subseteq \{c, d\}$ but $\{c\} \notin \mathcal{F}$. The completed space $(\Omega, \tilde{\mathcal{F}}, \tilde{P})$ uses

$$\tilde{\mathcal{F}} = \{\phi, \Omega, \{a, b\}, \{c, d\}, \{c\}, \{d\}, \{a, b, c\}, \{a, b, d\}\}$$

and defines $\tilde{P} : \tilde{\mathcal{F}} \to [0, 1]$ by

$$\tilde{P}[\{a, b\}] = \tilde{P}[\{a, b, c\}] = \tilde{P}[\{a, b, d\}] = 1, \quad \tilde{P}[\{c, d\}] = \tilde{P}[\{c\}] = \tilde{P}[\{d\}] = 0$$

N. A useful probability space

Define $\mathcal{B}([0, 1]) = \{A \in \mathcal{B}(\mathbb{R}) : A \subseteq [0, 1]\}$. The triplet $([0, 1], \mathcal{B}([0, 1]), \mu)$, where μ represents the Borel measure restricted to Borel subsets of $[0, 1]$, can be shown to be a probability space. That is, we can define

$$\Omega = [0, 1]$$
$$\mathcal{F} = \mathcal{B}([0, 1])$$

and we define $P : \mathcal{B}([0, 1]) \to [0, 1]$ by

$$P[A] = \mu(A) \quad \forall A \in \mathcal{B}([0, 1])$$

On this probability space, the identity random variable $U : [0, 1] \to \mathbb{R}$, defined by $U(\omega) = \omega$ for all $\omega \in [0, 1]$, has the Unif$[0, 1]$ distribution:

$$F_U(u) = P[U \leq u] = \begin{cases} 1 & \text{if } u > 1 \\ u & \text{if } 0 \leq u \leq 1 \\ 0 & \text{if } u < 0 \end{cases}$$

The triplet $([0,1], \mathcal{B}([0,1]), \mu)$ is perhaps the simplest probability space that houses a Unif$[0,1]$ random variable. We now argue that this probability space can be used to model almost any probability problem of practical interest.

Any probability space that houses just one Unif$[0,1]$ random variable must also house a countably infinite sequence of i.i.d. Unif$[0,1]$ random variables. This is because there exists a bijective function $f : [0,1] \to [0,1]^{\mathbb{N}}$ that is measurable and that has a measurable inverse such that for any random variable U that takes values in $[0,1]$ and satisfies $U \sim$ Unif$[0,1]$ we have

$$f(U) = (f_1(U), f_2(U), f_3(U), \ldots)$$

where $\{f_i(U)\}_{i=1}^{\infty}$ are i.i.d. Unif$[0,1]$ random variables.[27]

Computer simulations use at most a countably infinite number of calls to a random number generator rand(). Such calls can be modeled on any probability space that contains a single random variable $U \sim$ Unif$[0,1]$: The simulation uses $f_1(U)$ as the result of the first call to rand(), $f_2(U)$ as the result of the second call to rand(), and so on. Arguably, any output variable computed by the simulation must use measurable operations and hence can be viewed as a measurable function of U. It can be shown that a Unif$[0,1]$ random variable can be used to construct a random variable of any desired distribution. Also, it can be shown that an infinite sequence of i.i.d. Unif$[0,1]$ random variables can be used to construct a countably infinite number of mutually independent Brownian motion processes. Overall, $([0,1], \mathcal{B}([0,1]), \mu)$ can be used as the probability space for most practical probability problems in the world of science, engineering, and economics.

The probability space $([0,1], \mathcal{B}([0,1]), \mu)$ cannot house an *uncountably infinite* collection of i.i.d. Unif$[0,1]$ random variables. There are larger spaces that can (see next subsection).

The probability space $([0,1], \mathcal{B}([0,1]), \mu)$ is incomplete. Its completion is $([0,1], \mathcal{L}([0,1]), \lambda)$, where $\mathcal{L}([0,1])$ is the sigma algebra of Lebesgue measurable subsets of $[0,1]$ and λ is the Lebesgue measure restricted to $\mathcal{L}([0,1])$.

O. Product spaces

Let I be a nonempty set that shall be used as a (possibly uncountably infinite) index set. For each $i \in I$, let $(\Omega_i, \mathcal{F}_i)$ be a measurable space. The *product space* $(\Omega, \mathcal{F})$ is constructed as follows. Define the following set of tuples:

$$\Omega = \times_{i \in I} \Omega_i = \{(\omega_i)_{i \in I} : \omega_i \in \Omega_i \quad \forall i \in I\}$$

To define $\mathcal{F}$, let $\mathcal{C}$ be the collection of all sets of the type $\times_{i \in I} A_i$ where $A_i \in \mathcal{F}_i$ for all $i \in I$, and $A_i = \Omega_i$ for all but at most one $i \in I$. By convention, we define $\times_{i \in I} A_i$ as the empty set ϕ if $A_i = \phi$ for at least one $i \in I$. The collection $\mathcal{C}$ is called a *cylinder set*. Define $\mathcal{F} = \sigma(\mathcal{C})$. Set $\mathcal{F}$ is called the *product sigma algebra on* Ω and is denoted $\mathcal{F} = \otimes_{i \in I} \mathcal{F}_i$, where the notation $\otimes$ means this product space construction is used.[28] The $\otimes$ notation is also used to represent the product space $(\Omega, \mathcal{F})$ itself:

$$(\Omega, \mathcal{F}) = \otimes_{i \in I} (\Omega_i, \mathcal{F}_i) \tag{55}$$

For example, let $I = \{1,2,3,4\}$. Suppose we have four measurable spaces $(\Omega_i, \mathcal{F}_i)$ for $i \in \{1,2,3,4\}$. Let $\mathcal{C}$ be the corresponding cylinder set. A set is in $\mathcal{C}$ if and only if it has one of four forms:

$$\begin{aligned}
&A_1 \times \Omega_2 \times \Omega_3 \times \Omega_4 \quad \text{for some set } A_1 \in \mathcal{F}_1 \\
&\Omega_1 \times A_2 \times \Omega_3 \times \Omega_4 \quad \text{for some set } A_2 \in \mathcal{F}_2 \\
&\Omega_1 \times \Omega_2 \times A_3 \times \Omega_4 \quad \text{for some set } A_3 \in \mathcal{F}_3 \\
&\Omega_1 \times \Omega_2 \times \Omega_3 \times A_4 \quad \text{for some set } A_4 \in \mathcal{F}_4
\end{aligned}$$

In general, $\mathcal{C}$ is *not* a sigma algebra, which is why we define $\mathcal{F} = \sigma(\mathcal{C})$. Indeed, let $A_1 \in \mathcal{F}_1$ be a set such that $A_1 \neq \Omega_1$ and $A_1 \neq \phi$. Likewise, let $A_2 \in \mathcal{F}_2$ be a set such that $A_2 \neq \Omega_2$ and $A_2 \neq \phi$. Then

$$\begin{aligned}
&A_1 \times \Omega_2 \times \Omega_3 \times \Omega_4 \in \mathcal{C} \\
&\Omega_1 \times A_2 \times \Omega_3 \times \Omega_4 \in \mathcal{C}
\end{aligned}$$

However, the union of these two sets in $\mathcal{C}$ does not have one of the four required forms and hence is not a set in $\mathcal{C}$:

$$(A_1 \times \Omega_2 \times \Omega_3 \times \Omega_4) \cup (\Omega_1 \times A_2 \times \Omega_3 \times \Omega_4) \notin \mathcal{C}$$

[27]There is an isomorphism $f : [0,1] \to [0,1]^{\mathbb{N}}$ such that if $U \sim$ Unif$[0,1]$ then $f(U)$ is a countably infinite sequence of i.i.d. Unif$[0,1]$ variables, see Theorem A.47 in [9] and Chapter 13 of [10]. Related: It can be shown that if A is an uncountably infinite Borel subset of some Borel space (such as an uncountably infinite Borel subset of $\mathbb{R}$ or $\mathbb{R}^n$ or $\mathbb{R}^{\mathbb{N}}$) then there is a bijection $f : [0,1] \to A$ that is measurable and has a measurable inverse. This is a result of Kuratowski in [11], see also statement and proof in Theorem 3.3.13 of [6]. The set $\mathcal{B}(\mathbb{R}^{\mathbb{N}})$ is the sigma algebra generated by sets of the type $A_1 \times A_2 \times \cdots \times A_n \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \cdots$ for some positive integer n and some sets $A_1, ..., A_n \in \mathcal{B}(\mathbb{R})$.

[28]Caveat: The set $\times_{i \in I} \mathcal{F}_i$ is not even the same *type* of thing as $\otimes_{i \in I} \mathcal{F}_i$. An element of $\times_{i \in I} \mathcal{F}_i$ is a tuple of sets. An element of $\otimes_{i \in I} \mathcal{F}_i$ is a set of tuples. For example, suppose $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$. Then $(A_1, A_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ while $A_1 \times A_2 \in \mathcal{F}_1 \otimes \mathcal{F}_2$.

The union contains tuples of the form $(a, b, c, d) \in \Omega_1 \times \Omega_2 \times \Omega_3 \times \Omega_4$ such that either $a \in A_1$ or $b \in A_2$. This example with $I = \{1, 2, 3, 4\}$ is also useful to illustrate our convention about the empty set:

$$\phi \times \Omega_2 \times \Omega_3 \times \Omega_4 = \phi$$
$$\Omega_1 \times \phi \times \Omega_3 \times \Omega_4 = \phi$$
$$A_1 \times A_2 \times \phi \times \Omega_4 = \phi$$

When I is an uncountably infinite index set and the product space $(\Omega, \mathcal{F})$ is defined by (55), a function $X : \Omega \to \mathbb{R}$ has *uncountably many variables* ω_i. In particular, each $(\omega_i)_{i\in I} \in \Omega$ produces a real number $X((\omega_i)_{i\in I})$. If such a function $X : \Omega \to \mathbb{R}$ is Borel measurable then it can be shown to have the following unusual property: If $X : \Omega \to \mathbb{R}$ is Borel measurable, then there is a countable set $J \subseteq I$ such that for all $(\omega_i)_{i\in I} \in \Omega$, the value $X((\omega_i)_{i\in I})$ is determined just by knowing ω_j for all $j \in J$. That is, on this product space $(\Omega, \mathcal{F})$, a Borel measurable function of uncountably many variables really is a function of only countably many of those variables.[29]

Let I be a general nonempty index set and let $(\Omega, \mathcal{F})$ be defined by (55). If for each $i \in I$ we have a probability measure $P_i : \mathcal{F}_i \to [0, 1]$, so $(\Omega_i, \mathcal{F}_i, P_i)$ is a probability space, the *Kolmogorov extension theorem* ensures existence of a unique probability measure $P : \mathcal{F} \to [0, 1]$ that treats events that occur on any finite collection of the individual spaces as being mutually independent when viewed on the larger space (see, for example, [3]). The resulting *product probability space* is also represented by the $\otimes$ notation:

$$(\Omega, \mathcal{F}, P) = \otimes_{i\in I}(\Omega_i, \mathcal{F}_i, P_i)$$

where $(\Omega, \mathcal{F})$ is the same as in (55). The most useful example is when $I = \{1, 2\}$, so $(\Omega, \mathcal{F}, P) = (\Omega_1, \mathcal{F}_1, P_1) \otimes (\Omega_2, \mathcal{F}_2, P_2)$ and $P : \mathcal{F}_1 \otimes \mathcal{F}_2 \to [0, 1]$ satisfies

$$P[A_1 \times A_2] = P_1[A_1]P_2[A_2] \quad \forall A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2$$

and P is defined on sets in $\mathcal{F}_1 \otimes \mathcal{F}_2$ that are not of the type $A_1 \times A_2$ in a way that does not violate the axioms of probability. When I is uncountably infinite, this extension theorem formally proves existence of a probability space that houses uncountably many i.i.d. Unif$[0, 1]$ random variables, being random variables X_i such that $X_i \sim$ Unif$[0, 1]$ for each $i \in I$ and such that every finite collection of these random variables are mutually independent. Arguably, no practical situation would ever require so much independence.

[29] See related Exercise 1.1.22 in [12] and Section 3.13d in [7] and note that X is determined by knowing the indicator functions $1_{\{X\leq x\}}$ for $x \in \mathbb{Q}$.

APPENDIX E – A FUNCTION THAT VIOLATES COUNTABLE SUMMABILITY

This section constructs a function $P : \mathcal{F} \rightarrow [0,1]$ that satisfies Axioms 1 and 2 of probability, and also satisfies *finite summability* (as defined in Section I-D), but does not satisfy countable summability (Axiom 3). Existence of this example *proves that it is impossible to prove* countable summability using only the first two axioms together with finite summability. It turns out that the example also implies it is impossible to prove the countably infinite version of the union bound using only the finite version, that is, it is impossible to prove (6) using only (5). This is advanced material. Nobody is allowed to read this appendix. Stop reading now. There is a monster here.

Let $\mathbb{N} = \{1, 2, 3, \ldots\}$ denote the set of positive integers. Define the sample space Ω and the sigma algebra $\mathcal{F}$ by

$$\Omega = \mathbb{N}$$
$$\mathcal{F} = Pow(\mathbb{N})$$

where $Pow(\mathbb{N})$ denotes the set of all subsets of $\mathbb{N}$.

The construction of $P : Pow(\mathbb{N}) \rightarrow [0,1]$ in this appendix uses the advanced concept of a *Banach limit*. Let B denote the set of all real-valued bounded sequences $\{a_i\}_{i=1}^{\infty}$, that is

$$B = \left\{ \{a_i\}_{i=1}^{\infty} : a_i \in \mathbb{R} \quad \forall i \in \mathbb{N} \text{ and } \sup_{i \in \mathbb{N}} |a_i| < \infty \right\}$$

Using a set theory axiom called the *axiom of choice*, it can be shown that there exists a function $\psi : B \rightarrow \mathbb{R}$ that satisfies the following properties:

- Limit-preserving:
$$\liminf_{i \rightarrow \infty} a_i \leq \psi\left(\{a_i\}_{i=1}^{\infty}\right) \leq \limsup_{i \rightarrow \infty} a_i \quad \forall \{a_i\}_{i=1}^{\infty} \in B$$
- Linearity: For all $\lambda, \mu \in \mathbb{R}$ and all $\{a_i\}_{i=1}^{\infty}$ and $\{b_i\}_{i=1}^{\infty}$ in B we have
$$\psi\left(\{\lambda a_i + \mu b_i\}_{i=1}^{\infty}\right) = \lambda \psi\left(\{a_i\}_{i=1}^{\infty}\right) + \mu \psi\left(\{b_i\}_{i=1}^{\infty}\right)$$

The limit-preserving property of ψ implies that for all $\{a_i\}_{i=1}^{\infty} \in B$ we have

$$\psi\left(\{a_i\}_{i=1}^{\infty}\right) = \lim_{i \rightarrow \infty} a_i$$

whenever the limit on the right-hand-side exists. The value $\psi(\{a_i\}_{i=1}^{\infty})$ is called a *Banach limit* of the sequence $\{a_i\}_{i=1}^{\infty}$, and is defined for every real-valued bounded sequence.

We now construct $P : Pow(\mathbb{N}) \rightarrow [0,1]$. For each $A \subseteq \mathbb{N}$ define

$$|A \cap \{1, \ldots, n\}|$$

as the number of elements in the intersection of A and the set of positive integers from 1 to n, inclusive. Note that

$$0 \leq |A \cap \{1, \ldots, n\}| \leq n \quad \forall A \subseteq \mathbb{N}$$

Define $P : Pow(\mathbb{N}) \rightarrow [0,1]$ by

$$P[A] = \psi\left(\left\{ \frac{|A \cap \{1, ..., n\}|}{n} \right\}_{n=1}^{\infty} \right)$$

In particular, for any $A \subseteq \mathbb{N}$ we have

$$P[A] = \lim_{n \rightarrow \infty} \frac{|A \cap \{1, ..., n\}|}{n} \tag{56}$$

whenever the limit on the right-hand-side exists. Usage of the sophisticated function ψ ensures $P[A]$ is defined for all $A \subseteq \mathbb{N}$, including sets A for which the limit on the right-hand-side of (56) does not exist. Specifically, the limit-preserving property of ψ implies

$$\liminf_{n \rightarrow \infty} \frac{|A \cap \{1, ..., n\}|}{n} \leq P[A] \leq \limsup_{n \rightarrow \infty} \frac{|A \cap \{1, ..., n\}|}{n} \quad \forall A \subseteq \mathbb{N} \tag{57}$$

The limit-preserving and linearity properties of ψ can be used to show that P satisfies the first two axioms of probability and also satisfies the following *axiom of finite summability*: If A and B are disjoint subsets of $\mathbb{N}$ then[30]

$$P[A \cup B] = P[A] + P[B]$$

Indeed:

[30]Using induction, it follows that the axiom of finite summability (as stated here for the union of two disjoint elements of $\mathcal{F}$) is equivalent to (2).

- Axiom 1: Fix $A \subseteq \mathbb{N}$. We have for all positive integers n that
$$0 \leq \frac{|A \cap \{1, ..., n\}|}{n}$$
and so
$$0 \leq \liminf \frac{|A \cap \{1, ..., n\}|}{n} \overset{(a)}{\leq} P[A]$$
where inequality (a) holds by (57). So P satisfies Axiom 1.
- Axiom 2: We have
$$\frac{|\mathbb{N} \cap \{1, ..., n\}|}{n} = \frac{|\{1, ..., n\}|}{n} = 1 \quad \forall n \in \{1, 2, 3, ...\}$$
Therefore, (56) implies $P[\mathbb{N}] = 1$. So P satisfies Axiom 2.
- Finite summability: Let A and B be disjoint subsets of $\mathbb{N}$. Then for all positive integers n we have
$$\frac{|(A \cup B) \cap \{1, ..., n\}|}{n} = \frac{|A \cap \{1, ..., n\}|}{n} + \frac{|B \cap \{1, ..., n\}|}{n}$$
and linearity of ψ implies $P[A \cup B] = P[A] + P[B]$.

For example, it can be shown from (56) that

$$\begin{aligned}
&P[\mathbb{N}] = 1 \\
&P[\{\text{even positive integers}\}] = 1/2 \\
&P[\{\text{odd positive integers}\}] = 1/2 \\
&P[\{3n + 8 : n \in \mathbb{N}\}] = 1/3 \\
&P[\{n^2 : n \in \mathbb{N}\}] = 0
\end{aligned}$$

Further, if A is any finite subset of $\mathbb{N}$ then $|A \cap \{1, \ldots, n\}| \leq |A|$, where $|A|$ is the (finite) number of elements in A, so

$$0 \leq P[A] \leq \limsup_{n\to\infty} \frac{|A \cap \{1, ..., n\}|}{n} \leq \limsup_{n\to\infty} \frac{|A|}{n} = 0$$

That is, if A is a finite subset of $\mathbb{N}$ then $P[A] = 0$. Thus

$$\sum_{i=1}^{\infty} P[\{i\}] = \sum_{i=1}^{\infty} 0 = 0$$

On the other hand

$$P[\cup_{i=1}^{\infty} \{i\}] = P[\mathbb{N}] = 1$$

and so

$$\underbrace{P[\cup_{i=1}^{\infty} \{i\}]}_{1} > \underbrace{\sum_{i=1}^{\infty} P[\{i\}]}_{0} \tag{58}$$

Since $1 \neq 0$, this function P violates countable summability. That is, P does not satisfy Axiom 3 and so P is not a probability measure.

The inequality (58) is also a violation of the countable version of the union bound stated in (6), even though the finite version (5) can be shown to hold for this function P. Further, since $\cup_{i=1}^{n} \{i\} \nearrow \mathbb{N}$ and

$$\lim_{n\to\infty} P[\cup_{i=1}^{n} \{i\}] = 0 \neq 1 = P[\mathbb{N}]$$

we see that this function P violates continuity of probability. Several standard properties of a CDF also fail: If we define $X : \mathbb{N} \to \mathbb{R}$ by $X(i) = i$ for $i \in \mathbb{N}$, then X trivially satisfies the measurability requirement (11) because $\mathcal{F} = Pow(\mathbb{N})$. However

$$P[X \leq n] = \sum_{i=1}^{n} P[\{i\}] = 0 \quad \forall n \in \mathbb{N}$$

and so $\lim_{n\to\infty} P[X \leq n] = 0$ for this unusual function P.[31] Similarly, we can define a function $Y : \mathbb{N} \to \mathbb{R}$ such that $\lim_{n\to\infty} P[Y \leq -n] = 1$. We can define another function $Z : \mathbb{N} \to \mathbb{R}$ such that right-continuity fails for $P[Z \leq x]$ as a function of $x \in \mathbb{R}$. This is the monster at the end of the notes.

[31]The proof of the standard CDF property $\lim_{n\to\infty} P[X \leq n] = 1$ crucially uses continuity of probability, see item 27 in Section III-H.

REFERENCES

[1] A. Leon-Garcia. *Probability, Statistics, and Random Processes For Electrical Engineering, 3rd ed.* Pearson Education Inc., 2008.
[2] S. Ross. *Introduction to Probability Models*. Academic Press, 8th edition, Dec. 2002.
[3] P. Billingsley. *Probability and Measure, 3rd edition*. New York: John Wiley & Sons, 1995.
[4] O. Kallenberg. *Foundations of Modern Probability, 2nd ed., Probability and its Applications*. Springer-Verlag, 2002.
[5] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.
[6] S. M. Srivastava. *A Course on Borel Sets*. Springer New York, NY, 1998.
[7] D. Williams. *Probability with Martingales*. Cambridge Mathematical Textbooks, Cambridge University Press, 1991.
[8] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrscheinlichkeitstheorie verw. Geiete*, vol. 5:pp. 119–122, 1981.
[9] L. Breiman. *Probability*. SIAM, Philadelphia, 2nd edition, 1968-92.
[10] R. M. Dudley. *Real Analysis and Probability*. Wadsworth Pub. Co., Pacific Grove, CA, 1989.
[11] K. Kuratowski. Sur une généralisation de la notion d'homéomorphie. *Fund. Math*, 22:206–220, 1934.
[12] A. Dembo. Probability theory: Stat 310 / math 230. *Stanford University; unpublished notes*, 2013.