PROBABILITY THEORY

Dinh Huu Nguyen, Fall 2014

Abstract: lecture notes for Probability Theory at John von Neumann Institute, Vietnam, Fall 2014.

Contents

Whether probability is frequency of uncertain phenomena or it is strength of personal beliefs, probability theory studies and predicts the patterns of such random occurrences.

1. Preliminaries

1.1. **Set Theory.**

We begin with the most fundamental object in mathematics, a set.

Definition 1.1. A set X is a collection of distinct, definite objects.

Each object x of X is called an element and written $x \in X$. Each subcollection U of X is called a subset and written $U \subset X$. Each subset usually contains elements of some property P. The empty set is denoted by $\varnothing$. A finite set can be enumerated $X = \{x_1, \dots, x_n\}$. A countably infinite set can also be indexed $X = \{x_1, x_2, x_3, ...\}$.

Example 1.2. The set of all natural numbers is $\mathbb{N} = \{0, 1, 2, \dots\}$, the set of all positive natural numbers is $\mathbb{N}^+ = \{1, 2, 3, \dots\}$, and the set of all real numbers is $\mathbb{R}$. Surely $\mathbb{N}^+ \subset \mathbb{N} \subset \mathbb{R}$.

Example 1.3. The set of all possible outcomes when we roll a die is $X = \{1, 2, 3, 4, 5, 6\}$.

Example 1.4. The set of all possible outcomes when we toss a coin is $X = \{H, T\}$.

Example 1.5. The set of all subsets of X is its power set $\mathbb{P}(X)$.

1.2. **Set Operations.** Below are basic set operations and identities, all of which can be seen and verified by Venn diagrams.

- complement: if $U \subset X$ then $U^c = \{x \in X \,|\, x \notin U\}$.
- union: $A \bigcup B = \{x \,|\, x \in A \text{ or } x \in B\}$.
- intersection: $A \bigcap B = \{x \,|\, x \in A \text{ and } x \in B\}$.
- disjoint: if $A \bigcap B = \varnothing$ then we say A and B are disjoint.
- disjoint union: if A and B are disjoint then we write their union as $A \bigsqcup B$.
- partition: if $\bigsqcup\limits_{i \in I} A_i = X$ then we say the A_i partition X.
- De Morgan's law: $(A \bigcup B)^c = A^c \bigcap B^c$ and $(A \bigcap B)^c = A^c \bigcup B^c$
- relative complement of B in A: $A \backslash B = \{\text{all } x \text{ in } A \text{ but not in } B\}$.
- symmetric difference: $A \vartriangle B = (A \bigcup B) \backslash (A \bigcap B) = A \backslash B \bigcup B \backslash A$

1.3. **Maps between Sets.** It is really important to consider the relationships, or maps between sets beside looking within each set.

Definition 1.6. A map f between X and Y is a law that assigns each $x \in X$ a unique element $f(x) \in Y$.

Such map is often denoted by $X \xrightarrow{f} Y$. We call X and Y the domain and codomain of f. For any subset $U \subset X$, the set $f(U) = \{f(x), x \in U\} \subset Y$ is called the image of U. For any subset $V \subset Y$, the set $X^{-1}(V) = \{X \in V\} = \{x, f(x) \in V\} \subset X$ is called the inverse image of V.

Example 1.7. For every nonempty set X, the map $X \xrightarrow{f} X, x \mapsto x$ is called the identity map and often denoted by id_X.

Example 1.8. For every nonempty sets X and Y, the map $X \xrightarrow{f} Y, x \mapsto y_0$ for some $y_0 \in Y$ is called a constant map and often denoted by c.

Example 1.9. If $U \subset X$ is a subset then we can define $U \longrightarrow X, u \mapsto u$. This is an embedding of U into X as itself. Or we can define $X \xrightarrow{1_U} \mathbb{R}, x \mapsto 1$ if $x \in U$ and $x \mapsto 0$ if $x \notin U$. This is the indicator function of U, it lets us know when an element $x \in X$ is in U.

Definition 1.10. Given $X \xrightarrow{f} Y$ and $Y \xrightarrow{g} Z$ we define their composition to be $X \xrightarrow{gf} Z, x \mapsto g(f(x))$.

Below is the picture of composition

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[dr, "gf"] & Y \arrow[d, "g"] \\
& Z
\end{tikzcd}
\qquad
\begin{tikzcd}
x \arrow[r] \arrow[dr] & f(x) \arrow[d] \\
& g(f(x))
\end{tikzcd}$$

Definition 1.11. A map $X \xrightarrow{f} Y$ is called injective (or one-to-one) if $f(x) \neq f(x')$ whenever $x \neq x' \in X$. It is called surjective (or onto) if $f(X) = \{f(x), \text{ all } x \in X\} = Y$. It is called bijective (or one-to-one and onto) if it is both injective and surjective. In this case, we can define an inverse map $Y \xrightarrow{g} X, y = f(x) \mapsto x$ such that $gf = id_X$ and $fg = id_Y$. This inverse map is unique and denoted by f^{-1}.

Example 1.12. If $m \leq n$ then there exists an injection $\{x_1, \ldots, x_m\} \xrightarrow{f} \{y_1, \ldots, y_n\}$. In general, an injection $X \xrightarrow{f} Y$ allows us to view X as a subset of Y. Conversely, each subset $U \subset X$ is an injection $U \xrightarrow{f} X$.

Example 1.13. If $m < n$ then there does not exist any surjection $\{x_1, \ldots, x_m\} \xrightarrow{f} \{y_1, \ldots, y_n\}$. The same does not hold for infinite sets, as we can have a surjection $5\mathbb{Z} \longrightarrow \mathbb{Z}, 5z \mapsto z$ despite the fact that $5\mathbb{Z} \subsetneq \mathbb{Z}$.

Already maps allow us to rigorously compare the cardinalities of sets. We say that $|X| \leq |Y|$ if there exists an injection $X \xrightarrow{f} Y$, that $|X| \geq |Y|$ if there exists a surjection $X \xrightarrow{f} Y$, and that $|X| = |Y|$ if there exists a bijection $X \xrightarrow{f} Y$.

Example 1.14. $5\mathbb{Z}, \mathbb{Z}, \mathbb{Z} \times \mathbb{Z}, \mathbb{N}, \mathbb{N}^+, \mathbb{Q}$ all have the same cardinality.

Exercise 1.15. Show that $|\mathbb{Z}| < |\mathbb{R}|$ with strict inequality.

1.4. **Counting Permutations and Combinations.** It is much easier to compare sizes of finite sets, we just need to count the number of elements in each set. Below are some formulas to do so.

1.4.1. *Permutations of n Objects.* The number of ways to permute n objects, or equally the number of ways to arrange n objects in order is $P(n, n) = n!$.

Example 1.16. The number of ways to line up ten people for distinct pictures is $10! = 3628800$.

1.4.2. *k-Permutations of n Objects.* If we only use k of those n objects then that number is

$$\begin{aligned} P(n,k) &= n \cdot (n-1) \cdots (n-(k+1)) \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

Example 1.17. The number of distinct pictures of four people out of ten people is $P(10, 4) = 5040$.

1.4.3. *Combinations.* The number of ways to choose k objects out of n objects with disregard for order is $C(n,k) = \frac{n!}{k!(n-k)!}$

Example 1.18. The number of teams of four people out of ten people is $C(10,4) = 210$.

1.4.4. *Partition.* Combinations are a special case of partitions: they are the number of ways to partition a set of size n into two sets of size k and size $n-k$. If we partition a set of size n into r sets of sizes $k_1, \dots, k_r$ with $k_1 + \dots + k_r = n$ then there are

$$\begin{aligned} C(n,k_1,\dots,k_r) &= \frac{n!}{k_1!(n-k_1)!}\frac{(n-k_1)!}{k_2!(n-k_1-k_2)!}\frac{(n-k_1-k_2)!}{k_3!(n-k_1-k_2-k_3)!}\cdots\frac{(n-k_1-\dots-k_{r-1})!}{k_r!(n-k_1-\dots-k_{r-1}-k_r)!} \\ &= \frac{n!}{k_1!...k_r!} \end{aligned}$$

ways.

Example 1.19. If we are to divide a collection of 5 blue balls and 15 red balls into 2 sets of 7 balls and 3 sets of 2 balls, find

$$P(\text{each set has a blue ball}) = \frac{5!C(15,6,6,1,1,1)}{C(20,7,7,2,2,2)}$$

Exercise 1.20. If 8 rooks are placed on the chessboard randomly, what is the probability that they are safe from each other? Hint: to place them so that they are safe from each other, go row by row.

Exercise 1.21. Compute the coefficient for $a^k b^{n-k}$ in the expansion of $(a+b)^n$. Hint: it is $C(n,k)$ for the number of ways to choose k a's and $n-k$ b's in $(a+b)(a+b)...(a+b)(a+b)$.

Exercise 1.22. A poker hand has five cards. Calculate

a. the number of all poker hands.
b. the number of all straights such as 1,2,3,4,5 or 10, J, Q, K, A.
c. the number of all flushes such as $2\spadesuit, 4\spadesuit, 7\spadesuit, 8\spadesuit, 10\spadesuit$.
d. the number of all straight flushes such as $2\spadesuit, 3\spadesuit, 4\spadesuit, 5\spadesuit, 6\spadesuit$.
e. the number of all full houses such as $5\spadesuit, 5\diamondsuit, 5\clubsuit, 9\spadesuit, 9\heartsuit$.
f. the number of all 4 of a kinds such as $6\spadesuit, 6\diamondsuit, 6\clubsuit, 6\heartsuit, 7\spadesuit$.
g. rank these hands by their probabilities.

2. Classical Probability

2.1. **Discrete and Continuous Probability Models.**

We begin with an elementary example.

Example 2.1. Suppose we toss a coin. Then we will get either head or tail so the set of all possible outcomes is $\Omega = \{H, T\}$. Some people will say the chance of getting each is $\frac{1}{2}$, which is equivalent to defining a map

$$\begin{aligned} \Omega &\xrightarrow{P} [0,1] \\ H &\mapsto \frac{1}{2} \\ T &\mapsto \frac{1}{2} \end{aligned}$$

Some other people will say the chance of getting H is $\frac{1}{3}$ and the chance of getting T is $\frac{2}{3}$, which is equivalent to defining a map

$$\begin{aligned} \Omega &\xrightarrow{P} [0,1] \\ H &\mapsto \frac{1}{3} \\ T &\mapsto \frac{2}{3} \end{aligned}$$

Neither group is wrong, only whose model best describes reality when we toss the coin 1000 times matters.

In probability theory, the set X of all possible outcomes is changed to Ω and called a sample space. The elements x are changed to w and called outcomes. The subsets U are changed to A and called events.

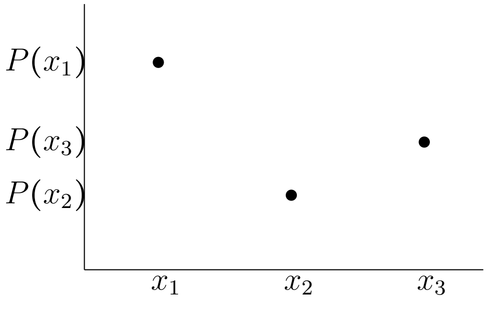
Definition 2.2. Given a sample space Ω, a map $\mathbb{P}(\Omega) \xrightarrow{P} [0,1]$ is called a probability law if it satisfies

1. $P(\Omega) = 1$
2. $P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} P(A_i)$ if the A_i are disjoint.

These two axioms imply additional properties that can be verified by Venn diagrams

3. $P(\varnothing) = 0$
4. $P(A^c) = 1 - P(A)$.
5. $P(A) \leq P(B)$ if $A \subseteq B$
6. $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$
7. $P(A \bigcup B) \leq P(A) + P(B)$

Property (2) is called countable additivity, it implies finite additivity as we can use $A_{n+1} = A_{n+2} = \ldots = \varnothing$ for any n. Together (Ω, P) is called a probability model, sometimes we drop P and just write Ω. When Ω is finite or countably infinite, we call (Ω, P) a discrete model, in which case we have the assignment $x_i \mapsto P(x_i)$. Each $P(x_i)$ is called the probability mass of x_i and we can graph them.



Example 2.3. When we toss a coin, the sample space is $\Omega = \{H, T\}$, the outcomes are H and T, and the events are $\varnothing, \{H\}, \{T\}$ and $\{H, T\}$. If the coin is fair, we can define a

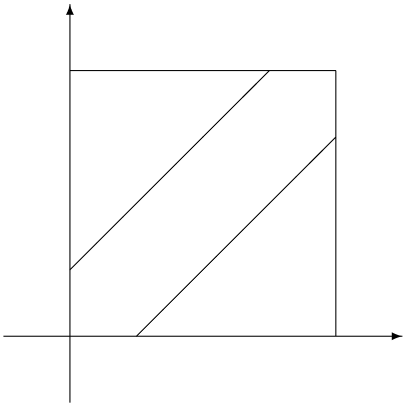
probability law with $P(\varnothing) = 0, P(H) = 1/2, P(T) = 1/2$ and $P(\{H,T\}) = 1$. If the coin is unfair, we can replace $1/2$ by some $0 \leq p \leq 1$.

Example 2.4. Toss a fair coin three times, create a probability model to find $P(\text{first two tosses are tails})$ and $P(\text{at least one toss is tail})$. Hint: write out all outcomes in Ω. Since the coin is fair, all outcomes should have the same probability $\frac{1}{8}$.

So some probability laws are uniform, i.e. all outcomes have the same probability. When Ω is finite, $P(A) = \frac{|A|}{|\Omega|}$ for any event $A \in \mathbb{P}(\Omega)$. This formula suggests how to define probability law for a more general Ω. Consider throwing dart at a board. Suppose the dart is equally likely to land on any point on the board. Drawing from the case of finite Ω, we can define $P(\text{event}) = \frac{\text{area of event}}{\text{area of board}}$. Surely $P(\text{a point}) = 0$ but this does not imply $P(\Omega) = 0$ by countable additivity because Ω is uncountable.

Example 2.5. Two people plan to meet. Each could be late for up to 1 hour and if the other has to wait for more than 15' then that person will leave. Compute

$$P(\text{they actually meet}) = P((x,y) \mid |x-y| \leq \frac{1}{4}, 0 \leq x, y \leq 1)$$



Exercise 2.6. Show that if $P(\text{meeting at any time}) > 0$ then $P(\Omega) = \infty$ and P is an invalid probability law. More generally, show that if $\{x_\alpha\}_{\alpha\in A}$ is an uncountable collection of positive real numbers then $\sum\limits_{\alpha \in A} x_\alpha = \infty$. Hint: consider $A_n = \{x_\alpha \geq \frac{1}{n}\}$.

2.2. **Conditional Probability.** Given (Ω, P), an inherent problem in probability theory is to determine the probability of one event given another event.

Example 2.7. Suppose we know the first person will show up between 0pm and 0:15pm. What is the probability that they will meet? If A is the diagonal strip and B is the leftmost vertical strip then via picture, this is

$$\frac{\text{area}(A \bigcap B)}{\text{area(B)}} = \frac{\text{area}(A \bigcap B)}{\text{area(square)}} \frac{\text{area(square)}}{\text{area(B)}} = \frac{P(A \bigcap B)}{P(B)}$$

This suggest to us a definition of probability law given an event B.

Definition 2.8. Given probability model (Ω, P) and event $B \in \mathbb{P}(\Omega)$, we define the probability law given B as

$$\mathbb{P}(\Omega) \xrightarrow{P(\ |B)} [0,1]$$
$$A \mapsto P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

We still need to verify this is a probability law,

1. $P(\Omega \,|\, B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$
2. $P(\bigcup A_i \,|\, B) = \frac{P((\bigcup A_i) \cap B)}{P(B)} = \frac{P(\bigcup (A_i \cap B))}{P(B)} = \frac{\sum P(A_i \cap B)}{P(B)} = \sum \frac{P(A_i \cap B)}{P(B)} = \sum P(A_i \,|\, B)$ for disjoint $\{A_i\}_{i \in I}$ in $\mathbb{P}(\Omega)$

Example 2.9. Toss a fair die twice and collect the outcomes. If A_i is the event that max equals i and B is the event that min equals 2, calculate $P(A_i \,|\, B)$.

Example 2.10. 500 patients test for a disease of likelihood 0.001. Unfortunately their blood samples are pooled together. Find

a. P(pool of blood tests positive). Hint: this is $1 - P$(pool of blood tests negative) $= 1 - P$(all healthy) $= 1 - 0.999^{500}$.
b. P(there are more than one carriers | pool of blood tests positive). Hint: name A = {more than one carriers} and B = {pool of blood tests positive} then B = {at least one carrier} and $A \bigcap B = A$. It is easier to calculate $P(B \backslash A)$ to get $P(A)$ than to calculate $P(A)$ directly.

Calculation of probability of some event via probability of its complement is one technique we often employ in probability theory. Here are some more.

2.3. **Multiplication Rule.** Another common problem in probability theory is to calculate the likelihood $P(\bigcap_{i=1}^{n} A_i)$ of intersection of events $A_1, \ldots, A_n$, especially when they seemingly occur in sequential order.

Proposition 2.11. *If (Ω, P) is a probability model and $A_1, \ldots, A_n$ is a finite sequence of events then*

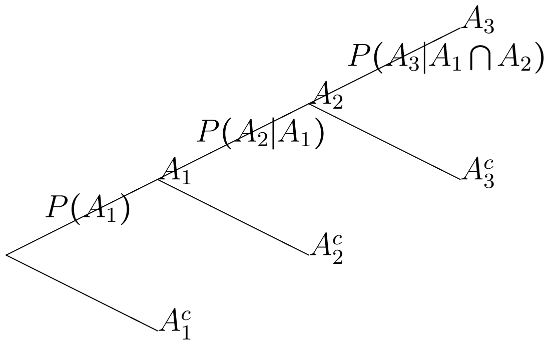
$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(A_2 \,|\, A_1)P(A_3 \,|\, A_1 \bigcap A_2) \cdots P(A_n \,|\, \bigcap_{i=1}^{n-1} A_i)$$

Proof. We move from right to left

$$P(A_1)P(A_2 \,|\, A_1)P(A_3 \,|\, A_1 \bigcap A_2) \cdots P(A_n \,|\, \bigcap_{i=1}^{n-1} A_i) = P(A_1)\frac{P(A_1 \cap A_2)}{P(A_1)}\frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdots \frac{P(\bigcap_{i=1}^{n} A_i)}{P(\bigcap_{i=1}^{n-1} A_i)}$$
$$= P(\bigcap_{i=1}^{n} A_i)$$

□

The picture looks like this



Example 2.12. Draw three cards out of the deck, what are the chances we get no heart?

a. Either we count all such hands and divide that number by the number of all triplets.
b. Or we let A_i be the event that the i card is not a heart and use multiplication rule.

Exercise 2.13. If the test in 2.10 is inaccurate with a 0.99 probability of returning positive when it is present and a 0.10 probability of returning positive when it is not present, calculate

a. P(false positive). Hint: let A_1 be the event of being healthy and A_2 be the event of testing positive.
b. P(false negative). Hint: let A_1 be the event of being sick and A_2 be the event of testing negative.

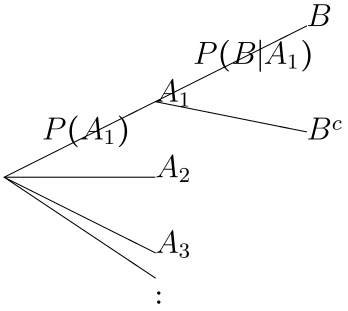
2.4. **Total Probability Law.** A common technique to calculate probability of some event B is to consider it in all the cases that cover Ω.

Proposition 2.14. *(total probability law) If* (Ω, P) *is a probability model and* $\Omega = \bigsqcup_{i=1}^{n} A_i$ *then for any event* B

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B\,|\,A_i)$$

Proof. Clearly $P(B) = P(B \bigcap \Omega) = P(B \bigcap (\bigsqcup_{i=1}^{n} A_i)) = P(\bigsqcup_{i=1}^{n}(B \bigcap A_i)) = \sum_{i=1}^{n} P(B \bigcap A_i) = \sum_{i=1}^{n} P(A_i)P(B\,|\,A_i)$. □

The picture looks like this



Example 2.15. You get to roll the die again if the first roll is less than 3. Calculate $P(\text{sum} \geq 4)$. Hint: let A_i be the event that the first roll is i and let B be the event that the sum is at least 4.

2.5. **Bayes' Rule.** Given a partition $\Omega = \bigsqcup_{i=1}^{n} A_i$, one common problem is to calculate the probability of A_i given some event B when the $P(B \mid A_i)$ are readily available.

Proposition 2.16. *(Bayes' rule) If* (Ω, P) *is a probability model and* $\Omega = \bigsqcup_{i=1}^{n} A_i$ *then for any event* B

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)}$$

Proof. We have

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + \cdots + P(A_n)P(B \mid A_n)}$$

by symmetry of intersection for the second equality and total probability law for the third equality. □

Example 2.17. Continuing with 2.10, calculate $P(\text{patient is carrier} \mid \text{blood tests positive})$. Hint: let A_1 be the event he is a carrier, A_2 be the event he is not a carrier, and B be the event his blood tests positive to use the given $P(A_i)$ and $P(B \mid A_i)$.

2.6. **Independence of Events.** Another common problem is to determine when two events A, B are independent. What we mean by independence is not even clear. Intuitively, we feel that A is independent of B if the probability of A does not change with knowledge of B, i.e. $P(A) = P(A \mid B) = P(A \mid B^c)$ and vice versa. But this means $P(A) = P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ and $P(B) = P(B \mid A) = \frac{P(B \cap A)}{P(A)}$. This leads us to choose the following definition for independence of events.

Definition 2.18. Two events A and B are said to be independent if $P(A \cap B) = P(A)P(B)$. More generally, events $A_1, A_2, \ldots, A_n$ are said to be independent if $P(\bigcap_{i \in U} A_i) = \prod_{i \in U} P(A_i)$ for any subset $U \subseteq \{1, 2, \ldots, n\}$.

Warning: this is far from the case A and B are disjoint.

Example 2.19. Roll a die twice, determine if the following events are independent,

a. first roll equals 1 and second roll equals 2. Hint: write out the sample space of all ij for $1 \le i, j \le 6$.
b. first roll equals 1 and sum equals 5.
c. max equals 2 and min equals 2.

Example 2.20. Again we toss a fair coin twice and let A_1 be the event that the first toss is tail, A_2 be the event that the second toss is tail and A_3 be the event that the two tosses are different. Determine if these events are independent.

Hint: $P(A_3 \,|\, A_1) = P(A_3) = \frac{1/4}{1/4} = \frac{1}{2}, P(A_3 \,|\, A_2) = \frac{1}{2}$, and $P(A_1 \bigcap A_2 \bigcap A_3) = 0 \neq \frac{1}{2}\frac{1}{2}\frac{1}{2} = P(A_1)P(A_2)P(A_3)$.

2.7. **Exercises.** pages 53-70: 8, 14, 16, 24, 30, 49, 58.

3. Measures and Measure Spaces

3.1. **Basic Concepts.** Discrete probability models require mostly combinatorial considerations. However, the presence of continuous probability models and other entities as in example 2.5 pushed probability theory toward analysis and so it was relaid in its modern foundations by Andrey Kolmogorov.

Definition 3.1. A measurable space $(X, \mathcal{F})$ is a set X together with a collection $\mathcal{F}$ of subsets of X that satisfies

1. $\mathcal{F}$ is nonempty.
2. (closure under complementation) if $U \in \mathcal{F}$ then $U^c \in \mathcal{F}$.
3. (closure under countable unions) if $\{U_i\}_{i \in \mathbb{N}} \in \mathcal{F}$ then $\bigcup\limits_{i \in \mathbb{N}} U_i \in \mathcal{F}$.

It follows from these axioms that $\mathcal{F}$ is also closed under countable intersection. The collection $\mathcal{F}$ is called a σ-algebra and the elements in $\mathcal{F}$ are called measurable, they are the subsets we will measure.

Example 3.2. Every nonempty σ-algebra $\mathcal{F}$ for X must contain some subset $U \subset X$, hence it contains $U^c, \varnothing$ and X. So the smallest σ-algebra for any set X is $\{\varnothing, X\}$. It is called trivial σ-algebra. Apparently the largest σ-algebra for any set X is $\mathbb{P}(X)$. It is called discrete σ-algebra.

Example 3.3. Every finite set $X = \{x_1, \ldots, x_n\}$ has a σ-algebra

$$\mathcal{F}_i = \{\varnothing, \{x_1, \ldots, x_i\}, \{x_{i+1}, \ldots, x_n\}, X\}$$

Example 3.4. (Borel σ-algebra) The Euclidean topology $\mathcal{E}$ on $\mathbb{R}$ induces a σ-algebra for $\mathbb{R}$ generated by all open sets through countable unions, countable intersections and relative complements. This σ-algebra is called Borel σ-algebra and denoted by $\mathcal{B}(X)$. More generally, if $\mathcal{T}$ is a topology for X then it induces a Borel σ-algebra $\mathcal{B}(\mathcal{T}) = \langle \mathcal{T} \rangle$ generated by open sets in $\mathcal{T}$ for X.

Definition 3.5. For two σ-algebras $\mathcal{F} \subset \mathcal{F}'$ of X, we say $\mathcal{F}'$ is a refinement of $\mathcal{F}$. A sequence $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \ldots$ of such σ-algebras is called a filtration.

Example 3.6. A filtration for $X = \{x_1, \ldots, x_6\}$ is $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \mathcal{F}_4$ where

$$\begin{aligned}
&\mathcal{F}_1 = \{\varnothing, X\}\\
&\mathcal{F}_2 = \{\varnothing, \{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}, X\}\\
&\mathcal{F}_3 = \{\varnothing, \{x_1, x_2, x_3, x_4\}, \{x_5, x_6\}, \{x_1, x_2\}, \{x_3, x_4\}, \{x_1, x_2, x_5, x_6\}, \{x_3, x_4, x_5, x_6\}, X\}\\
&\mathcal{F}_4 = \mathbb{P}(X)
\end{aligned}$$

Example 3.7. $\{\varnothing, \mathbb{R}\} \subset \mathcal{B}(\mathbb{R}) \subset \mathcal{L}(\mathbb{R}) \subset \mathbb{P}(\mathbb{R})$ is a filtration, where $\mathcal{L}(\mathbb{R})$ is the Lebesgue σ-algebra generated by all open and negligible sets in $\mathbb{R}$.

Once among measurable sets, we consider their relationships, i.e. maps between them that respect their σ-algebras.

Definition 3.8. A map $(X, \mathcal{F}) \xrightarrow{f} (Y, \mathcal{G})$ between measurable spaces is called a measurable map if $f^{-1}(V) \in \mathcal{F}$ for all $V \in \mathcal{G}$.

Surely composition of two measurable maps is measurable.

$$\begin{tikzcd}
(X,\mathcal{F}) \arrow[r, "f"] \arrow[dr, "gf"] & (Y,\mathcal{G}) \arrow[d, "g"'] \\
 & (Z,\mathcal{Z})
\end{tikzcd}$$

Warning: a topology $\mathcal{T}$ for X may not be a σ-algebra and a σ-algebra $\mathcal{F}$ may not be a topology. A continuous map $X \xrightarrow{f} Y$ may not a a measurable map and vice versa.

Example 3.9. Any constant map $(X, \mathcal{F}) \xrightarrow{c} (Y, \mathcal{G})$ is measurable, as $c^{-1}(V)$ is either X or $\varnothing$ for any $V \in \mathcal{G}$.

Example 3.10. Any map $(X, \mathbb{P}(X)) \xrightarrow{f} (Y, \mathcal{G})$ is measurable, as $f^{-1}(V) \in \mathbb{P}(X)$ for any $V \in \mathcal{G}$.

Example 3.11. If $(X, \mathcal{F}') \xrightarrow{f} (Y, \mathcal{G}')$ is a measurable map and $\mathcal{F} \subset \mathcal{F}' \subset \mathcal{F}''$ then $(X, \mathcal{F}'') \xrightarrow{f} (Y, \mathcal{G})$ is also measurable while $(X, \mathcal{F}) \xrightarrow{f} (Y, \mathcal{G})$ may not be measurable. Conversely, if $\mathcal{G} \subset \mathcal{G}' \subset \mathcal{G}''$ then $(X, \mathcal{F}) \xrightarrow{f} (Y, \mathcal{G})$ is also measurable while $(X, \mathcal{F}) \xrightarrow{f} (Y, \mathcal{G}'')$ may not be measurable.

Exercise 3.12. Given X_2 in example 3.6 and $Y_{10} = \{y_1, \ldots, y_{10}\}$ with σ-algebra $\mathbb{P}(Y_{10})$, determine if the following maps are measurable.

a.

$$\begin{aligned}
(X_2, \mathcal{F}_2) &\xrightarrow{f} (Y_{10}, \mathbb{P}(Y_{10}))\\
x_1, x_2, x_3, x_4 &\mapsto y_1\\
x_5, x_6 &\mapsto y_5
\end{aligned}$$

Hint: consider inverse image of four types of measurable subsets in $\mathbb{P}(Y_{10})$.

b.

$$\begin{aligned} (X_2, \mathcal{F}_2) &\xrightarrow{f} (Y_{10}, \mathbb{P}(Y_{10})) \\ x_1, x_2, x_3, x_4 &\mapsto y_1 \\ x_5 &\mapsto y_5 \\ x_6 &\mapsto y_{10} \end{aligned}$$

Exercise 3.13. Determine if a continuous map $(X, \mathcal{T}_X) \xrightarrow{f} (Y, \mathcal{T}_Y)$ induces a measurable map $(X, \mathcal{B}(\mathcal{T}_X)) \xrightarrow{f} (Y, \mathcal{B}(\mathcal{T}_Y))$.

We are now ready for a measure, which makes precise the notion of size.

Definition 3.14. A measure μ on $(X, \mathcal{F})$ is a map $\mathcal{F} \xrightarrow{\mu} [0, \infty]$ that satisfies

1. (countable additivity) $\mu\left(\bigsqcup\limits_{i=1}^{\infty} U_i\right) = \sum\limits_{i=1}^{\infty} \mu(U_i)$ for any countable disjoint $\{U_i\}_{i=1}^{\infty}$ in $\mathcal{F}$.
2. (negligible empty set) $\mu(\varnothing) = 0$.

If $\mu(X) < \infty$ then we say μ is a finite measure. If X is the countable union of measurable sets of finite measures then we say μ is σ-finite. Every finite measure is σ-finite. Each measurable set $U \in \mathcal{F}$ is called negligible if $\mu(U) = 0$ and it is called almost sure if $\mu(U) = \mu(X)$. By definition the empty set is negligible and the whole space is almost sure.

Definition 3.15. A measure space $(X, \mathcal{F}, \mu)$ is a measurable space $(X, \mathcal{F})$ equipped with a measure μ.

Example 3.16. If we define

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R})) &\xrightarrow{\mu_B} [0, \infty] \\ (a, b) &\mapsto b - a \end{aligned}$$

and extend it to all of $\mathcal{B}(\mathbb{R})$ then it is a σ-finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and called the Borel measure. For any measurable set $U \in \mathcal{B}(\mathbb{R})$,

$$\mu_B(U) = \int\limits_{\mathbb{R}} 1_U(s) ds = \int\limits_{U} ds$$

the usual integral in calculus. Recall the indicator function in example 1.9.

Example 3.17. The measure

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R})) &\xrightarrow{\nu} [0, \infty) \\ U &\mapsto \mu_B([a, b] \bigcap U) = \int\limits_{\mathbb{R}} 1_{[a,b] \cap U}(s) ds = \int\limits_{U} 1_{[a,b]}(s) ds = \int\limits_{[a,b] \cap U} ds \end{aligned}$$

is a finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Example 3.18. Consider the map

$$\begin{aligned}(\mathbb{R},\mathbb{P}(\mathbb{R})) &\xrightarrow{\mu_{\mathcal{L}^*}} [0,\infty]\\ A &\mapsto \inf\{\sum_{i=1}^{\infty} b_i - a_i \mid \bigcup_{i=1}^{\infty}(a_i,b_i) = A\}\end{aligned}$$

This μ_L^* is called the *Lebesgue outer measure* on $\mathbb{R}$. A set $U \subset \mathbb{R}$ is called Lebesgue measurable if $\mu_L^*(A) = \mu_L^*(A \bigcap U) + \mu_L^*(A \bigcap U^c)$ for all $A \subset \mathbb{R}$. One can verify that the set $\mathcal{L}(\mathbb{R}) = \{\text{all such Lebesgue measurable sets}\}$ form a σ-algebra. Lastly, if we define

$$\begin{aligned}(\mathbb{R},\mathcal{L}(\mathbb{R})) &\xrightarrow{\mu_L} [0,\infty]\\ U &\mapsto \mu_L^*(U)\end{aligned}$$

then it is a σ-finite measure on $(\mathbb{R},\mathcal{L}(\mathbb{R}))$ and called the Lebesgue measure. For any Borel measurable set $U \in \mathcal{B}(\mathbb{R}) \subset \mathcal{L}(\mathbb{R})$, the Lebesgue measure coincides with the Borel measure

$$\int_A d\mu_L = \mu_L(U) = \mu_B(U) = \int\limits_U ds$$

Many times we simply write a measure space $(X,\mathcal{F},\mu)$ as X and omit $\mathcal{F},\mu$. As this is a probability course, we are only interested in the case $\mu(X) = 1$. We call μ a probability measure and $(X,\mathcal{F},\mu)$ a probability space. The notation then becomes $(\Omega,\mathcal{F},P)$, Ω is called the sample space, the elements $w \in \Omega$ are called outcomes, and the elements $F \in \mathcal{F}$ are called events whose likelihoods are given by P. At this moment we realize measure theory formalizes and generalizes the notion of probability model in section 2. It allows for the case of incomplete knowledge, as $\mathcal{F}$ may not be $\mathbb{P}(X)$ and many events only become known over time.

Exercise 3.19. Prove the following statements.

a. (monotonicity) If $U_1 \subset U_2$ are measurable then $\mu(U_1) \leq \mu(U_2)$.

b. (subadditivity) If $\{U_i\}_{i=1}^{\infty}$ are measurable then show that $\mu\left(\bigcup\limits_{i=1}^{\infty} U_i\right) \leq \sum\limits_{i=1}^{\infty}\mu(U_i)$.

c. (continuity from below) If $U_1 \subset U_2 \subset U_3 \subset \dots$ are measurable then $\mu\left(\bigcup\limits_{i=1}^{\infty} U_i\right) = \lim\limits_{i\to\infty}\mu(U_i)$.

d. (continuity from above) If $U_1 \supset U_2 \supset U_3 \supset \dots$ are measurable and at least one U_i has finite measure then $\mu\left(\bigcap\limits_{i=1}^{\infty} U_i\right) = \lim\limits_{i\to\infty}\mu(U_i)$. This property is false without the assumption that one U_i has finite measure, for example consider $U_i = [i,\infty) \subset \mathbb{R}$ for $i \geq 1$.

e. If $\{U_i\}_{i=1}^{\infty}$ is a sequence of measurable set with $\sum\limits_{i=1}^{\infty}\mu(U_i) < \infty$ then the set of points which belong to infinitely many U_i has measure zero. Hint: show that this set is $\bigcap\limits_{n=1}^{\infty}(\bigcup\limits_{k\geq n} U_k)$.

f. Find a nonBorel set.

g. Find a nonmeasurable set.

In general, a measurable space $(X,\mathcal{F})$ will have more than one measure. A measure μ on $(X,\mathcal{F})$ is said to be dominated by another measure ν, written $\mu \ll \nu$, if $\mu(U) = 0$ whenever $\nu(U) = 0$ for $U \in \mathcal{F}$. And μ is said to be absolutely continuous with respect to ν if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $\mu(U) < \epsilon$ whenever $\nu(U) < \delta$ for $U \in \mathcal{F}$. The equivalence between domination and absolute continuity is shown in the next proposition for its own sake and has little to do with our course.

Proposition 3.20. *A finite measure μ on $(X,\mathcal{F})$ is dominated by ν if and only it is absolutely continuous with respect to ν.*

Proof. Suppose μ is absolutely continuous with respect to ν. If $\mu(U) > 0$ for $U \in \mathcal{F}$, take $\epsilon \in (0,\mu(U)]$ then there exists a $\delta > 0$ such that $\nu(U) > \delta$. Hence $\mu(U) = 0$ whenever $\nu(U) = 0$ and μ is dominated by ν. For the converse, suppose μ is not absolutely continuous with respect to ν. Then there exists an $\epsilon > 0$ such that for every $\delta_n = \frac{1}{2^n}, n \geq 1$ there exists $U_n \in \mathcal{F}$ with $\nu(U_n) < \delta_n$ and yet $\mu(U_n) > \epsilon$. Now $\nu(\bigcup\limits_{k\geq n} U_k) \leq \sum\limits_{k\geq n} \nu(U_k) < \sum\limits_{k\geq n} \delta_n = \sum\limits_{k\geq n} \frac{1}{2^k} = \frac{1}{2^{n-1}}$. So $\nu(\bigcap\limits_{n\geq 1}(\bigcup\limits_{k\geq n} U_k)) = \lim\limits_{n\to\infty} \nu(\bigcup\limits_{k\geq n} U_k) < \lim\limits_{n\to\infty} \frac{1}{2^{n-1}} = 0$ by continuity from above in 3.19. Meanwhile $\mu(\bigcap\limits_{n\geq 1}(\bigcup\limits_{k\geq n} U_k)) = \lim\limits_{n\to\infty} \mu(\bigcup\limits_{k\geq n} U_k) \geq \limsup\limits_{n\to\infty}\{\mu(U_k), k \geq n\} \geq \lim\limits_{n\to\infty} \epsilon = \epsilon$ again by continuity from above in 3.19 and μ is not dominated by ν. □

The converse does not hold without the assumption of finiteness. Only probability measures are considered in this course, for which domination is the same as absolute continuity. We present an important theorem what will be used later.

Theorem 3.21. *(Radon-Nikodym theorem) If a σ-finite measure μ on $(X,\mathcal{F})$ is absolutely continuous with respect to another σ-finite measure ν then there exists a nonnegative measurable function $(X,\mathcal{F}) \xrightarrow{f} ([0,\infty),\mathcal{B}([0,\infty)))$ such that $\mu(U) = \int\limits_U f d\nu$ for any $U \in \mathcal{F}$. The function f is unique up to a ν-negligible set in X.*

Proof. literature. □

In Leibitz's notation, people will write $d\mu = f d\nu$ and $f = \frac{d\mu}{d\nu}$ will be called the Radon-Nikodym derivative of μ with respect to ν. Note that the theorem does not hold without σ-finiteness for ν.

3.2. **Conditional Measure.**

Given a measure space $(X,\mathcal{F},\mu)$ and an event $G \in \mathcal{F}$, we want to define conditional measure of an event $F \in \mathcal{F}$ given G. We do this in similar fashion to how we defined conditional probability of an event A given an event B in 2.8.

Definition 3.22. For $(X,\mathcal{F},P)$ and $G \in \mathcal{F}$, we define the conditional probability measure given G as

$$\mathcal{F} \xrightarrow{P(\ |G)} [0,1]$$
$$F \mapsto P(F|G) = \frac{P(F\bigcap G)}{P(G)}$$

As with conditional probability law, one can verify that this is a measure on $(X,\mathcal{F})$.

3.3. **Change of Measures.**

If $(X,\mathcal{F},\mu)$ is a measure space and $(X,\mathcal{F}) \xrightarrow{f} (Y,\mathcal{G})$ is a measurable map then we can define a measure

$$(Y,\mathcal{G}) \xrightarrow{f_*(\mu)} [0,\infty]$$
$$V \mapsto f_*(\mu)(V) = \mu(f^{-1}(V))$$

$$\begin{tikzcd}
(X,\mathcal{F}) \arrow[r, "f"] \arrow[dr, "\mu"'] & (Y,\mathcal{G}) \arrow[d, "f_*(\mu)"] & f^{-1}(V) \arrow[dr, "\mu"'] & & V \arrow[ll, "f^{-1}"'] \arrow[d, "f_*(\mu)"] \\
 & {[0,\infty]} & & & \mu(f^{-1}(V)) = f_*(\mu)(V)
\end{tikzcd}$$

This new measure $f_*(\mu)$ is called the pushforward of μ by f.

Example 3.23. If $(\Omega,\mathcal{F},P)$ is a probability space and $(\Omega,\mathcal{F}) \xrightarrow{f} (\mathbb{R},\mathcal{L}(\mathbb{R}))$ is a measurable function then $f_*(P)$ is a probability measure on $(\mathbb{R},\mathcal{L}(\mathbb{R}))$ beside μ_L.

$$\begin{tikzcd}
(\Omega,\mathcal{F}) \arrow[r, "f"] \arrow[dr, "P"'] & (\mathbb{R},\mathcal{L}(\mathbb{R})) \arrow[d, shift right, "\mu_L"'] \arrow[d, shift left, "f_*(P)"] \\
 & {[0,1]}
\end{tikzcd}$$

Exercise 3.24. Show that $f_*(\mu)$ is indeed a measure on $(Y,\mathcal{G})$ and conclude that $(Y,\mathcal{G},f_*(\mu))$ is a measure space.

Exercise 3.25. If $(X,\mathcal{F}) \xrightarrow{f} (Y,\mathcal{G})$ is measurable and $(Y,\mathcal{G}) \xrightarrow{\mu} [0,\infty]$ is a measure, show that the composition $(X,\mathcal{F}) \xrightarrow{f^*(\mu)} [0,\infty], U \mapsto \mu(f(U))$ is not a measure. Unlike pushforward, the pullback $f^*(\mu)$ of a measure is not a measure.

$$\begin{tikzcd}
(X,\mathcal{F}) \arrow[r, "f"] \arrow[dr, "f^*(\mu)"'] & (Y,\mathcal{G}) \arrow[d, "\mu"] & U \arrow[rr, "f"] \arrow[drr, "f^*(\mu)"'] & & f(U) \arrow[d, "\mu"] \\
 & {[0,\infty]} & & & f^*(\mu)(U) = \mu(f(U))
\end{tikzcd}$$

4. Random Variables

In many probability models, the outcomes are either numerical or associated with numerical values. For example, in 2.19 the event A of first roll equaling 1 is not a number, rather it is a subset of outcomes $\{11,12,13,14,15,16\}$ which is attributed a number 1. As such, we realize $P(A) = P(f^{-1}(1)) = P(\{11,12,13,14,15,16\})$ if f is the function $\Omega \xrightarrow{f} \{1,2,3,4,5,6\}, ij \mapsto i$. We generalize this situation.

Definition 4.1. An S-valued random variable is a measurable map $(\Omega, \mathcal{F}, P) \xrightarrow{X} (S, \mathcal{S})$ between a probability space $(\Omega, \mathcal{F}, P)$ and a measurable space $(S, \mathcal{S})$.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"'] & (S, \mathcal{S}) \\
{[0,1]} &
\end{tikzcd}
\qquad
\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
{[0,1]} &
\end{tikzcd}$$

Most common is the case $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then X is simply called random variable. Note that the domain of X is Ω while the domain of P is $\mathcal{F}$; we often abuse notations.

Example 4.2. Any $c \in \mathbb{R}$ defines a constant random variable X

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
{[0,1]} &
\end{tikzcd}
\qquad
\begin{tikzcd}
w \arrow[rr, "X"] & & c
\end{tikzcd}$$

Example 4.3. We can model the game of tossing a coin and winning \$1 for each head and losing \$1 for each tail as

$$\begin{aligned}
\Omega &= \{H, T\} \\
\mathcal{F} &= \mathbb{P}(\Omega) \\
P(\varnothing) = 0, P(\Omega) &= 1, P(H) = p, P(T) = 1 - p \\
(\Omega, \mathcal{F}, P) &\xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
H &\mapsto 1 \\
T &\mapsto -1
\end{aligned}$$

Then

$$P(\text{winning } \$1) = P(X = 1) = P(X^{-1}(1)) = P(H) = p$$

while

$$P(\text{winning } \$7) = P(X = 7) = P(X^{-1}(7)) = P(\varnothing) = 0$$

Example 4.4. These previous examples generalize to the indicator function

$$(\Omega, \mathcal{F}, P) \xrightarrow{1_F} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

$$\omega \mapsto \begin{cases} 1 \text{ if } \omega \in F \\ 0 \text{ if } \omega \notin F \end{cases}$$

to identify an event $F \in \mathcal{F}$ and its complement. One can verify that 1_F is measurable, hence it is a random variable with

$$P(1_F = 1) = P(1_F^{-1}(1)) = P(F)$$

while

$$P(1_F = 0) = P(F^c)$$

Example 4.5. The constant random variable $(\Omega, \mathcal{F}, P) \xrightarrow{c} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in example 4.2 can be written as $c \cdot 1_\Omega$ since $c \cdot 1_\Omega(\omega) = c \cdot 1 = c$.

While these random variables are simple, others are trickier.

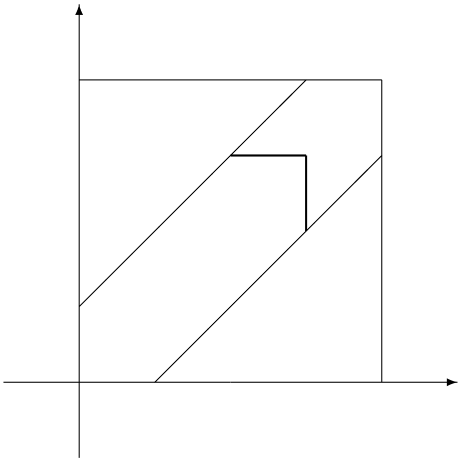
Example 4.6. If X is to mark the meeting time of two people in example 2.5 then we must set up

$$\begin{aligned} \Omega &= [0,1] \times [0,1] \\ \mathcal{F} &= \mathcal{B}([0,1] \times [0,1]) \\ P &= P_{\mathcal{B}} \text{ the Borel measure} \end{aligned}$$

$$\begin{aligned} (\Omega, \mathcal{F}, P_{\mathcal{B}}) &\xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ (x,y) &\mapsto \max(x,y) \end{aligned}$$

Then

$$\begin{aligned} P(\text{meeting at 0:45}) &= P(X = 0.75) \\ &= P(X^{-1}(0.75)) \\ &= P(\text{two line segments in picture}) \\ &= \text{area of two line segments} \\ &= 0 \end{aligned}$$



It is worth noting that linear combinations of random variables from the same sample space to $\mathbb{R}$ are also random variables. That means the set of all random variables form a vector space.

Exercise 4.7. Show that the set $RV((\Omega, \mathcal{F}, P), (\mathbb{R}, \mathcal{B}(\mathbb{R})))$ of all random variables from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ form a vector space over $\mathbb{R}$.

Example 4.8. If two people are at the betting table in example 4.3 then their combined winning against the house is

$$\begin{aligned} (\Omega, \mathbb{P}(\Omega), P) &\xrightarrow{X_1+X_2} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ H &\mapsto X_1(H) + X_2(H) = 2 \\ T &\mapsto X_1(T) + X_2(T) = -2 \end{aligned}$$

If the first player raises his stake to \$6 per toss and the second person lowers his bet to 70 cents per toss then their winning is $(\Omega, \mathbb{P}(\Omega), P) \xrightarrow{6X_1+0.7X_2} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Each random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ induces a probability measure

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R})) &\xrightarrow{X_*(P)} [0,1] \\ A &\mapsto X_*(P)(A) = P(X^{-1}(A)) \end{aligned}$$

This lets us define a function that watches the accumulation of probability.

Definition 4.9. For a random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we define its cumulative distribution function to be

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R})) &\xrightarrow{F_X} [0,1] \\ x &\mapsto X_*(P)((-\infty, x]) = P(X^{-1}((-\infty, x])) = P(\{\omega \,|\, X(\omega) \le x\}) \end{aligned}$$

$$\begin{tikzcd} (\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[dl, "F_X"'] \\ {[0,1]} & \end{tikzcd} \qquad \begin{tikzcd} X^{-1}((-\infty, x]) \arrow[d, "P"'] & x \arrow[dl, "F_X"'] \\ P(X^{-1}((-\infty, x])) & \end{tikzcd}$$

This cumulative distribution function is abbreviated as CDF. It is surely nondecreasing because $P(X \in (-\infty, x]) \le P(X \in (-\infty, x'])$ if $x < x'$. We will use the notations $\{\omega, X(\omega) \le x\}, \{X^{-1}((-\infty, x])\}, \{X \le x\}$ interchangeably.

Example 4.10. If X models the game of tossing coin in example 4.3 then

$$F_X(x) = \begin{cases} 0 \text{ if } -\infty < x < -1 \\ 1 - p \text{ if } -1 \le x < 1 \\ 1 \text{ if } x \ge 1 \end{cases}$$

Example 4.11. Back to example 4.6

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P(\text{they meet at or before } x) \\ &= P(\{(u,v), \max\{u,v\} \le x) \\ &= \text{area of part of the strip below the two segments} \end{aligned}$$

In particular, $F_X(0.75) = \frac{10}{32}$.

Proposition 4.12. *Every cumulative distribution function F_X has the following properties*
1. F_X is nondecreasing.
2. $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to+\infty} F_X(x) = 1$.
3. F_X is right continuous everywhere.
4. F_X is left continuous at x iff $X_*(P)(\{x\}) = 0$.

Proof. Proving (1) and (2) is straightforward. On one hand

$$\begin{aligned} F_X(x) &= P(X^{-1}((-\infty, x]))) \\ &= P(X^{-1}(\bigcap_{x'\to x^+} (-\infty, x'])) \\ &= P(\bigcap_{x'\to x^+} X^{-1}((\infty, x'])) \\ &= \lim_{x'\to x^+} P(X^{-1}((-\infty, x'])) \\ &= \lim_{x'\to x^+} F_X(x') \end{aligned}$$

by continuity from above in 3.19. So F_X is right continuous everywhere. On the other hand

$$\begin{aligned} F_X(x) &= P(X^{-1}((-\infty, x])) \\ &= P(X^{-1}((-\infty, x)\bigsqcup\{x\})) \\ &= P(X^{-1}(-\infty, x)) + P(X^{-1}(\{x\})) \\ &= P(X^{-1}(\bigcup_{x'\to x^-} (\infty, x'])) + X_*(P)(x) \\ &= P(\bigcup_{x'\to x^-} X^{-1}((\infty, x'])) + X_*(P)(x) \\ &= \lim_{x'\to x^-} P(X^{-1}((-\infty, x'])) + X_*(P)(x) \\ &= \lim_{x'\to x^-} F(x') + X_*(P)(x) \end{aligned}$$

by continuity from below in 3.19. So $F_X(x) = \lim_{x'\to x^-} F(x')$ and F_X is left continuous iff $X_*(P)(\{x\}) = 0$. □

Together, (3) and (4) mean F_X is continuous iff $X_*(P)(\{x\}) = 0$ for all x. We will use continuity of F_X to define and study two classes of random variables.

Associated with each random variable X are two sequences of invariants that shed much insight into its behavior.

Definition 4.13. For a random variable X, we define its n^{th} moment, $n \geq 0$ to be

$$\mu_n(X) = \int_{\mathbb{R}} x^n dX_*(P) = \int_{\Omega} X(\omega)^n dP$$

if the integrals converge absolutely.

The zero$^{\text{th}}$ moment is 1. The first moment is also called mean and denoted by μ_X or called expected value and denoted by $E(X)$. If Ω has some sort of geometric shape then one sees that $E(X)$ is the center of gravity for the mass under X over Ω.

Definition 4.14. For a random variable X, we define its n^{th} central moment, $n \geq 0$ to be

$$\mu_n(X,\mu) = \int_{\mathbb{R}} (x - \mu_X)^n dX_*(P) = \int_{\Omega} (X(\omega) - \mu_X)^n dP$$

if μ_X exists and if the integrals converge absolutely.

Again the zero$^{\text{th}}$ moment is 1. The first central moment is 0. The second central moment is also called variance and denoted by $\text{var}(X)$. One can see that $\text{var}(X) = E((X - \mu_X)^2)$ the first moment of $(X - \mu_X)^2$. Its square root is called standard deviation and denoted by $\text{sd}(X)$ or σ_X.

In general, we can define the n^{th} moment of X around any number a to be

$$\mu_n(X,a) = \int_{\mathbb{R}} (x - a)^n dX_*(P) = \int_{\Omega} (X(\omega) - a)^n dP$$

Among these, we will care about mean, variance and standard deviation the most, while we will care about other moments much later and we will not care about other central moments or moments about $a \neq 0$ at all. Though their definitions require understanding integration over a general measure space, they are easy for us to give now and they are here for us to come back to when we stray too far later.

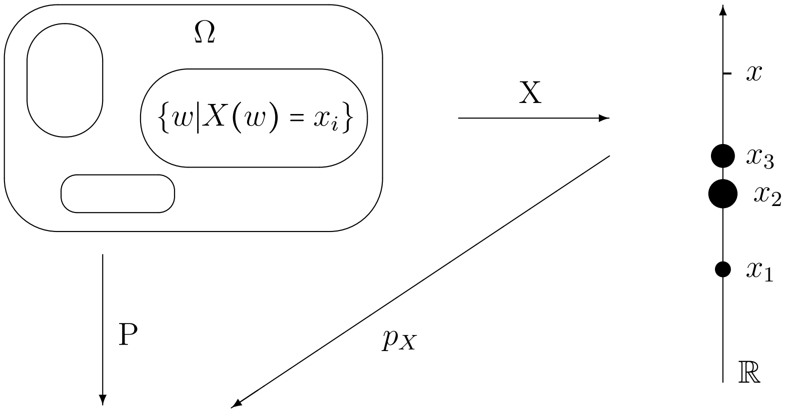
4.1. **Discrete Random Variable.** If X takes countably many values $x_i, i \in \mathbb{N}^+$ then $1 = P(\Omega) = P(X^{-1}\{x_i, i \in \mathbb{N}^+\})) = \sum_i P(X^{-1}(\{x_i\}) = \sum_i X_*(P)(\{x_i\})$ implies $X_*(P)(\{x_i\}) > 0$ for some i and so F_X is discontinuous by proposition 4.12. We define our first class of random variables.

Definition 4.15. A random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called discrete if it takes at most countably many values x_i in $\mathbb{R}$.

If X is discrete then the function

$$\begin{aligned}&\mathbb{R} \xrightarrow{p_X} [0,1]\\ &x \mapsto p_X(x) = P(X = x) = P(X^{-1}(x)) = P(\{w \,|\, X(w) = x)\end{aligned}$$

is 0 everywhere except over those x_i. It is called probability mass function of X and abbreviated as pmf. One sees that the concept of probability mass has been applied to numerical values, so that each x now has probability mass $P(\{w \in \Omega\} \,|\, X(w) = x)$ while each w may or may not have probability mass as defined in 2.1. Here is one visualization



$[0,1] \ni p_X(x_i) = P(\{w\,|\,X(w) = x_i\})$ $\qquad$ $p_X(x_i) =$ mass at x_i

The probability mass function p_X completely determines the cumulative density function F_X via $F_X(x) = P(X \le x) = \sum_{x_i \le x} p_X(x_i)$ and vice versa.

Example 4.16. If a die has two 4's and no 6 and X scores its toss then

$$p_X(x) = \begin{cases} \frac{1}{6}, x = 1 \\ \frac{1}{6}, x = 2 \\ \frac{1}{6}, x = 3 \\ \frac{2}{6}, x = 4 \\ \frac{1}{6}, x = 5 \\ 0, \text{ all other } x \end{cases} \qquad F_X(x) = \begin{cases} 0, x \in (-\infty, 1) \\ \frac{1}{6}, x \in [1,2) \\ \frac{2}{6}, x \in [2,3) \\ \frac{3}{6}, x \in [3,4) \\ \frac{5}{6}, x \in [4,5) \\ 1, x \in [5, \infty) \end{cases}$$

Following definition 4.13 and definition 4.14, we get

$$\mu_1(X) = \sum_i x_i p_X(x_i)$$
$$\mu_n(X) = \sum_i x_i^n p_X(x_i), n > 1$$
$$\mathrm{var}(X) = \sum_i (x_i - \mu_X)^2 p_X(x_i)$$
$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$

If one places a weight $p_X(x)$ at each x then mean $E(X)$ is the center of mass of $\mathbb{R}$ above while variance var(X) is the mean of the squared deviation of X from this center of mass and standard deviation measures the dispersion of X around its mean. In calculating the n^{th} moment of X we raise the moment arm x to the n^{th} power. Why did we not have $P(X^n = x_i^n)$ in the definition of n^{th} moment of X?

Example 4.17. If the coin in example 4.3 is fair and one wins \$1 for every head with probability $p = \frac{1}{2}$ and loses \$1 for every tail with probability $1 - p = \frac{1}{2}$ then $\mu = 0$ and $var = \frac{3}{4}$. If one wins \$1 for every head and loses \$0 for every tail then $\mu = \frac{1}{2}$ while $var = \frac{1}{4}$. In the second case, the variance is smaller because there is less deviation from the mean.

Example 4.18. For an event $F \in \mathcal{F}$ the indicator function 1_F in example 4.4 has

$$\begin{aligned} E(1_F) &= 1P(1_F = 1) + 0P(1_F = 0) \\ &= P(F) \\ \operatorname{var}(1_F) &= (1 - P(F))^2 P(F) + (0 - P(F))^2 P(F^c) \\ &= (1 - P(F))^2 P(F) + (0 - P(F))^2 (1 - P(F)) \\ &= P(F)(1 - P(F)) \\ F_{1_F} &= \begin{cases} P(F^c) \text{ if } x \in (-\infty, 1) \\ 1 \text{ if } x \in [1, \infty) \end{cases} \end{aligned}$$

For any random variable X, if we choose $F = \{X \le x\}$ then $E(1_F) = P(F) = P(X \le x) = F_X(x)$. Hence we have equal maps $\mathbb{R} \underset{F_X(-)}{\overset{E(1_{X \le -})}{\rightrightarrows}} \mathbb{R}$.

Exercise 4.19. Let X be a discrete random variable with cumulative distribution function

$$F_X(x) \begin{cases} 0 \text{ if } x \in (-\infty, 0) \\ 0.1 \text{ if } x \in [0, 1) \\ 0.3 \text{ if } x \in [1, 2) \\ 0.8 \text{ if } x \in [2, 3) \\ 0.9 \text{ if } x \in [3, 5) \\ 1 \text{ if } x \in [5, \infty) \end{cases}$$

a. Graph $F_X(x)$. What contributes to the jumps?
b. Find pmf p_X and plot it. Now one can view this as tossing a die with one face of 0, two faces of 1, five faces of 2, one face of 3, and one face of 5.
c. Find $E(X)$ and $\operatorname{var}(X)$.
d. Find $E((X - 0.5)(X + 0.5))$.

4.2. **Continuous Random Variable.** If X takes uncountably many values then either $X_*(P)(x) = 0$ for all x or $X_*(P)(\{x_i\}) \ne 0$ for at most countably many x_i. The second scenario would mean F_X is discontinuous. So having uncountable image is a necessary but not sufficient condition for continuity of F_X. We define our other class of random variables.

Definition 4.20. A random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{L}(\mathbb{R}))$ is called continuous if its cumulative distribution function F_X is continuous.

General continuous random variables are hard to work with unless we know more about them.

Definition 4.21. A continuous random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{L}(\mathbb{R}))$ is called absolutely continuous if $X_*(P)$ is absolutely continuous with respect to μ_L.

By Radon-Nikodym theorem 3.21, there exists a nonnegative measurable function $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{f} ([0, \infty), \mathcal{B}([0, \infty)))$ such that $X_*(P)(A) = \int_A f(s)ds$ for all $A \in \mathcal{B}(\mathbb{R})$.

This Radon Nikodym derivative $f = \frac{dX_*(P)}{d\mu_L}$ is called the probability density function of X, abbreviated as pdf and denoted by f_X. Now the probabilities of events $\{\omega, X(\omega) \in A\}$ are given as areas under a curve f_X over A

$$P(X \in A) = X_*(P)(A) = \int\limits_A f_X(s)ds$$

and this is what makes absolutely continuous random variables more tractable than general continuous random variables, especially since we are more familiar with integration of functions over $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L)$. In probability theory and in this course, when we say continuous random variables, we mean absolutely continuous random variables equipped with probability density functions f_X.

Surely $X_*(P)(\{x\}) = \int\limits_{\{x\}} f_X(s)ds = 0$ while $P(\Omega) = X_*(P)(\mathbb{R}) = \int\limits_{\mathbb{R}} f_X(s)ds = 1$. Furthermore, the probability density function f_X completely determines the cumulative density function F_X and vice versa

$$F_X(x) = P(X \leq x) = \int\limits_{(-\infty,x]} f_X(s)ds$$

Following definition 4.13 and definition 4.14, we get

$$\mu_1(X) = \int\limits_{\mathbb{R}} x dX_*(P) = \int\limits_{\mathbb{R}} s f_X(s)ds$$

$$\mu_n(X) = \int\limits_{\mathbb{R}} x^n dX_*(P) = \int\limits_{\mathbb{R}} s^n f_X(s)ds$$

$$\mathrm{var}(X) = \int\limits_{\mathbb{R}} (x - \mu_X)^2 dX_*(P) = \int\limits_{\mathbb{R}} (s - \mu_X)^2 f_X(s)ds$$

$$\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$$

Analogous to the discrete case, $E(X)$ is the s-coordinate of the center of mass of the area of uniform density under the curve f_X.

Remark 4.22. There are random variables X that are neither discreet nor continuous.

Exercise 4.23. Let X be a continuous random variable with probability density function

$$f_X(s) = \begin{cases} 0 \text{ if } x \in (\infty, 0) \\ s \text{ if } x \in [0,1) \\ cs \text{ if } x \in [1,3) \\ 0 \text{ if } x \in [3,\infty) \end{cases}$$

a. Graph f_X.
b. Find c so that f_X is indeed a probability density function.
c. Find $F_X(5), F_X(7)$ and $P(X^2 - 12X + 35 > 0)$.
d. Find $E(X)$ and $\mathrm{var}(X)$.

4.3. **Common Properties.** Now we delve more into the properties of both discrete and continuous random variables.

Proposition 4.24. *The following properties hold for both discrete and continuous random variables.*

1. (constancy) $E(c) = c$.
2. (monotonicity) $E(X) \leq E(Y)$ *if* $X \leq Y$ *almost everywhere.*
3. (absolute value) $|E(X)| \leq E(|X|)$.
4. (linearity) $E(aX + bY) = aE(X) + bE(Y)$.
5. $var(X) = E(X^2) - E(X)^2$.
6. $var(aX + b) = a^2 var(X)$.

Proof. Straightforward. □

Exercise 4.25. Prove proposition 4.24

Example 4.26. If $Y = \frac{X_1}{7} + \frac{X_2}{8} + b$ then

$$\mu(Y) = \frac{E(X_1)}{7} + \frac{E(X_2)}{8} + b$$

$$\text{var}(Y) = var\left(\frac{X_1}{7} + \frac{X_2}{8}\right) = \frac{\text{var}(X_1)}{7^2} + \frac{\text{var}(X_2)}{8^2}$$

Example 4.27. If the X_i are independent identically distributed random variables with mean μ and standard deviation σ then their average $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$ has

$$\mu(\overline{X}_n) = \mu$$

$$\text{var}(\overline{X}_n) = \frac{1}{n^2}\text{var}(X_1 + \cdots + X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\sigma(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$$

One sees that the variance of the mean decreases when n increases. Imagine n players sitting at the coin toss table in example 4.3 and $\overline{X}$ their average winning.

To each pair $X, Y \in RV((\Omega, \mathcal{F}, P), (\mathbb{R}, \mathcal{B}(\mathbb{R})))$ we associate a quantity that speaks volumes about their relationship.

Definition 4.28. For two random variables $X, Y \in RV((\Omega, \mathcal{F}, P), (\mathbb{R}, \mathcal{B}(\mathbb{R})))$, we define their covariance to be $\text{cov}(X, Y) = E((X - E(X))(Y - E(Y)))$.

When $Y = X$ then

$$\begin{aligned}
\text{cov}(X, X) &= E((X - E(X))(X - E(X))) \\
&= E(X^2 - 2E(X)X + E(X)^2) \\
&= E(X^2) - 2E(X)E(X) + E(X)^2 \\
&= E(X^2) - E(X)^2 \\
&= \text{var}(X)
\end{aligned}$$

by properties in 4.24. What is more, covariance defines a map

$$\begin{aligned} RV((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))\times RV((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R}))) &\xrightarrow{\operatorname{cov}(-,-)} \mathbb{R} \\ (X,Y) &\mapsto \operatorname{cov}(X,Y) \end{aligned}$$

Exercise 4.29. Verify that covariance satisfies similar properties to those of an inner product on $RV((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$.
1. (symmetry) $\operatorname{cov}(X,Y)=\operatorname{cov}(Y,X)$.
2. (linearity) $\operatorname{cov}(aX+bY,Z)=a\operatorname{cov}(X,Z)+b\operatorname{cov}(Y,Z)$.
3. (positive semidefiniteness) $\operatorname{cov}(X,X)\geq 0$ with equality iff X is constant.

Therefore, $\operatorname{cov}(-,-)$ is not quite an inner product for $RV((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$. However, it is an inner product for the subspace $RVF((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$ of all random variables with finite second moment and zero mean. This inner product induces the norm $\|X\|=\sqrt{\operatorname{cov}(X,X)}=\sqrt{\operatorname{var}(X)}=\sigma_X$ on $RVF((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$.

Covariance also induces another useful quantity.

Definition 4.30. For two random variables X,Y with nonzero finite second moment, we define their correlation coefficient $\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}$.

Cauchy-Schwarz inequality says that $-1\leq\rho(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\langle X,Y\rangle}{\|X\|\|Y\|}\leq 1$. Hence we can use this correlation coefficient to define angle.

Definition 4.31. For two random variables X,Y with nonzero finite second moment, we define their angle $\angle(X,Y)=\arccos(\rho(X,Y))$.

When $\rho(X,Y)=1$ and $\angle(X,Y)=0^\circ$, X and Y are said to be correlated. When $\rho(X,Y)=0$ and $\angle(X,Y)=90^\circ$, X and Y are said to be uncorrelated. When $\rho(X,Y)=-1$ and $\angle(X,Y)=180^\circ$, X and Y are said to be anticorrelated.

Example 4.32. Let X be the number of heads and Y be the number of tails in n coin tosses. Then

$$\begin{gathered} X+Y=n \\ E(X)+E(Y)=n \\ Y-E(Y)=-(X-E(X)) \end{gathered}$$

Thus

$$\begin{aligned} \operatorname{cov}(X,Y) &= E((X-E(X))(Y-E(Y))) \\ &= E((X-E(X))(-(X-E(X)))) \\ &= -E((X-E(X))^2) \\ &= -\operatorname{var}(X) \end{aligned}$$

Hence

$$\begin{aligned} \rho(X,Y) &= \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} \\ &= \frac{-\operatorname{var}(X)}{\sqrt{\operatorname{var}(X)\operatorname{var}(X)}} \\ &= -1 \end{aligned}$$

and X, Y are anticorrelated as we would expect.

Below are some properties of variance and covariance.

Proposition 4.33. *All random variables satisfy*

1. $cov(X, Y) = E(XY) - E(X)E(Y)$.
2. $var(X_i + X_j) = var(X_i) + var(X_j) + 2cov(X_i, X_j)$. *More generally,* $var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(X_i) + \sum_{i \neq j} cov(X_i, X_j)$.
3. $var(aX + bY) = a^2 var(X) + b^2 var(Y) + 2ab cov(X, Y)$.

Proof. The first statement follows from straightforward expansion. The second statement follows from taking $\langle X_i + X_j, X_i + X_j \rangle$. The third statement follows from the second statement, proposition 4.24, and exercise 4.29. □

4.4. Exercises.

pages 122: 16, 21.
pages 184: 2.

4.5. Common Random Variables.

We now go through the most common random variables.

4.5.1. *Bernoulli Random Variable.* (X, p) counts 1 for a successful trial with likelihood p (oftentimes called parameter) and counts 0 otherwise with likelihood $1 - p$. Success here could be tossing a tail, picking a blue ball, winning a game or living past 60 years, etc. Then X is just an indicator function 1_F where $F = \{s\}$ in example 4.18 with

$$\begin{aligned} p_X(x) &= \begin{cases} p \text{ if } x = 1 \\ 1 - p \text{ if } x = 0 \\ 0 \text{ if } x = k \neq 0, 1 \end{cases} \\ \mu(X) &= p \\ \mathrm{var}(X) &= p(1 - p) \\ \sigma(X) &= \sqrt{p(1 - p)} \end{aligned}$$

$$\begin{tikzcd}
(\Omega = \{\text{s, f}\}, \mathbb{P}(\Omega)) \arrow[r, "X"] \arrow[d, "P"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) & \text{success} \arrow[r] \arrow[d] & 1 & \text{failure} \arrow[r] \arrow[d] & 0 \\
{[0, 1]} & & p & & 1 - p &
\end{tikzcd}$$

Example 4.34. We can model the game of picking any of the 4 blue balls to win \$1 and picking any of the 6 red balls to lose \$1 as follows

$$\begin{tikzcd}
(\{\text{4 blue balls, 6 red balls}\}, \mathcal{F}'') \arrow[r, "f"] \arrow[d, "P''"'] \arrow[drr, "X''"] & (\{\text{blue, red}\}, \mathcal{F}') \arrow[r, "g"] \arrow[dl, "P'"'] \arrow[dr, "X'"] & (\{\text{success, failure}\}, \mathcal{F}) \arrow[d, "x"] \arrow[dll, "P"] \\
{[0,1]} & & \mathbb{R}
\end{tikzcd}$$

Note that $f, g,$ and X map outcomes while P, P', and P'' draw from $\mathcal{F}, \mathcal{F}'$, and $\mathcal{F}''$. For example $\{\text{blue ball}\} \notin \mathcal{F}''$ and P'' does not map a blue ball to 4/10.

4.5.2. *Binomial Random Variable.* (X, n, p) counts the number of successes in n Bernoulli trials $X_1, \ldots, X_n$ with same success rate p. If Ω is the domain of the X_i then

$$\Omega' = \{\text{all sequences of s and f of length } n\} = \Omega \times \cdots \times \Omega$$

is the domain of X and $(\Omega', \mathbb{P}(\Omega'), P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ has

$$\begin{aligned}
p_X(k) &= \begin{cases} C(n,k)p^k(1-p)^{n-k} \text{ if } 0 \le k \le n \\ 0 \text{ if } k > n \end{cases} \\
\mu(X) &= np \\
\operatorname{var}(X) &= np(1-p)
\end{aligned}$$

This binomial random variable X is not $X_1 + \cdots + X_n$, they do not have the same domain.

Example 4.35. To find the probability of getting 4 tails in a sequence of 10 coin tosses given the probability of getting tails is p, we define a measurable space, a probability measure for that sample space, and a random variable for that probability space. Let Ω be the set of all sequences of tosses of length 10, i.e.

$$\Omega = \{HHHHHHHHHH, HH \ldots HT, \ldots, TTTTTTTTTT\}$$

then $(\Omega, \mathbb{P}(\Omega))$ is a measurable space. Let $(\Omega, \mathbb{P}(\Omega)) \xrightarrow{P} [0,1]$ be the probability measure defined by $P(w) = p^k(1-p)^{10-k}$ where k is the number of tails in each sequence w. Verify that $P(\Omega) = 1$. Lastly, let $(\Omega, \mathbb{P}(\Omega), P) \xrightarrow{X} \mathbb{R}$ be the random variable that counts the number of tails in each w. Then X is discrete with pmf $p_X(k) = C(10,k)p^k(1-p)^{10-k}$ and we have the following diagram,

$$\begin{tikzcd}
(\Omega, \mathbb{P}(\Omega)) \arrow[r, "X"] \arrow[d, "P"'] & \mathbb{R} & X^{-1}(4) \arrow[r] \arrow[d] & 4 \\
{[0,1]} & & C(10,4)p^4(1-p)^6 &
\end{tikzcd}$$

4.5.3. *Geometric Random Variable.* (X, p) marks where the first successful Bernoulli trial with success rate p is. It is discrete with pmf

$$p_X(k) = P(X = k) = (1-p)^{k-1}p$$

Example 4.36. If T stands for tail, then

$$P(\text{getting first tail on } 10^{\text{th}} \text{ flip}) = (\frac{1}{2})^9(\frac{1}{2})$$
$$P(\text{getting third tail on } 10^{\text{th}} \text{ flip}) = C(9,2)(\frac{1}{2})^3(\frac{1}{2})^7$$

where $C(9,2)$ comes from the number of ways to place the first two tails.

Exercise 4.37. Define the measure space $(\Omega, \mathcal{F}, P)$, verify $P(\Omega) = 1$ and define the geometric random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Calculate $\mu(X)$ and $\text{var}(X)$.

4.5.4. *Poisson Random Variable.* (X, λ) allows us to calculate chances of a certain number of customers per hour or other occurrences per time interval with $\mu_X = \text{var}(X) = \lambda$ as well as estimate probabilities for the similar binomial random variable (X, n, p). It is discrete with pmf

$$p_X(k) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

so everything depends on this λ. If asked per 2 hours, replace λ by $\lambda' = 2\lambda$ since the mean doubles.

Example 4.38. The number of customers that walk into KFC is Poisson distributed with mean $\lambda = 5$ customers/hour.

a. $P(1 \text{ customer from 2pm to 3pm}) = e^{-5}\frac{5^1}{1!}$.
b. $P(2 \text{ customers from 3pm to 4pm}) = e^{-5}\frac{5^2}{2!}$.
c. $P(3 \text{ customers from 2pm to 4pm}) = e^{-10}\frac{10^3}{3!}$. For the number of customers over 2 hours, you can consider X' with $\lambda' = 10$.
d. $P(1 \text{ from 2pm to 3pm and 2 from 3pm to 4pm given 3 from 2pm to 4pm}) = P(\text{A}|\text{B}) = \frac{ab}{c}$ (What are events A and B?)

Example 4.39. If a coin toss turns up tail with probability $p = P(\text{tail}) = \frac{1}{3}$ then

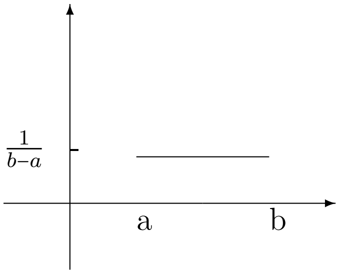
a. $P(\text{getting 40 tails in 100 tosses}) = P(\text{binomial } X(100, \frac{1}{3}) = 40) = C(100, 40)(\frac{1}{3})^{40}(\frac{2}{3})^{60}$ precisely.
b. If we think of n tosses as a time interval and the number of tails in there as the number of customers, then we can approximate the above probability with a Poisson random variable X'. This X' has $\lambda = np = \frac{100}{3}$. So $P(\text{getting 40 tails in 100 tosses}) \approx P(X' = 40) = e^{-\frac{100}{3}}\frac{(\frac{100}{3})^{40}}{40!}$.

Exercise 4.40. Define the measure space $(\Omega, \mathcal{F}, P)$, verify $P(\Omega) = 1$ and define the Poisson random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

4.5.5. *Uniform Random Variable.* (X, a, b) is perhaps the nicest of all continuous random variables, so it is examined firstly. Imagine throwing darts at the real line such that they will land randomly at any point within the stretch $[a, b]$. Then X is uniformly distributed over $[a, b]$ with pdf

$$f_X(s) = \begin{cases} \frac{1}{b-a} \text{ if } s \in [a,b] \\ 0 \text{ if } s \notin [a,b] \end{cases}$$

Can you compute $\mu(X)$ and var(X)?



Example 4.41. Find c so that the function

$$f(s) = \begin{cases} c \text{ if } s \in [2,5] \\ 0 \text{ if } s \notin [2,5] \end{cases}$$

is a pdf. Given X with such pdf, find $\mu(X)$ and var(X).

Example 4.42. One can generalize this to the game of throwing darts at a disk in the plane, then one would have a 2-dimensional uniform random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$.

Exercise 4.43. Define the measure space $(\Omega, \mathcal{F}, P)$, verify $P(\Omega) = 1$ and define the uniform random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Hint: you may try

$$\begin{aligned} \Omega &= [a,b] \\ \mathcal{F} &= \mathcal{B}([a,b]) \\ P(-) &= \frac{\mu_B(-)}{b-a} \text{ where } \mu_B \text{ is the Borel measure} \\ X &= id_{[a,b]} \end{aligned}$$

Or you may try

$$\begin{aligned} \Omega &= \mathbb{R} \\ \mathcal{F} &= \mathcal{B}(\mathbb{R}) \\ P(-) &= \frac{\mu_B(- \cap [a,b])}{b-a} = \int_{- \cap [a,b]} \frac{1}{b-a} dt \\ X &= id_{\mathbb{R}} \end{aligned}$$

4.5.6. *Exponential Random Variable.* (X, λ) models longevity with density function

$$f_X(s) = \begin{cases} \lambda e^{-\lambda s} \text{ if } s \geq 0 \\ 0 \text{ if } s < 0 \text{ (no chance of lasting a negative number of years)} \end{cases}$$

Easy integrations give all these

$$E(X) = \frac{1}{\lambda} \text{ and } \operatorname{var}(X) = \frac{1}{\lambda^2}$$

$$F_X(x) = P(X \leq x) = P(\text{lasting less than } x \text{ years}) = 1 - e^{-\lambda x}$$

$$S_X(x) = P(X > x) = P(\text{surviving at least } x \text{ years}) = e^{-\lambda x}$$

Beside average lifetime μ_X, median lifetime x_m is the moment when both probabilities $F_X(x_m) = S_X(x_m) = e^{-\lambda x_m} = 0.5$. Given one, you know the other. Again λ determines everything. A realistic model would not have constant but rather increasing hazard rate $\lambda(t)$ to account for aging, in which case you must calculate $S_X(x) = e^{-\int_0^x \lambda(t) dt}$.

Example 4.44. Lifetime of a battery is often exponentially distributed.

a. If its average lifetime is 2 years, find the probability that it will last 6 years.
b. Find the probability that it will last another 4 years given it has lasted 6 years.
c. If its median lifetime is 3 years, find the probability that it will die within 1 year.

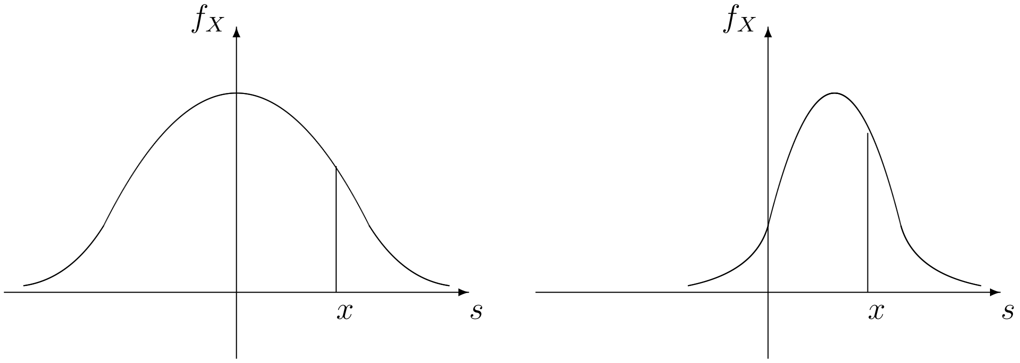
4.5.7. *Normal Random Variable.* (X, μ, σ) describes large sums of independent identically distributed random variables with mean μ and standard deviation σ. It is continuous with

$$f_X(s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(s-\mu)^2}{2\sigma^2}}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(s) ds = \text{area under bell curve over } (-\infty, x]$$

Nicest is the standard case $(X, 0, 1)$ where the standard normal table gives all values $F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$ for $x \in [0, 3.49]$. When $\mu \neq 0$ and $\sigma \neq 1$ you can reduce to the standard case by making a change of variables $Z = \frac{X-\mu}{\sigma}$.

Note that a larger mean μ means a shift of the graph to the right while a larger standard deviation σ means a wider bell.



Example 4.45. Let $(X, 0, 1)$ be the standard normal distribution.

a. Find $P(X < 1)$.
b. Find $P(X \geq 2)$.
c. Find $P(X < -1)$. Standard normal table does not handle negative numbers, but this by symmetry is the same as $P(X > 1) = 1 - P(X \leq 1) = 1 - 0.8413$.
d. Find c such that $P(|X| \leq c) = 0.9$. This is $P(-c \leq X \leq c)$ so we have $P(X < -c) = P(X > c) = 0.05$ or $P(X \leq c) = 0.95$. So c must be about 1.64 or 1.65.
e. Given $(X, 1, 3)$, find $P(X \leq 7)$ by working with $Z = \frac{X-1}{3}$.

Example 4.46. Suppose the height of pines is normally distributed with mean 130m and standard deviation 10m.

a. Find the percentage of trees between 135.1m and 149.6m. Of course you have to standardize $(X, 130, 10)$.
b. Find the maximum height of the shortest 5% of trees. Note that our table does not handle negative z.

4.6. **Exercises.** pages 184: 5, 6, 9, 12.

4.7. **Conditional Probability for Random Variables.** For random variables

$$(\Omega, \mathcal{F}, P) \underset{Y}{\overset{X}{\rightrightarrows}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

we can ask about the distribution of X given an event $G \in \mathcal{F}$, or given some information about Y, or given some σ-algebra $\mathcal{G} \subset \mathcal{F}$ of information in the senses we will make more precise.

We begin with the usual condition on an event G of nonzero measure. By definition 3.22 we get the conditional measure $P(-\,|\,G) = \frac{P(-\cap G)}{P(G)}$. Afterward, we can compute $P(F\,|\,G), E(X\,|\,G), \text{var}(X)$, etc. by integration over either $(\Omega, \mathcal{F}, P(-\,|\,G))$ or over $(\mathbb{R}, \mathcal{B}(\mathbb{R}), X_* P(-\,|\,G))$. For example, when X is discrete, we can compute conditional probability, conditional expectation, conditional variance

$$P(X = x_i\,|\,G) = \frac{P(\{X = x_i\} \cap G)}{P(G)} = p_{X\,|\,G}(x_i)$$

$$E(X\,|\,G) = \sum_i x_i P(X = x_i\,|\,G) = \sum_i x_i p_{X\,|\,G}(x_i)$$

$$\text{var}(X\,|\,G) = \sum_i (x_i - E(X\,|\,G))^2 p_{X\,|\,G}(x_i)$$

Without more information about P, X, G, we stop here.

Exercise 4.47. Verify that

a. $\sum_{x_i} p_{X\,|\,G}(x_i) = 1$ so that $p_{X\,|\,G}$ is in fact a pmf.
b. (trivial condition) $p_{X\,|\,\Omega}(x_i) = p_X(x_i)$ for all x_i.
c. (total probability law) $p_X(x) = \sum_{i=1}^n P(F_i) p_{X\,|\,F_i}(x)$ for any partition $\Omega = \bigsqcup_{i=1}^n F_i$.
d. (total expectation law) $E(X) = \sum_i^n P(F_i) E(X\,|\,F_i)$ for any partition $\Omega = \bigsqcup_{i=1}^n F_i$.

Example 4.48. When $G = \{Y = y\}$ for discrete X, Y, compute $P(X \in A \mid Y = y)$ and $E(X \mid Y = y)$ as far as you can. We will be able to calculate this more explicitly when there is more information about X, Y. In the real world, X and Y could be the prices of two stocks.

Example 4.49. Compute $P(F \mid Y = y)$ for $F \in \mathcal{F}$ as far as you can.

When $G = \{Y = y\}$ we change the notation $p_{X|G}$ to $p_{X|Y}$. We get maps

$$\begin{gathered}
(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{P(F \mid Y = -)} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
y \mapsto P(F \mid Y = y) \\
(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{E(X \mid Y = -)} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
y \mapsto E(X \mid Y = y) \\
(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{\operatorname{var}(X \mid Y = -)} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
y \mapsto \operatorname{var}(X \mid Y = y)
\end{gathered}$$

Whether the first map is measurable depends of F, while the second map and third map are measurable for all X, Y.

Definition 4.50. We define conditional expectation $E(X \mid Y)$ of X given Y to be the composition

$$\begin{gathered}
(\Omega, \mathcal{F}, P) \xrightarrow{Y} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{E(X \mid Y = -)} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
\omega \mapsto Y(\omega) \mapsto E(X \mid Y = Y(\omega))
\end{gathered}$$

and conditional variance $\operatorname{var}(X \mid Y)$ of X given Y to be the composition

$$\begin{gathered}
(\Omega, \mathcal{F}, P) \xrightarrow{Y} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{\operatorname{var}(X \mid Y = -)} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
\omega \mapsto Y(\omega) \mapsto \operatorname{var}(X \mid Y = Y(\omega))
\end{gathered}$$

$$\begin{tikzcd}
(\Omega, \mathcal{F}, P) \arrow[r, "Y"] \arrow[dr, "E(X \mid Y)"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[d, "E(X \mid Y = -)"] & (\Omega, \mathcal{F}, P) \arrow[r, "Y"] \arrow[dr, "\operatorname{var}(X \mid Y)"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[d, "\operatorname{var}(X \mid Y = -)"] \\
& (\mathbb{R}, \mathcal{B}(\mathbb{R})) & & (\mathbb{R}, \mathcal{B}(\mathbb{R}))
\end{tikzcd}$$

Both $E(X \mid Y)$ and $\operatorname{var}(X \mid Y)$ are random variables and with them come their invariants $E(E(X \mid Y)), \operatorname{var}(E(X \mid Y)), E(\operatorname{var}(X \mid Y)), \operatorname{var}(\operatorname{var}(X \mid Y))$, etc. One can think of $E(X \mid Y)$ as the expected value of one stock given information about another stock.

When X is continuous, the situation is more sophisticated. The pdf f_X does not help us calculate the conditional measure

$$\begin{gathered}
(\Omega, \mathcal{F}) \xrightarrow{P(- \mid G)} [0, 1] \\
F \mapsto \frac{P(F \cap G)}{P(G)}
\end{gathered}$$

because $F \bigcap G$ and G may not be of the form $X^{-1}(A)$ for any $A \in \mathcal{B}(\mathbb{R})$. Therefore f_X is unhelpful in computation of $P(X \in A|G)$ or $E(X|G)$. We need a conditional pdf $f_{X|G}$.

Theorem 4.51. *(conditional pdf given event) If $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a continuous random variable and $G \in \mathcal{F}$ then there exists a nonnegative measurable function $\mathbb{R} \xrightarrow{f} [0, \infty)$ such that $P(X \in A|G) = \int\limits_A f(s)ds$ for all $A \in \mathcal{B}(\mathbb{R})$. Moreover, this f is unique up to a μ_L-negligible set in $\mathbb{R}$.*

Proof. The pushforward $X_*(P(-|G))$ is a finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If $\mu_L(A) = 0$ for $A \in \mathcal{B}(\mathbb{R})$ then

$$\begin{aligned}
X_*(P(-|G))(A) &= P(-|G)(X^{-1}(A)) \\
&= P(X^{-1}(A))|G) \\
&= \frac{P(X^{-1}(A) \bigcap G)}{P(G)} \\
&\leq \frac{P(X^{-1}(A))}{P(G)} \\
&= \frac{\int\limits_A f_X(s)ds}{P(G)} \\
&= \frac{0}{P(G)} \\
&= 0
\end{aligned}$$

So $X_*(P(-|G))$ is absolutely continuous with respect to μ_L. By Radon-Nikodym theorem 3.21, there exists a nonnegative measurable function $\mathbb{R} \xrightarrow{f} [0, \infty)$ as desired and it is unique up to a μ_L-negligible set in $\mathbb{R}$.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[dr, "P(-|G)"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[d, shift right, "\mu_L"'] \arrow[d, shift left, "X_*(P(-|G))"] \\
 & {[0, \infty)}
\end{tikzcd}$$

$$\begin{tikzcd}
A \arrow[d, "\mu_L"'] & X^{-1}(A) \arrow[dr, "P(-|G)"'] & A \arrow[l, "X^{-1}"'] \arrow[d, "X_*(P(-|G))"] \\
\mu_L(A) & & X_*(P(-|G))(A)
\end{tikzcd}$$

□

Definition 4.52. We define conditional pdf $f_{X|G}$ of X given G to be the function f in theorem 4.51.

With this conditional pdf $f_{X|G}$ we can compute conditional probability, conditional expectation, conditional variance

$$P(X \in A|G) = P(X^{-1}(A)|G) = X_*(P(-|G))(A) = \int\limits_A f_{X|G}(s)ds$$

$$E(X|G) = \int\limits_{\mathbb{R}} sdX_*(P(-|G)) = \int\limits_{\mathbb{R}} sf_{X|G}(s)ds$$

$$\operatorname{var}(X\,|\,G) = \int\limits_{\mathbb{R}} (s - E(X\,|\,G))^2 dX_*(P(-\,|\,G)) = \int\limits_{\mathbb{R}} (s - E(X\,|\,G))^2 f_{X\,|\,G}(s)ds$$

Note that we still can not get the conditional measure

$$(\Omega, \mathcal{F}) \xrightarrow{P(-\,|\,G)} [0,1]$$
$$F \mapsto P(F\,|\,G)$$

because some F may not be of the form $X^{-1}(A)$ for any $A \in \mathcal{B}(\mathbb{R})$.

Exercise 4.53. Verify that

1. $\int\limits_{\mathbb{R}} f_{X\,|\,G}(s)ds = 1$.
2. (trivial condition) $f_{X\,|\,\Omega} = f_X$ and hence $P(X \in A\,|\,\Omega) = P(X \in A)$.
3. $f_{X\,|\,\{X \in B\}}(s) = \begin{cases} \frac{f_X(s)}{P(X \in B)} \text{ if } s \in B \\ 0 \text{ if } s \in B^c \end{cases}$ for any $B \in \mathcal{B}(\mathbb{R})$. This relates f_X and $f_{X\,|\,G}$ when $G = \{X \in B\}$.
4. (total probability law) $f_x = \sum\limits_{i=1}^{n} P(X \in B_i) f_{X|\{X \in B_i\}}$ for any partition $\mathbb{R} = \bigsqcup\limits_{i=1}^{n} B_i$ with $P(X \in B_i) > 0$.
5. (total expectation law) $E(X) = \sum\limits_{i}^{n} P(G_i)E(X\,|\,G_i)$ for any partition $\Omega = \bigsqcup\limits_{i=1}^{n} G_i$.

Example 4.54. When $G = \{Y \in B\}$ of positive measure for continuous X, Y, compute $P(X \in A\,|\,Y \in B)$ and $E(X\,|\,Y \in B)$ as far as you can. We will be able to calculate this more explicitly when there is more information about X, Y.

When $G = \{Y = y\}$ has measure 0, and this happens for example when Y is continuous, we can not define conditional measure $P(-\,|\,Y = y)$ by the classical way $\frac{P(- \cap Y = y)}{P(Y = y)}$. We can not even assume its existence to find $f_{X\,|\,Y=y}$ and work backward to $P(-\,|\,Y = y), E(X\,|\,Y = y), E(X\,|\,Y), \operatorname{var}(X\,|\,Y = y), \operatorname{var}(X\,|\,Y)$ as we did above. So we use another result from Radon-Nikodym theorem.

Theorem 4.55. *(conditional expectation given $Y = y$) If $(\Omega, \mathcal{F}, P) \overset{X}{\underset{Y}{\rightrightarrows}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are continuous random variables then there exists a nonnegative measurable function $\mathbb{R} \xrightarrow{f} [0, \infty)$ such that $\int\limits_{Y^{-1}(A)} X(\omega)dP = \int\limits_{A} f(s)d(Y_*(P))$ for all $A \in \mathcal{B}(\mathbb{R})$. Moreover, this f is unique up to a $Y_*(P)$-negligible set in $\mathbb{R}$.*

Proof. The pushforward $Y_*(P)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. If we define

$$(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{\mu} (-\infty, \infty)$$
$$A \mapsto \int\limits_{Y^{-1}(A)} X(\omega)dP$$

then μ is a σ-finite signed measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. And if $Y_*(P)(A) = P(Y^{-1}(A)) = 0$ for $A \in \mathcal{B}(\mathbb{R})$ then $\mu(A) = \int\limits_{Y^{-1}(A)} X(\omega)dP = 0$. So μ is absolutely continuous with respect to

$Y_*(P)$. By Radon-Nikodym theorem 3.21, there exists a nonnegative measurable function $\mathbb{R} \xrightarrow{f} [0,\infty)$ as desired and it is unique up to a $Y_*(P)$-negligible set in $\mathbb{R}$.

$$\begin{tikzcd}
(\Omega,\mathcal{F}) \arrow[r, shift left, "X"] \arrow[r, shift right, "Y"'] \arrow[dr, "P"'] & (\mathbb{R},\mathcal{B}(\mathbb{R})) \arrow[d, shift right, "\mu"'] \arrow[d, shift left, "Y_*(P)"] \\
 & (-\infty,\infty)
\end{tikzcd}
\quad
\begin{tikzcd}
Y^{-1}(A) \arrow[dr, "\int X(\omega)dP"'] & A \arrow[l, "Y^{-1}"'] \arrow[d, "\mu"'] \\
 & \mu(A)
\end{tikzcd}
\quad
\begin{tikzcd}
Y^{-1}(A) \arrow[dr, "P"'] & A \arrow[l, "Y^{-1}"'] \arrow[d, "Y_*(P)"] \\
 & Y_*(P)(A)
\end{tikzcd}$$

□

Definition 4.56. We define conditional expectation $E(X\,|\,Y=-)$ of X given $\{Y=-\}$ to be the function f in theorem 4.55.

Taking a cue from unconditional probability $P(F)=E(1_F)$ for all $F\in\mathcal{F}$ in example 4.18, we define conditional probability measure given event $\{Y=y\}$ via conditional expectation.

Definition 4.57. We define conditional measure given $Y=y$ as

$$(\Omega,\mathcal{F}) \xrightarrow{P(-\,|\,Y=y)} [0,1]$$
$$F \mapsto E(1_F\,|\,Y=y)$$

One needs to verify that this $P(-\,|\,Y=y)$ is actually a probability measure on $(\Omega,\mathcal{F})$. Then we have conditional probability, conditional expectation and conditional variance

$$P(F\,|\,Y=y) = E(1_F\,|\,Y=y) \text{ for all } F\in\mathcal{F}$$

$$E(X\,|\,Y=y) = \int_\Omega X(\omega)dP(-\,|\,Y=y) \text{ again}$$

$$\operatorname{var}(X\,|\,Y=y) = \int_\Omega (X(\omega)-E(X\,|\,Y=y))^2 dP(-\,|\,Y=y)$$

We get maps

$$(\mathbb{R},\mathcal{B}(\mathbb{R})) \xrightarrow{P(F\,|\,Y=-)} (\mathbb{R},\mathcal{B}(\mathbb{R}))$$
$$y \mapsto P(F\,|\,Y=y)$$
$$(\mathbb{R},\mathcal{B}(\mathbb{R})) \xrightarrow{E(X\,|\,Y=-)} (\mathbb{R},\mathcal{B}(\mathbb{R}))$$
$$y \mapsto E(X\,|\,Y=y)$$
$$(\mathbb{R},\mathcal{B}(\mathbb{R})) \xrightarrow{\operatorname{var}(X\,|\,Y=-)} (\mathbb{R},\mathcal{B}(\mathbb{R}))$$
$$y \mapsto \operatorname{var}(X\,|\,Y=y)$$

Whether the map first map is measurable again depends on F, while the second map and third map are measurable for all X, Y. So we repeat what we did in the discrete case.

Definition 4.58. We define conditional expectation $E(X\,|\,Y)$ to be the composition

$$(\Omega,\mathcal{F},P)\xrightarrow{Y}(\mathbb{R},\mathcal{B}(\mathbb{R}))\xrightarrow{E(X\,|\,Y=-)}(\mathbb{R},\mathcal{B}(\mathbb{R}))$$

and conditional variance $\operatorname{var}(X\,|\,Y)$ to be the composition

$$(\Omega,\mathcal{F},P)\xrightarrow{Y}(\mathbb{R},\mathcal{B}(\mathbb{R}))\xrightarrow{\operatorname{var}(X\,|\,Y=-)}(\mathbb{R},\mathcal{B}(\mathbb{R}))$$

Again both $E(X\,|\,Y)$ and $\operatorname{var}(X\,|\,Y)$ are random variables and with them come their invariants $E(E(X\,|\,Y)), \operatorname{var}(E(X\,|\,Y)), E(\operatorname{var}(X\,|\,Y)), \operatorname{var}(\operatorname{var}(X\,|\,Y))$, etc.

Example 4.59. Surely $E(Y\,|\,Y)(\omega) = E(Y\,|\,-)(Y)(\omega) = E(Y\,|\,Y = Y(\omega)) = Y(\omega)$, so $E(Y\,|\,Y) = Y$.

Exercise 4.60. Verify that

a. (total expectation law) $E(X) = \sum\limits_{y_i} P(Y = y_i)E(X\,|\,Y = y_i)$ for discrete X, Y.

b. (total expectation law) $E(X) = \int\limits_{\mathbb{R}} E(X\,|\,Y = y)f_Y(y)dy$ for continuous X, Y.

There is another way to define $E(X\,|\,Y)$ and $\operatorname{var}(X\,|\,Y)$ by conditioning on a σ-algebra.

Theorem 4.61. *(conditional expectation given σ-algebra) If $(\Omega,\mathcal{F},P)\xrightarrow{X}(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is a random variable and $\mathcal{G}\subset\mathcal{F}$ is a σ-subalgebra then there exists a nonnegative measurable function $(\Omega,\mathcal{G})\xrightarrow{f}([0,\infty),\mathcal{B}([0,\infty))), \omega\mapsto f(\omega)$ such that $\int\limits_G X(\omega)dP = \int\limits_G f(\omega)dP$ for all $G\in\mathcal{G}$. Moreover, this f is unique up to a P-negligible set in Ω.*

Proof. Note that P is a probability measure on $(\Omega,\mathcal{G})$. If we define

$$\begin{aligned}(\Omega,\mathcal{G}) &\xrightarrow{\mu} (-\infty,\infty)\\ G &\mapsto \int\limits_G X(\omega)dP\end{aligned}$$

then μ is a σ-finite signed measure on $(\Omega,\mathcal{G})$. And if $P(G) = 0$ then $\mu(G) = \int\limits_G X(\omega)dP = 0$. So μ is absolutely continuous with respect to P. By Radon-Nikodym theorem 3.21, there exists an nonnegative measurable function $(\Omega,\mathcal{G})\xrightarrow{f}([0,\infty),\mathcal{B}([0,\infty)))$ as desired and it is unique up to a P-negligible set in Ω. See the proofs of theorem 4.51 and theorem 4.55. □

Definition 4.62. We define conditional expectation $E(X\,|\,\mathcal{G})$ of X given $\mathcal{G}\subset\mathcal{F}$ to be the function f in theorem 4.61.

For each $\omega\in\Omega$, one can think of $E(X\,|\,\mathcal{G})(\omega)$ as the expected value of stock ω given knowledge $\mathcal{G}$. Since any measurable map $(\Omega,\mathcal{G})\xrightarrow{X}(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is also measurable with respect to $(\mathcal{F},\mathcal{B}(\mathbb{R}))$ by example 3.11, we can view $RV((\Omega,\mathcal{G},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$ as a subspace of $RV((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$. Then we can think of $E(X\,|\,\mathcal{G})$ as the projection of X onto $RV((\Omega,\mathcal{G},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$. While $\int\limits_G X(\omega)dP = \int\limits_G E(X\,|\,\mathcal{G})(\omega)dP$ for all $G\in\mathcal{G}$, uniqueness in Radon-Nikodym theorem implies $E(X\,|\,\mathcal{G}) = X$ only when X is measurable with respect to $(\mathcal{G},\mathcal{B}(\mathbb{R}))$, or in other words only when X is already in the subspace $RV((\Omega,\mathcal{G},P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$.

Example 4.63. When $\mathcal{G} = \mathcal{F}$ we have $E(X\,|\,\mathcal{F}) = X$. Or when $\mathcal{G} = \sigma(X) = \langle X^{-1}(\mathcal{B}(\mathbb{R}))\rangle$ the σ-algebra generated by all events about X, we have $E(X\,|\,\sigma(X)) = X$ because X is surely measurable with respect to $(\sigma(X), \mathcal{B}(\mathbb{R}))$. Both equalities agree with our intuition and with interpretation of projection.

All three theorems in this subsection provide the existence of the Radon-Nikodym derivatives $f_{X\,|\,G}, E(X\,|\,Y = -)$ and $E(X\,|\,\mathcal{G})$ but none spells out what each really is. In theorem 4.51

$$X_*(P(-\,|\,G))(A) = \int_A 1_{\mathbb{R}}(s) dX_*(P(-\,|\,G)) = \int_A f_{X\,|\,G}(s) ds$$

for $A \in \mathcal{B}(\mathbb{R})$, so people will write

$$\frac{dX_*(P(-\,|\,G))}{ds} = f_{X\,|\,G}(s)$$

Or

$$X_*(P)(A) = \int_A 1_{\mathbb{R}}(s) dX_*(P) = \int_A f_X(s) ds$$

for $A \in \mathcal{B}(\mathbb{R})$, so

$$\frac{dX_*(P)}{ds} = f_X(s)$$

Hence we gain expressions such as

$$\frac{dX_*(P(-\,|\,G))}{dX_*(P)} = \frac{f_{X\,|\,G}(s)}{f_X(s)}$$

that are often useful in computation.

Exercise 4.64. Verify that

a. (total expectation law) $E(X\,|\,\mathcal{G}) = \sum_i E(X\,|\,F_i) 1_{F_i}$ for any partition $\Omega = \bigsqcup_i F_i$ that generate $\mathcal{G}$.
b. In particular, when $\mathcal{G} = \{\varnothing, G, G^c, \Omega\}$ we get $E(X\,|\,\mathcal{G}) = E(X\,|\,G)1_G + E(X\,|\,G^c)1_{G^c}$. This relates $E(X\,|\,\mathcal{G})$ with $E(X\,|\,G)$.
c. (trivial condition) And when $G = \Omega$, we get $E(X\,|\,\mathcal{G}) = E(X\,|\,\Omega)1_\Omega = E(X)1_\Omega$.

Again we define conditional probability and conditional variance given $\mathcal{G}$ via conditional expectation.

Definition 4.65. For $\mathcal{G} \subset \mathcal{F} \ni F$ we define conditional probability $P(F\,|\,\mathcal{G}) = E(1_F\,|\,\mathcal{G})$ and conditional variance $\mathrm{var}(X\,|\,\mathcal{G}) = E(X^2\,|\,\mathcal{G}) - E(X\,|\,\mathcal{G})^2$.

By definition, all three $E(X\,|\,\mathcal{G}), P(F\,|\,\mathcal{G}), \mathrm{var}(X\,|\,\mathcal{G})$ are random variables. For fixed $F \in \mathcal{F}$, the random variable

$$(\Omega, \mathcal{G}, P) \xrightarrow{P(F\,|\,\mathcal{G})} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$
$$\omega \mapsto P(F\,|\,\mathcal{G})(\omega) = E(1_F\,|\,\mathcal{G})(\omega)$$

satisfies $\int\limits_G P(F\,|\,\mathcal{G})(\omega)dP = \int\limits_G 1_F(\omega)dP$ for all $G \in \mathcal{G}$. And for certain fixed ω, the map

$$\begin{aligned} \mathcal{F} &\xrightarrow{P(-|\mathcal{G})(\omega)} [0,1] \\ F &\mapsto P(F\,|\,\mathcal{G})(\omega) \end{aligned}$$

is a probability measure. This is how we can define conditional measure given a σ-algebra. Overall, we have a map

$$\begin{aligned} \mathcal{F} \times \Omega &\xrightarrow{P(-|\mathcal{G})} [0,1] \\ (F,\omega) &\mapsto P(F\,|\,\mathcal{G})(\omega) \end{aligned}$$

Example 4.66. For $\mathcal{G} = \{\varnothing, \Omega\} \subset \mathcal{F}$ and $F \in \mathcal{F}$, we get the constant random variable

$$\begin{aligned} (\Omega, \mathcal{F}, P) &\xrightarrow{P(F|\mathcal{G})} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ \omega &\mapsto E(1_F\,|\,\mathcal{G})(\omega) = E(1_F)1_\Omega(\omega) = P(F) \end{aligned}$$

by exercise 4.64.

For all our work in laying the foundation for conditional expectation, we reap the following important result with ease.

Proposition 4.67. *(law of iterated expectations) If X, Y are random variables and $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ are σ-algebras then $E(E(X\,|\,\mathcal{G})\,|\,\mathcal{H}) = E(X\,|\,\mathcal{H})$. In particular, $E(E(X\,|\,\mathcal{G})) = E(X)$ and $E(E(X\,|\,Y)) = E(X)$.*

Proof. For the first statement, we observe that $E(X\,|\,\mathcal{H})$ is measurable with respect to $(\mathcal{H}, \mathcal{B}(\mathbb{R}))$ by its very definition. Moreover

$$\int\limits_H E(X\,|\,\mathcal{G})(\omega)dP = \int\limits_H X(\omega)dP = \int\limits_H E(X\,|\,\mathcal{H})(\omega)dP$$

for all $H \in \mathcal{H}$ by defining properties of $E(X\,|\,\mathcal{G})$ and $E(X\,|\,\mathcal{H})$. Hence

$$E(X\,|\,\mathcal{H}) = E(E(X\,|\,\mathcal{G})\,|\,\mathcal{H})$$

by uniqueness of the latter.

For the second statement, let $\mathcal{H} = \{\varnothing, \Omega\}$, then

$$E(E(X\,|\,\mathcal{G}))1_\Omega = E(E(X\,|\,\mathcal{G})\,|\,\mathcal{H}) = E(X\,|\,\mathcal{H}) = E(X)1_\Omega$$

by exercise 4.64 and the first statement. For $\omega \in \Omega$ we achieve

$$E(E(X\,|\,\mathcal{G})) = E(E(X\,|\,\mathcal{G}))1_\Omega(\omega) = E(X)1_\Omega(\omega) = E(X)$$

The third statement follows from the second statement with $\mathcal{G} = \sigma(Y)$. We can also prove the third statement directly. In the discrete case

$$\begin{aligned}
E(E(X\,|\,Y)) &= \sum_y E(X\,|\,Y = y)P(Y = y) \\
&= \sum_y (\sum_x xP(X = x\,|\,Y = y)P(Y = y) \\
&= \sum_y \sum_x xP(X = x, Y = y) \\
&= \sum_x x \sum_y P(X = x, Y = y) \\
&= \sum_x xP(X = x) = E(X)
\end{aligned}$$

If Y is continuous then

$$\begin{aligned}
E(E(X\,|\,Y)) &= \int_{\mathbb{R}} E(X\,|\,Y)(y)f_Y(y)dy \\
&= \int_{\mathbb{R}} E(X\,|\,Y)(y)dY_*(P) \\
&= \int_{\Omega} X(\omega)dP \\
&= \int_{\mathbb{R}} xdX_*(P) \\
&= \int_{\mathbb{R}} xf_X(x)dx \\
&= E(X)
\end{aligned}$$

Along the way, we use the defining property of $E(X\,|\,Y)$ plus the facts $\frac{dY_*(P)}{dt} = f_Y(y)$ and $\frac{dX_*(P)}{dx} = f_X(x)$. □

This proposition will let us compute some examples later. What it means in linear algebra is projecting a vector onto a subspace, then projecting the result onto a possibly smaller subspace is the same as projecting onto the smaller space. What does it mean in finance? We will prove this proposition again directly when X, Y are joint. Now, we peruse an example as antidote to all these definitions and claims.

Example 4.68. We toss a coin whose probability of heads is given by a random variable Y for n times and count the number X of heads. So Y is a measurable probability measure. If $(\Omega, \mathcal{F}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ where $\Omega = \{$sequences of H and T of length $n\}$ and $\mathcal{F} = \mathbb{P}(\Omega))$ then we can think of Y as

$$\begin{gathered}
(\Omega, \mathcal{F}) \xrightarrow{Y} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\
\text{HH...TT} \mapsto p^k(1-p)^{n-k}
\end{gathered}$$

a. For any $\omega \in \Omega$ we have

$$\begin{aligned} E(X\,|\,Y)(\omega) &= E(X\,|\,Y = Y(\omega)) \\ &= nY(\omega) \end{aligned}$$

so $E(X\,|\,Y) = nY$ as random variables.

b. For any $\omega \in \Omega$ we have

$$\begin{aligned} \mathrm{var}(X\,|\,Y)(\omega) &= \mathrm{var}(X\,|\,Y = Y(\omega)) \\ &= nY(\omega)(1 - Y(\omega)) \end{aligned}$$

so $\mathrm{var}(X\,|\,Y) = nY(1-Y)$ as random variables.

Suppose Y is uniformly distributed over $[0,1]$ with $E(Y) = \frac{1}{2}$ and $\mathrm{var}(Y) = \frac{1}{12}$.

c. By the law of iterated expectations

$$\begin{aligned} E(E(X\,|\,Y)) &= E(nY) \\ &= nE(Y) \\ &= \frac{n}{2} \\ &= E(X) \end{aligned}$$

d.

$$\begin{aligned} \mathrm{var}(E(X\,|\,Y)) &= \mathrm{var}(nY) \\ &= n^2\mathrm{var}(Y) \\ &= \frac{n^2}{12} \end{aligned}$$

e.

$$\begin{aligned} E(\mathrm{var}(X\,|\,Y)) &= E(nY(1-Y)) \\ &= n(E(Y) - E(Y^2)) \\ &= n(E(Y) - (\mathrm{var}(Y) + E(Y)^2)) \\ &= n(\frac{1}{2} - \frac{1}{12} - \frac{1}{4}) \\ &= \frac{n}{6} \end{aligned}$$

f. By exercise 4.73

$$\begin{aligned} \mathrm{var}(X) &= E(\mathrm{var}(X\,|\,Y)) + \mathrm{var}(E(X\,|\,Y)) \\ &= \frac{n}{6} + \frac{n^2}{12} \end{aligned}$$

g.

$$\begin{aligned}\operatorname{var}(\operatorname{var}(X\,|\,Y)) &= \operatorname{var}(nY(1-Y))\\ &= n^2\operatorname{var}(Y-Y^2)\\ &= n^2(\operatorname{var}(Y)+\operatorname{var}(Y^2)-2\operatorname{cov}(Y,Y^2))\\ &= \frac{n^2}{12}+n^2\operatorname{var}(Y^2)-2n^2(E(Y^3)-E(Y)E(Y^2))\end{aligned}$$

Exercise 4.69. Can we force the case $\{Y=y\}$ of measure 0 for continuous Y by opening up an interval $(y-\delta,y+\delta)$ so that $\{y-\delta<Y<y+\delta\}$ has positive measure? More precisely, for each $f_{X\,|\,\{y-\delta<Y<y+\delta\}}(s)$ given in definition 4.52, does $\lim\limits_{\delta\to 0} f_{X\,|\,\{y-\delta<Y<y+\delta\}}(s)$ exist? If so, we call it $f_{X\,|\,Y=y}(s)$ and from there define $P(X\in A\,|\,Y=y), E(X\,|\,Y=y), \operatorname{var}(X\,|\,Y=y)$ as usual.

Exercise 4.70. If the limit above does not exist, let $G=\{y-\delta<Y<y+\delta\}$ for $\delta>0$.

a. Does $\lim\limits_{\delta\to 0} P(X\in A\,|\,G)=\lim\limits_{\delta\to 0}\int_A f_{X\,|\,G}(s)ds$ exist? If so, we call it $P(X\in A\,|\,Y=y)$.
b. Does $\lim\limits_{\delta\to 0} E(X\,|\,G)=\lim\limits_{\delta\to 0}\int_{\mathbb{R}} sf_{X\,|\,G}(s)ds$ exist? If so, we call it $E(X\,|\,Y=y)$ and further define $E(X\,|\,Y)=E(X\,|\,Y=-)\circ Y$.
c. Does $\lim\limits_{\delta\to 0}\operatorname{var}(X\,|\,G)=\lim\limits_{\delta\to 0}\int_{\mathbb{R}}(s-E(X\,|\,G)^2 f_{X\,|\,G}(s)ds$ exist? If so, we call it $\operatorname{var}(X\,|\,Y=y)$ and further define $\operatorname{var}(X\,|\,Y)=\operatorname{var}(X\,|\,Y=-)\circ Y$.

   Note that the difference between this exercise and the previous exercise is taking limit of integral versus taking integral of limit.

Exercise 4.71. Can you relate $P(X\in A\,|\,Y=y), E(X\,|\,Y=y), \operatorname{var}(X\,|\,Y=y)$ taken after definition 4.57 with $P(X\in A\,|\,y), E(X\,|\,Y=y), \operatorname{var}(X\,|\,Y=y)$ in exercise 4.70?

Exercise 4.72. By example 4.59 and example 4.63, $E(Y\,|\,\sigma(Y))=E(Y\,|\,Y)$ a special case. Can you show $E(X\,|\,\sigma(Y))=E(X\,|\,Y)$ and $\operatorname{var}(X\,|\,\sigma(Y))=\operatorname{var}(X\,|\,Y)$ in general? A positive answer means the distribution of X given Y is the same as the distribution of X given information generated by Y. Hint: show $E(X\,|\,Y)$ is measurable with respect to $(\sigma(Y),\mathcal{B}(\mathbb{R}))$ and satisfies the defining equality in theorem 4.61.

Exercise 4.73. Recall from exercise 4.29 that $(RVF((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R}))),\operatorname{cov}(-,-))$ is an inner product space. If

$$\begin{aligned}RV((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R}))) &\xrightarrow{E(-\,|\,\sigma(Y))} RV((\Omega,\sigma(Y),P),(\mathbb{R},\mathcal{B}(\mathbb{R})))\\ X &\mapsto E(X\,|\,\sigma(Y))=E(X\,|\,Y)\end{aligned}$$

is really a projection then it restricts to a projection

$$RVF((\Omega,\mathcal{F},P),(\mathbb{R},\mathcal{B}(\mathbb{R}))) \xrightarrow{E(-\,|\,\sigma(Y))} RVF((\Omega,\sigma(Y),P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$$

Can you show the following statements for $X,Y\in RVF((\Omega,\sigma(Y),P),(\mathbb{R},\mathcal{B}(\mathbb{R})))$?

a. $E(X\,|\,Y)=0$ whenever X,Y are perpendicular.
b. $E(X-E(X\,|\,Y)\,|\,Y)=0$. Hint: use proposition 4.67.
c. $E(X-E(X\,|\,Y))=0$. Hint: use proposition 4.67.
d. $\operatorname{var}(X-E(X\,|\,Y))=E(\operatorname{var}(X\,|\,Y))$. Hint: use (c) and proposition 4.67.

e. $E(E(X\,|\,Y)(X{-}E(X\,|\,Y))) = 0$. Hint: use proposition 4.67 and the fact $E(E(X\,|\,Y)Z\,|\,Y) = E(X\,|\,Y)E(Z\,|\,Y)$ due to $E(X\,|\,Y)$ being constant given Y.
f. Deduce from (e) that $\mathrm{cov}(E(X\,|\,Y), X - E(X\,|\,Y)) = 0$, that is $E(X\,|\,Y)$ and $X - E(X\,|\,Y)$ are perpendicular.
g. (Pythagorean theorem) Deduce $\mathrm{var}(X) = \mathrm{var}(X - E(X\,|\,Y)) + \mathrm{var}(E(X\,|\,Y))$.
h. (total expectation law) Deduce $\mathrm{var}(X) = E(\mathrm{var}(X\,|\,Y)) + \mathrm{var}(E(X\,|\,Y))$.

4.8. **Sequential Random Variables.** In the same problem there may be a string of variables $X_1, X_2, X_3 \dots$, one often related to the previous others.

Example 4.74. Suppose the number of donors arriving at your food drive per hour is Poisson distributed with mean 4 donors.

a. If the number of donors between 4-5pm and 5-6pm are independent, find the probability that 5 donors arrive between 4-6pm. Use $\lambda' = 2 \cdot 4 = 8$,
$$P(5 \text{ donors arrive between 4-6pm}) = \tfrac{e^{-8}8^5}{5!}$$
b. Given 5 donors come between 4-6pm, find the probability that 3 donors come between 4-5pm and 2 donors come between 5-6pm.
c. Assume the number of items they bring is Poisson distributed with mean 2 items/donor and that they are independent of each other. On average, how many donors must enter until one with at least 7 items arrives? Set Poisson $(Y, 2)$ to be the number of items per donor. Set Bernoulli (X, p) with success rate p equal the probability that a donor comes in with at least 7 items and and failure rate $1 - p$. Then

$$\begin{aligned} p &= P(\text{success}) \\ &= P(Y \geq 7) \\ &= 1 - P(Y = 0) - P(Y = 1) - P(Y = 2) - P(Y = 3) - P(Y = 4) - P(Y = 5) - P(Y = 6) \end{aligned}$$

Set geometric (X', p) to mark where the first successful trial appears, then $E(X') = \frac{1}{p}$ is our answer.

4.9. **Joint Random Variables.** When there are two or more random variables $X_1, \dots, X_n$ on the same probability space, they may jointly define interesting events whose probability we want to calculate. This leads us to consider the multivariate random variable $(X_1, \dots, X_n)$. For simplicity, we consider bivariate random variable (X, Y) by joining only two variables X and Y.

$$\begin{tikzcd} (\Omega, \mathcal{F}) \arrow[r, shift left, "X"] \arrow[r, shift right, "Y"'] \arrow[d, "P"] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ {[0,1]} & \end{tikzcd} \qquad \begin{tikzcd} (\Omega, \mathcal{F}) \arrow[r, "{(X,Y)}"] \arrow[d, "P"] & (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)) \\ {[0,1]} & \end{tikzcd}$$

Definition 4.75. Given two random variables $(\Omega, \mathcal{F}, P) \underset{Y}{\overset{X}{\rightrightarrows}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we define their joint random variable to be $(\Omega, \mathcal{F}, P) \xrightarrow{(X,Y)} (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ if (X, Y) is measurable, or equivalently if $(X, Y)^{-1}(U) \in \mathcal{F}$ for any $U \in \mathcal{B}(\mathbb{R}^2)$.

We say X and Y are joint. Again the distribution function for (X,Y) is most readily defined, regardless whether X,Y are discrete or continuous.

Definition 4.76. We define the joint cumulative distribution function of (X,Y) to be $F_{X,Y}(x,y) = P(X \le x, Y \le y)$.

Definition 4.77. A joint random variable $(\Omega, \mathcal{F}, P) \xrightarrow{(X,Y)} (\mathbb{R}^2, \mathcal{L}(\mathbb{R}^2))$ is called discrete if it takes at most countably many values (x_i, y_i) in $\mathbb{R}^2$.

Again the function

$$\begin{aligned} &\mathbb{R}^2 \xrightarrow{p_{X,Y}} [0,1] \\ &(x,y) \mapsto p_{X,Y}(x,y) = P((X,Y)^{-1}(x,y)) = P(\{\omega \mid X(\omega) = x, Y(\omega) = y\}) \end{aligned}$$

is 0 everywhere except over those (x_i, y_i). It is called probability mass function of X and abbreviated as pmf. Clearly (X,Y) is discrete iff X and Y are discrete and

$$\begin{aligned} p_{X,Y}(x,y) &= P(X = x, Y = y) \\ &= P(X = x \cap Y = y) \\ &= P(\{w \in \Omega \,|\, X(w) = x, Y(w) = y)\}) \end{aligned}$$

We should not expect $p_{X,Y}(x,y)$ to depend on the size $p_X(x)$ of event $\{X = x\}$ and the size $p_Y(y)$ of event $\{Y = y\}$. It depends on how the two events intersect. From $p_{X,Y}$ we can recover the marginal p_X, p_Y

$$p_X(x) = \sum_{\text{all } y} p_{X,Y}(x,y)$$

$$p_Y(y) = \sum_{\text{all } x} p_{X,Y}(x,y)$$

The probability mass function $p_{X,Y}$ completely determines the cumulative density function $F_{X,Y}$ via

$$\begin{aligned} F_{X,Y}(x,y) &= P(X \le x, Y \le y) \\ &= \sum_{x_i \le x, y_i \le y} p_{X,Y}(x_i, y_i) \end{aligned}$$

and vice versa.

Example 4.78. If (X,Y) has joint probability mass function $p_{X,Y}(k,l) = \frac{\lambda^k \mu^l}{k! l!} e^{-(\lambda+\mu)}$ then

$$\begin{aligned} p_X(k) &= \sum_{l=0}^{\infty} p_{X,Y}(k,l) \\ &= e^{-(\lambda+\mu)} \frac{\lambda^k}{k!} \sum_{l=0}^{\infty} \frac{\mu^l}{l!} \\ &= e^{-(\lambda+\mu)} \frac{\lambda^k}{k!} e^{\mu} \\ &= e^{-\lambda} \frac{\lambda^k}{k!} \end{aligned}$$

Definition 4.79. A joint random variable $(\Omega, \mathcal{F}, P) \xrightarrow{(X,Y)} (\mathbb{R}^2, \mathcal{L}(\mathbb{R}^2))$ is called continuous if its cumulative distribution function $F_{X,Y}$ is continuous.

Again general continuous joint random variables are hard to work with unless we know more about them.

Definition 4.80. A joint random variable $(\Omega, \mathcal{F}, P) \xrightarrow{(X,Y)} (\mathbb{R}^2, \mathcal{L}(\mathbb{R}^2))$ is called absolutely continuous if $(X,Y)_*(P)$ is absolutely continuous with respect to μ_L.

By Radon-Nikodym theorem 3.21, there exists a nonnegative measurable function $(\mathbb{R}^2, \mathcal{L}(\mathbb{R}^2)) \xrightarrow{\quad f \quad} ([0,\infty), \mathcal{L}([0,\infty)))$ such that $(X,Y)_*(P)(A) = \iint\limits_A f(s,t)dsdt$ for all $A \in \mathcal{L}(\mathbb{R}^2)$

Again this Radon-Nikodym derivative $f = \frac{d(X,Y)_*(P)}{d\mu_L}$ is called the probability density function of X, abbreviated by pdf and denoted by $f_{X,Y}$. Now the probability of events $\{\omega, (X,Y)(\omega) \in A\}$ are given by areas under a surface $f_{X,Y}$ over A

$$P((X,Y) \in A) = (X,Y)_*(A) = \int\limits_A f_{X,Y}(s,t)dsdt$$

and this is what makes absolutely continuous joint random variables more tractable then general continuous joint random variables. In probability theory and in this course, when we say continuous joint random variables, we mean absolutely continuous joint random variables equipped with probability density functions $f_{X,Y}$.

Surely $(X,Y)_*(P)(\{(x,y)\}) = \iint\limits_{\{(x,y)\}} f_{X,Y}(s,t)dsdt = 0$ while $P(\Omega) = (X,Y)_*(\mathbb{R}^2) = \iint\limits_{\mathbb{R}^2} f_{X,Y}(s,t)dsdt = 1$. Furthermore, the probability density function $f_{X,Y}$ completely determines the cumulative density function $F_{X,Y}$ and vice versa

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = \iint\limits_{(-\infty,x]\times(-\infty,y]} f_{X,Y}(s,t)dsdt$$

From $f_{X,Y}$ we can recover the marginal f_X and f_Y.

Theorem 4.81. *If the joint random variable (X,Y) has joint CDF F_X and joint pdf $f_{X,Y}$ then X and Y have*

$$f_X(s) = \int\limits_{-\infty}^{\infty} f_{X,Y}(s,t)dt$$

$$f_Y(t) = \int\limits_{-\infty}^{\infty} f_{X,Y}(s,t)ds$$

Proof. By definition

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \iint\limits_{\{(s,t), s \leq x\}} f_{X,Y}(s,t)dsdt \\ &= \int\limits_{-\infty}^{x} \left(\int\limits_{-\infty}^{\infty} f_{X,Y}(s,t)dt \right) ds \end{aligned}$$

Hence $f_X(s) = \int\limits_{-\infty}^{\infty} f_{X,Y}(s,t)dt$. The same goes for Y. □

Draw the graph of $f_{X,Y}$ in 3-D. Then $f_X(s)$ is the area under the graph of $f_{X,Y}$ above $\{s,t\}, t \in \mathbb{R}\}$.

Like random variables, there are joint random variables (X,Y) that are neither discrete nor continuous. Unlike random variables, a multivariate random variable (X,Y) has no notion of real-valued mean and variance. For example, neither definition $\iint\limits_{\mathbb{R}^2}(s,t)f_{X,Y}(x,y)dsdt$ or $\int\limits_{\Omega}(X,Y)(\omega)dP$ works.

Example 4.82. We return to the meeting in example 2.5.

a. Model this meeting with joint random variable. Hint: Let X and Y be the two arrival times. Since all arrival times between 0 and 1 are equally likely, X and Y are identical uniform random variables with $f_X = f_Y = 1$ over $[0,1]$. Next, consider the bivariate random variable (X,Y). Since no point (x,y) in the unit square is likelier than the other

$$f_{X,Y}(s,t) = \begin{cases} c \text{ if } (s,t) \in [0,1] \times [0,1] \\ 0 \text{ if } (s,t) \notin [0,1] \times [0,1] \end{cases}$$

This is nothing but a 2-dimensional uniform distribution. For this to be a joint pdf, c must be 1. Draw unit squares for f_X, f_Y and unit cube for $f_{X,Y}$.

b. Calculate the probability that they actually meet in example 2.5.

c. Calculate the chances they meet within 15' of appointment time (square of sides 0.25).

d. From $f_{X,Y}$, we recover the marginal

$$\begin{aligned} f_X(s) &= \int\limits_{-\infty}^{\infty} f_{X,Y}(s,t)dt \\ &= \int\limits_{0}^{1} 1dt \\ &= 1 \end{aligned}$$

for all $s \in [0,1]$ and

$$f_X(s) = 0$$

for all $s \notin [0,1]$. The same goes for f_Y

Example 4.83. If (X,Y) has joint pdf

$$f_{X,Y}(s,t) = \begin{cases} \frac{1}{t}e^{-\frac{s}{t}-t} \text{ if } s,t > 0 \\ 0 \text{ if } s \le 0 \text{ or } t \le 0 \end{cases}$$

then we recover

$$\begin{aligned} f_Y(t) &= \int_{-\infty}^{\infty} f_{X,Y}(s,t)ds \\ &= \int_{0}^{\infty} \frac{1}{t}e^{-\frac{s}{t}-t}ds \\ &= e^{-t} \end{aligned}$$

for all $t > 0$. Hence Y is an exponential random variable with $\lambda = 1$.

4.10. **Exercises.**

pages 122: 15, 16, 23.

4.11. **Conditional Probability for Joint Random Variables.** Now we can return to the matter of conditional probability for random variables with more information in hands.

When $X, Y, (X,Y)$ are discrete random variables with pmfs $p_X, p_Y, p_{X,Y}$, we can compute the conditional pmf for X given event $Y = y$ easily.

Theorem 4.84. *The conditional pmf $p_{X|Y}(x|y)$ of X with respect to the measure $P(-|Y = y)$ is $\frac{p_{X,Y}(x,y)}{p_Y(y)}$ for all x, y. Similarly, $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$.*

Proof. Clearly

$$\begin{aligned} p_{X|Y}(x|y) &= P(X = x|Y = y) \\ &= \frac{P(X = x \bigcap Y = y)}{P(Y = y)} \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \end{aligned}$$

The same goes for $p_{Y|Y}(y|x)$. □

It follows that

$$\begin{aligned} P(X = x|Y = y)P(Y = y) = P(X = x \bigcap Y = y) &= P(Y = y|X = x)P(X = x) \\ E(X|Y = y) &= \sum_x x p_{X|Y}(x|y) \\ \mathrm{var}(X|Y = y) &= \sum_x (x - E(X|Y = y))^2 p_{X|Y}(x|Y) \end{aligned}$$

The same goes for the distribution of Y given event $X = x$.

When $X, Y, (X,Y)$ are continuous with pdfs $f_X, f_Y, f_{X,Y}$, we have the following result.

Theorem 4.85. *The conditional pdf $f_{X|Y}(x|y)$ of X with respect to the measure $P(-|Y = y)$ is $\frac{f_{X,Y}(x,y)}{f_Y(y)}$. Similarly, $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.*

Proof. We must show that

$$\begin{aligned} P(-\,|\,Y=y)(X^{-1}(A)) &= X_*(P(-\,|\,Y=y))(A) \\ &= \int\limits_A \frac{f_{X,Y}(x,y)}{f_Y(y)}dx \end{aligned}$$

for all $A \in \mathcal{B}(\mathbb{R})$. By definition of conditional measure given event $Y=y$ for continuous X,Y in 4.57, we must show that $\int\limits_A \frac{f_{X,Y}(x,y)}{f_Y(y)}dx = E(1_{X^{-1}(A)}\,|\,Y=y)$ for all A. For any $B \in \mathcal{F}$, we have

$$\begin{aligned} \int\limits_B \left(\int\limits_A \frac{f_{X,Y}(x,y)}{f_Y(y)}dx \right) f_Y(y)dy &= \int\limits_B \left(\int\limits_A \frac{f_{X,Y}(x,y)}{f_Y(y)} f_Y(y)dx \right) dy \\ &= \int\limits_B \left(\int\limits_A f_{X,Y}(x,y)dx \right) dy \\ &= \iint\limits_{A\times B} f_{X,Y}(x,y)dxdy \\ &= P(X\in A, Y\in B) \\ &= \iint\limits_{X^{-1}(A)\times Y^{-1}(B)} 1_\Omega(\omega)dP \\ &= \int\limits_{Y^{-1}(B)} 1_{X^{-1}(A)}dP \end{aligned}$$

The claim follows by uniqueness of $E(1_{X^{-1}}(A)\,|\,Y=y)$. The same goes for $f_{Y|X}(y\,|\,x)$. □

With this conditional pdf $f_{X|Y}(x\,|\,y)$ of X given event $Y=y$ we can compute explicitly

$$\begin{aligned} P(X\in A\,|\,Y=y) &= \int\limits_A f_{X|Y}(x\,|\,y)dx \\ E(X\,|\,Y=y) &= \int\limits_{\mathbb{R}} x f_{X|Y}(x\,|\,y)dx \\ \mathrm{var}(X\,|\,Y=y) &= \int\limits_{\mathbb{R}} (x - E(X\,|\,Y=y))^2 f_{X|Y}(x\,|\,y)dx \end{aligned}$$

The same goes for the distribution of Y given event $X=x$.

Example 4.86. Suppose a thrown dart will land randomly on some point in a round board of radius r. If X and Y are the s and t coordinates of the landing point then (X,Y) is a continuous bivariate random variable.

a. We get the joint pdf

$$f_{X,Y}(s,t) = \begin{cases} \frac{1}{\pi r^2} \text{ if } s^2+t^2 \le r^2 \\ 0 \text{ if } s^2+t^2 > r^2 \end{cases}$$

b. We get the marginal pdf

$$\begin{aligned} f_Y(t) &= \int_{-\infty}^{\infty} f_{X,Y}(s,t)ds \\ &= \int_{s^2+t^2\leq r^2} \frac{1}{\pi r^2} ds \\ &= \int_{-\sqrt{r^2-t^2}}^{\sqrt{r^2-t^2}} \frac{1}{\pi r^2} ds \\ &= \frac{2}{\pi r^2}\sqrt{r^2-t^2} \end{aligned}$$

for $-r \leq t \leq r$ and

$$f_Y(t) = 0$$

for $t < -r$ or $t > r$. Note that Y is not uniform. The same goes for f_X and X.

c. We get the conditional pdf

$$f_{X\mid Y}(s\mid t) = \frac{f_{X,Y}(s,t)}{f_Y(t)} = \begin{cases} \frac{1}{2\sqrt{r^2-t^2}} \text{ if } -r \leq s \leq r \\ 0 \text{ if } s < -r \text{ or } s > r \end{cases}$$

Hence X is uniform given event $Y = t$, agreeing with what is seen on the board.

We give a direct proof of the Law of Iterated Expectations 4.67 in the case X, Y are joint.

Proposition 4.87. *(law of iterated expectations for joint random variables) If X, Y are joint random variables then $E(E(X\mid Y)) = E(X)$.*

Proof. In the discrete case

$$\begin{aligned} E(E(X\mid Y)) &= \sum_y E(X\mid Y=y)p_Y(y) \\ &= \sum_y (\sum_x x p_{X\mid Y}(x\mid y))p_Y(y) \\ &= \sum_y \sum_x x p_{X,Y}(x,y) \\ &= \sum_x x \sum_y p_{X,Y}(x,y) \\ &= \sum_x x p_X(x) \\ &= E(X) \end{aligned}$$

If Y is continuous then

$$\begin{aligned}
E(E(X\,|\,Y)) &= E(E(X\,|\,Y = -) \circ Y) \\
&= \int\limits_{\mathbb{R}} E(X\,|\,Y = -)(t) f_Y(t) dt \\
&= \int\limits_{\mathbb{R}} E(X\,|\,Y = t) f_Y(t) dt \\
&= \int\limits_{\mathbb{R}} \left(\int\limits_{\mathbb{R}} s f_{X\,|\,Y}(s\,|\,t) ds \right) f_Y(t) dt \\
&= \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}} s f_{X\,|\,Y}(s\,|\,t) f_Y(t) ds dt \\
&= \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}} s f_{X,Y}(s,t) ds dt \\
&= \int\limits_{\mathbb{R}} s \left(\int\limits_{\mathbb{R}} f_{X,Y}(s,t) dt \right) ds \\
&= \int\limits_{\mathbb{R}} s f_X(s) ds \\
&= E(X)
\end{aligned}$$

Note that we did use the fact $E(g \circ Y) = \int\limits_{\mathbb{R}} g(t) f_Y(t) dt$. □

4.12. **Independence of Joint Random Variables.** Next we address the question of independence of random variables. Intuitively, we think two random variables X, Y are independent if every event $F \in \sigma(X)$ about X is independent of any event $G \in \sigma(Y)$ about Y as defined in 2.18. We use this very intuition as definition.

Definition 4.88. Two random variables X, Y are said to be independent if every event $F \in \sigma(X)$ is independent of any event $G \in \sigma(Y)$ and vice versa.

So to verify independence of X, Y, we have to show $P(F\,|\,G) = P(F)$ and $P(G\,|\,F) = P(G)$ for all events $F \in \sigma(X), G \in \sigma(Y)$. But this is no practical means, especially for arbitrary X, Y. However, for joint X, Y there is a more feasible criterion.

Proposition 4.89. *Two joint discrete random variables X, Y are independent iff $p_X(x)p_Y(y) = p_{X,Y}(x,y)$ for all $x, y \in \mathbb{R}$. Two joint continuous random variables X, Y are independent iff $f_X(x)f_Y(y) = f_{X,Y}(x,y)$ for all $x, y \in \mathbb{R}$.*

.

Proof. If $p_X(x)p_Y(y) = p_{X,Y}(x,y)$ then by theorem 4.84

$$\begin{aligned} P(X = x \,|\, Y = y) &= p_{X\,|\,Y}(x\,|\,y) \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= p_X(x) \\ &= P(X = x) \end{aligned}$$

and

$$\begin{aligned} P(Y = y \,|\, X = x) &= p_{Y\,|\,X}(y\,|\,x) \\ &= \frac{p_{X,Y}(x,y)}{p_X(x)} \\ &= p_Y(y) \\ &= P(Y = y) \end{aligned}$$

for all x,y. So $\{X = x\}$ and $\{Y = y\}$ are independent for all x,y. Since they generate $\sigma(X)$ and $\sigma(Y)$, it follows that X,Y are independent. The converse is obvious.

If $f_X(x)f_Y(y) = f_{X,Y}(x,y)$ then by theorem 4.85

$$\begin{aligned} P(X \in A \,|\, Y = y) &= \int\limits_A f_{X\,|\,Y}(x\,|\,y)dx \\ &= \int\limits_A \frac{f_{X,Y}(x,y)}{f_Y(y)}dx \\ &= \int\limits_A f_X(x)dx \\ &= P(X \in A) \end{aligned}$$

and

$$\begin{aligned} P(Y \in B \,|\, X = x) &= \int\limits_B f_{Y\,|\,X}(y\,|\,x)dy \\ &= \int\limits_B \frac{f_{X,Y}(x,y)}{f_X(x)}dy \\ &= \int\limits_B f_Y(y)dy \\ &= P(Y \in B) \end{aligned}$$

for all $x, y \in \mathbb{R}$ and $A, B \in \mathcal{B}(\mathbb{R})$. So $\{X \in A\}$ and $\{Y \in B\}$ are independent for all $A, B \in \mathcal{B}(\mathbb{R})$. Since they generate $\sigma(X)$ and $\sigma(Y)$, it follows that X,Y are independent. The converse is obvious. □

By this proposition, we can use $p_X(x)p_Y(y) = p_{X,Y}(x,y)$ or $f_X(x)f_Y(y) = f_{X,Y}(x,y)$ as equivalent definition of independence for joint X,Y. These generalize to $p_{X_1}(x_1)\cdots p_{X_n}(x_n) = p_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ for joint discrete $X_1,\ldots,X_n$ and $f_{X_1}(x_1)\cdots f_{X_n}(x_n) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$

for continuous $X_1, \dots, X_n$. This is a lot more succinct than asking all events $A_1 \in \sigma(X_1), \dots, A_n \in \sigma(X_n)$ be pairwise disjoint.

Example 4.90. Suppose the number N of coin tosses is Poisson distributed with mean λ and the probability of tossing heads is p. If X is the number of heads in N tosses and Y is the number of tails then $X + Y = N$ so they are seemingly dependent. Upon a closer look

$$\begin{aligned} p_{X,Y}(x,y) &= P(X = x, Y = y) \\ &= P(X = x, Y = y \,|\, N = x + y) P(N = x + y) \\ &= C(x + y, x) p^x (1 - p)^y \frac{e^{-\lambda} \lambda^{x+y}}{(x + y)!} \\ &= \frac{(\lambda p)^x (\lambda(1 - p))^y e^{-\lambda}}{x! y!} \end{aligned}$$

while

$$\begin{aligned} p_X(x) &= P(X = x) \\ &= \sum_{n \ge x} P(X = x \,|\, N = n) P(N = n) \\ &= \sum_{n \ge x} C(n, x) p^x (1 - p)^{n-x} \frac{e^{-\lambda} \lambda^n}{n!} \\ &= \frac{(\lambda p)^x e^{-\lambda p}}{x!} \end{aligned}$$

and similarly

$$\begin{aligned} p_Y(y) &= P(Y = y) \\ &= \frac{(\lambda(1 - p))^y e^{-\lambda(1-p)}}{y!} \end{aligned}$$

Hence $p_{X,Y}(x,y) = p_X(x) p_Y(y)$ and X and Y are independent. Can you explain why? Hint: fix $N = 10$ and recalculate $p_X(x), p_Y(y), p_{X,Y}(x,y)$.

Exercise 4.91. There is a third criterion to judge independence of joint X, Y. Show that joint X, Y are independent iff $F_{X,Y}(x,y) = F_X(x) F_Y(y)$ for all x, y.

Exercise 4.92. If X_1, X_2, X_3 are independent and uniformly distributed over $[1, 4]$ and $X = \min(X_1, X_2, X_3)$, find $F_X(x)$ and $f_X(s)$. Hint: this X is the composition

$$\Omega \xrightarrow{(X_1, X_2, X_3)} \mathbb{R}^3 \xrightarrow{\min} \mathbb{R}$$

For $s \notin [1, 4], f_X(s) = 0$. For $s \in [1, 4]$ we go with

$$\begin{aligned} P(X > x) &= P(\text{all } X_1, X_2, X_3 > x) \\ &= P(X_1 > x) P(X_2 > x) P(X_3 > x) \\ &= \left(\frac{4 - x}{3} \right)^3 \end{aligned}$$

(why not $P(X \leq x)$?) So $F_X(x) = P(X \leq x) = 1 - \frac{(4-x)^3}{27}$ and $f_X(s) = \frac{(4-s)^2}{9}$. With this its mean, variance, etc. can be computed.

Proposition 4.93. *If X and Y are independent joint random variables then*
1. $E(XY) = E(X)E(Y)$.
2. $cov(X,Y) = 0$ and X, Y are uncorrelated. The converse is not true.
3. $var(X+Y) = var(X) + var(Y)$.

Proof. The first statement follows from direct computation of mean and proposition 4.89. For example

$$\begin{aligned} E(XY) &= \sum_x \sum_y xy\, p_{(X,Y)}(x,y) \\ &= \sum_x \sum_y xy\, p_X(x) p_Y(y) \\ &= \sum_x x\, p_X(x) \sum_y y\, p_Y(y) \\ &= E(X)E(Y) \end{aligned}$$

The second statement follows from the first statement and proposition 4.33. The third statement follows from the second statement and proposition 4.33. □

We complete the proof of proposition 4.93 with a counterexample.

Example 4.94. If X is uniformly distributed over $[-1,1]$ and $Y = X^2$ then

$$\begin{aligned} \mathrm{cov}(X,Y) &= \mathrm{cov}(X, X^2) \\ &= E(X \cdot X^2) - E(X)E(X^2) \\ &= E(X^3) - E(X)E(X^2) \\ &= 0 - 0 \cdot E(X^2) \\ &= 0 \end{aligned}$$

So X and Y are uncorrelated but clearly dependent.

Exercise 4.95. Show that if X, Y are independent joint continuous random variables then $E(X \mid Y) = E(X)1_\Omega$ constant.

4.13. **Exercises.**
pages 246-251: 7, 12, 22, 23.

4.14. **Change of Random Variables.** As seen in example 4.92, when we change a random variable X by a measurable function g we get another random variable $Y = g \circ X$ and naturally want to know it behavior. We begin a systematic survey of this composition.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"] \arrow[dr, "g \circ X"] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[d, "g"] \\
{[0,1]} & (\mathbb{R}, \mathcal{B}(\mathbb{R}))
\end{tikzcd}
\qquad
\begin{tikzcd}
\{w \mid X(w) = x\} \arrow[r] \arrow[d, "P"] \arrow[dr, "g \circ X"] & \{x \mid g(x) = y\} \arrow[d, "g"] \\
{[0,1]} & y
\end{tikzcd}$$

When X is discrete, Y is certainly discrete and we can compute the distribution of Y from p_X and g

$$\begin{aligned}
p_Y(y) &= P(Y = y)\\
&= P(g \circ X = y)\\
&= P(X \in g^{-1}(y))\\
&= \sum_{x, g(x) = y} P(X = x)\\
&= \sum_{x, g(x) = y} p_X(x)\\
F_Y(y) &= P(Y \leq y)\\
&= P(g \circ X \leq y)\\
&= \sum_{x, g(x) \leq y} P(X = x)\\
&= \sum_{x, g(x) \leq y} p_X(x)\\
E(Y) &= \sum_{y_i} y_i p_Y(y_i)\\
&= \sum_{y_i} \left(\sum_{x, g(x) = y_i} g(x) p_X(x) \right)
\end{aligned}$$

When X is continuous and g is discrete, Y is still discrete and we can compute the distribution of Y from f_X and g

$$\begin{aligned}
p_Y(y) &= P(Y = y)\\
&= P(g \circ X = y)\\
&= P(X \in g^{-1}(y))\\
&= \int_{g^{-1}(y)} f_X(s) ds\\
F_y(y) &= P(Y \leq y)\\
&= \sum_{y_i \leq y} P(Y = y_i)\\
&= \sum_{y_i \leq y} \int_{g^{-1}(y_i)} f_X(s) ds\\
E(Y) &= \sum_{y_i} y_i p_Y(y_i)\\
&= \sum_{y_i} \left(\int_{g^{-1}(y_i)} g(s) f_X(s) ds \right)
\end{aligned}$$

Example 4.96. Back to the dartboard in example 4.86 and assume the board has radius 1. Let X measure the distance of the dart to the center and

$$g(u) = \begin{cases} 0 \text{ if } -0.1 \leq u \leq 0.1 \\ 1 \text{ if } u < -0.1 \text{ or } u > 0.1 \end{cases}$$

Then $Y = g \circ X$ takes only two values 0 and 1. We can compute directly

$$\begin{aligned} p_Y(0) &= P(Y = 0) \\ &= P(g \circ X = 0) \\ &= P(X \leq 0.1) \\ &= 0.1^2 \end{aligned}$$

while

$$\begin{aligned} p_Y(1) &= P(Y = 1) \\ &= P(X > 0.1) \\ &= 1 - 0.1^2 \end{aligned}$$

Meanwhile

$$F_X(x) = P(X \leq x) = \begin{cases} 0 \text{ if } x < 0 \\ P(\text{landing in disc of radius } x) = \frac{\pi x^2}{\pi 1^2} = x^2 \text{ if } 0 \leq x \leq 1 \\ 1 \text{ if } x > 1 \end{cases}$$

and

$$f_X(s) = \begin{cases} 2s \text{ if } 0 \leq s \leq 1 \\ 0 \text{ if } s < 0 \text{ or } s > 1 \end{cases}$$

This means X is not uniform, as one can see by eyes that $P(0.1 \leq X \leq 0.2) \neq P(0.8 \leq X \leq 0.9)$. From here, one can compute $E(X) = \int_{\mathbb{R}} s f_X(s) ds = \frac{2}{3}$, etc.

When X is continuous and g is continuous, Y is also continuous and we can compute the distribution of Y from f_X and g

$$\begin{aligned}
P(Y \in A) &= P(g \circ X \in A) \\
&= P(X \in g^{-1}(A)) \\
&= \int\limits_{g^{-1}(A)} f_X(s)ds \\
F_Y(y) &= P(Y \leq y) \\
&= P(g \circ X \leq y) \\
&= P(X \in g^{-1}((-\infty, y])) \\
&= \int\limits_{g^{-1}((-\infty,y])} f_X(s)ds \\
E(Y) &= \int\limits_{\Omega} Y(\omega)dP \\
&= \int\limits_{\Omega} (g \circ X)(\omega)dP \\
&= \int\limits_{\mathbb{R}} g(s)f_X(s)ds
\end{aligned}$$

Note that we use the definition of expected value as integration of Y over Ω with respect to P, so the appearance of $g(s)$ inside the last integral is natural, see the discussion after 4.14. In all cases, $E(g \circ X)$ may not equal $g(E(X))$.

Example 4.97. If X is the distance in example 4.96 and $g(u) = u^2$ then $Y = X^2$ is continuous. Again $F_X(x) = x^2$ and $f_X(s) = 2s$ while

$$E(Y) = E(X^2) = \int\limits_{\mathbb{R}} s^2 f_X(s)ds = \frac{1}{2}$$

Thus

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(X^2 \leq y) \\
&= P(0 \leq X \leq \sqrt{y}) \\
&= F_X(\sqrt{y}) \\
&= \begin{cases} 0 \text{ if } y < 0 \\ y \text{ if } 0 \leq y \leq 1 \\ 1 \text{ if } y > 1 \end{cases}
\end{aligned}$$

Differentiation then gives

$$f_Y(t) = \begin{cases} 1 \text{ if } 0 \leq t \leq 1 \\ 0 \text{ if } t < 0 \text{ or } t > 1 \end{cases}$$

From here we can compute

$$E(Y) = \int_{\mathbb{R}} t f_Y(t) dt = \frac{1}{2}$$

again. Compare this with $g(E(X))$ in example 4.96.

Example 4.98. (probability integral transformation in statistics) If X is a continuous random variable and $g = F_X$ itself then $Y = g \circ X = F_X \circ X$ is uniformly distributed over $[0,1]$.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"] \arrow[dr, "Y"] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[d, "F_X"] \\
{[0,1]} & {[0,1]}
\end{tikzcd}$$

In example 4.97, we found f_Y via F_Y, which in turn was found via F_X and g. When g is nice enough we can derive f_Y directly from f_X and g.

Proposition 4.99. *If X is a continuous random variable and g is a strictly monotone differentiable measurable function on the range of X then $Y = g \circ X$ is a continuous random variable with*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Proof. If $g(x)$ is strictly increasing then $g'(x)$ is positive, hence $g^{-1'}(y) = \frac{1}{g'(g^{-1}(y))}$ is also positive and $g^{-1}(y)$ is strictly increasing. By the chain rule

$$\begin{aligned}
f_Y(y) &= F_Y'(y) \\
&= \frac{dP(g \circ X \leq y)}{dy} \\
&= \frac{dP(X \leq g^{-1}(y))}{dy} \\
&= \frac{dF_X(g^{-1}(y))}{dy} \\
&= F_X'(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} \\
&= F_X'(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))} \\
&= F_X'(g^{-1}(y)) \frac{1}{g'(g^{-1}(y))}
\end{aligned}$$

An alternate proof comes from the fact that $dP = f_X(x)dx = f_Y(y)dy$. Hence

$$f_Y(y) = f_X(x) \frac{dx}{dy} = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy}$$

The same argument plus absolute value works when g is strictly decreasing. □

Example 4.100. Again back to the dart board, the function $g(u) = u^2$ is increasing and differentiable over the range $[0,1]$ of X with $g^{-1}(y) = \sqrt{y}$ and $(g^{-1})'(y) = \frac{1}{2\sqrt{y}}$. Therefore Y has density function

$$f_Y(y) = \begin{cases} 2\sqrt{y}\frac{1}{2\sqrt{y}} = 1 \text{ if } 0 < y < 1 \\ 0 \text{ if } y < 0 \text{ or } y > 1 \end{cases}$$

which agrees with the calculation above.

Example 4.101. When we compute the n^{th} moment $E(X^n)$ we actually compute $E(g \circ X)$ where $g(u) = u^n$.

Example 4.102. If (X,Y) is a continuous bivariate random variable with pdf $f_{X,Y}$ and $\mathbb{R}^2 \xrightarrow{g} \mathbb{R}$ is a continuous function then $Z = g \circ (X,Y)$ is another continuous random variable. We can calculate

$$\begin{aligned} E(Z) &= \int_{\Omega} Z(\omega) dP \\ &= \int_{\Omega} (g \circ (X,Y))(\omega) dP \\ &= \int_{\Omega} g(X(\omega), Y(\omega)) dP \\ &= \iint_{\mathbb{R}^2} g(s,t) f_{X,Y}(s,t) ds dt \\ \text{var}(Z) &= E(Z^2) - E(Z)^2 \\ &= E((g \circ (X,Y))^2) - E(g \circ (X,Y))^2 \\ &= E(h \circ g \circ (X,Y)) - E(g \circ (X,Y))^2 \end{aligned}$$

where $h(u) = u^2$. The case of (X,Y) discrete with pmf $p_{X,Y}$ is similar and even simpler. For example

$$\begin{aligned} E(Z) &= E(g \circ (X,Y)) \\ &= \sum_{x,y} g(x,y) p_{X,Y}(x,y) \end{aligned}$$

Having gone this far, we can even calculate the distribution of $g \circ X$ given some condition. For example, we can calculate

$$P(g \circ X \in A | G) = P(X \in g^{-1}(A)) | G)$$

Or we can calculate

$$\begin{aligned}
E(g \circ X \,|\, Y = y) &= \int\limits_{\Omega} (g \circ X)(\omega) dP(- \,|\, Y = y) \\
&= \int\limits_{\mathbb{R}} g(s) dX_*(P(- \,|\, Y = y)) \\
&= \int\limits_{\mathbb{R}} t d(g \circ X)_*(P(- \,|\, Y = y))
\end{aligned}$$

by viewing $g \circ X$ as a random variable over $(\Omega, \mathcal{F}, P(- \,|\, Y = y))$, regardless of whether X, Y, g are discrete or continuous. Or we can calculate $E(g \circ X \,|\, \mathcal{G})$ as the unique random variable f such that $\int\limits_G f(\omega) dP = \int\limits_G (g \circ X)(\omega) dP$ for all $G \in \mathcal{G}$.

Another approach is to view the previous discussion as changes of measures. The random variable X induces the pushforward $X_*(P)$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The function g then induces another pushforward $g_*(X_*(P)) = (g \circ X)_*(P) = Y_*(P)$ of $X_*(P)$ by g. It is another probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. When X is continuous then $X_*(P)(A) = \int\limits_A f_X(s) ds$. If Y is also continuous then we hope to find $f_Y(t)$ from $f_X(s)$ and g such that $Y_*(P)(B) = \int\limits_B f_Y(t) dt$.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[d, "P"'] \arrow[dr, "Y"] & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[d, "g"] \arrow[dl, "X_*(P)"'] \\
{[0, \infty]} & (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[l, "Y_*(P)"]
\end{tikzcd}
\qquad
\begin{tikzcd}
X^{-1}(g^{-1}(B)) \arrow[d, "P"'] \arrow[dr, "Y^{-1}"] & g^{-1}(B) \arrow[l, "X^{-1}"'] \arrow[dl, "X_*(P)"'] \\
Y_*(P)(B) & B \arrow[l, "Y_*(P)"] \arrow[u, "g^{-1}"']
\end{tikzcd}$$

Example 4.103. Let X be the uniform random variable $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{id_{\mathbb{R}}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ in exercise 4.103 with $f_X(s) = 1_{[3,4]}$ and let $g(u) = 2u$ be the change of variables. Surely $g^{-1}(v) = \frac{v}{2}$ with $g^{-1'}(v) = \frac{1}{2}$.

a. If $A = [3.5, 6]$ then

$$\begin{aligned}
X_*(P)(A) &= P(X^{-1}(A)) \\
&= \int\limits_A f_X(s) ds \\
&= \int\limits_{[3.5,6]} 1_{[3,4]} ds \\
&= \mu_L([3.5, 6] \cap [3, 4]) \\
&= 0.5
\end{aligned}$$

b. If $B = [7, 12]$ then

$$\begin{aligned} Y_*(P)([7,12]) &= X_*(P)(g^{-1}([7,12])) \\ &= X_*(P)([3.5,6]) \\ &= 0.5 \end{aligned}$$

c.

$$\begin{aligned} Y_*(P)(A) &= X_*(P)(g^{-1}(A)) \\ &= X_*(P)(g^{-1}([3.5,6])) \\ &= X_*(P)([1.75,3]) \\ &= 0 \end{aligned}$$

d. Alternatively

$$\begin{aligned} f_Y(t) &= f_X(g^{-1}(t)) \left| \frac{dg^{-1}(t)}{dt} \right| \\ &= f_X\left(\frac{t}{2}\right)\frac{1}{2} \\ &= \frac{1}{2} 1_{[3,4]}\left(\frac{t}{2}\right) \\ &= \frac{1}{2} 1_{[6,8]} \end{aligned}$$

so

$$\begin{aligned} Y_*(P)(A) &= \int_A f_Y(t) dt \\ &= \int_{[3.5,6]} \frac{1}{2} 1_{[6,8]} dt \\ &= \frac{1}{2}\mu_B([3.5,6] \cap [6,8]) \\ &= 0 \end{aligned}$$

agreeing with (c).

Exercise 4.104. Compute $Y_*(P)([5.5, 6.5])$ from its definition and from $f_Y(t)$ in the previous example.

Exercise 4.105. Let $(\Omega, \mathcal{F}, P) \overset{X}{\underset{Y}{\rightrightarrows}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two joint random variables and let $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \overset{f}{\underset{g}{\rightrightarrows}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two joint measurable functions, i.e. $(f, g)^{-1}(B)$ is measurable for any measurable $B \in \mathcal{B}(\mathbb{R}^2)$.

$$\begin{tikzcd}
(\Omega,\mathcal{F}) \arrow[r, shift left, "Y"] \arrow[r, shift right, "X"'] \arrow[dr, "g\circ Y", "f\circ X"'] & (\mathbb{R},\mathcal{B}(\mathbb{R})) \arrow[d, shift right, "f"'] \arrow[d, shift left, "g"] \\
& (\mathbb{R},\mathcal{B}(\mathbb{R}))
\end{tikzcd}
\qquad
\begin{tikzcd}
(\Omega,\mathcal{F}) \arrow[r, "{(X,Y)}"] \arrow[dr, "{(f\circ X, g\circ Y)}"'] & (\mathbb{R}^2,\mathcal{B}(\mathbb{R}^2)) \arrow[d, "{(f,g)}"] & (f,g)^{-1}(B) \\
& (\mathbb{R}^2,\mathcal{B}(\mathbb{R}^2)) & B \arrow[u]
\end{tikzcd}$$

a. Show that $f\circ X, g\circ Y$ are also joint random variables.
b. Show that $f\circ X, g\circ Y$ are independent if X, Y are independent. Hint: use definition 4.88 instead of proposition 4.89.
c. Deduce that $E((f\circ X)(g\circ Y)) = E(f\circ X)E(g\circ Y)$ from proposition 4.93.

4.15. **Exercises.**
page 119: 14, 24, 31, 40.
page 184: 1, 8.
page 246: 1, 5, 8.

4.16. Other Functions of X.

4.16.1. *Moment Generating Function.* Recall that we define the k^{th} moment of a random variable X to be $\mu_k(X) = E(X^k), k \geq 0$. If X is discrete then $\mu_k(X) = \sum_{i=1}^{\infty} x_i^k p(x_i)$. If X is continuous with pdf f_X then $\mu_k(X) = \int_{-\infty}^{\infty} s^k f_X(s)ds$, provided this integral is finite. We now see it as change of variable X by $g(u) = u^k$,

$$\begin{tikzcd}
(\Omega,\mathcal{F}) \arrow[r, "X"] \arrow[d, "P"'] \arrow[dr, "X^k"] & (\mathbb{R},\mathcal{B}(\mathbb{R})) \arrow[d, "g"] \\
{[0,1]} & (\mathbb{R},\mathcal{B}(\mathbb{R}))
\end{tikzcd}$$

In either case, the first two moments determine variance $\text{var}(X) = \mu_2(X) - \mu_1(X)^2$ while together all moments determine X. We capture them nicely in the next definition.

Definition 4.106. For a random variable X, we define its moment generating function $\mathbb{R} \xrightarrow{g_X} \mathbb{R}, t \mapsto E(e^{tX})$ whenever this expectation exists.

We have

$$\begin{aligned} g_X(t) &= E(e^{tX}) \\ &= E(\sum_{k=0}^{\infty} \frac{X^k t^k}{k!}) \\ &= \sum_{k=0}^{\infty} \frac{E(X^k) t^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{\mu_k(X) t^k}{k!} \end{aligned}$$

in series expansion, which exists if each coefficient $\mu_n(X)$ is finite and the series converges at t. In that case, the generating function encodes all moments as its coefficients $\mu_k(X) = g_X^{(k)}(0)$ and gives another expression for variance

$$\mathrm{var}(X) = g_X''(0) - g_X'(0)^2$$

Beside computing moment generating function via moments, we can compute it via definition of expected value

$$g_X(t) = E(e^{tX}) = \begin{cases} \sum\limits_{k=1}^{\infty} e^{tx_k} p_X(x_k) \text{ when } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{ts} f_X(s) ds \text{ when } X \text{ is continuous} \end{cases}$$

These explicit formulas let us compute $g_X(t)$ for the common X in 4.5.

Example 4.107. If X is geometrically distributed with parameter p then

$$\begin{aligned} g_X(t) &= \sum_{k=1}^{\infty} e^{tk} (1-p)^{k-1} p \\ &= pe^t \sum_{k=1}^{\infty} (e^t)^{k-1} (1-p)^{k-1} \\ &= pe^t \sum_{k=0}^{\infty} (e^t)^k (1-p)^k \\ &= \frac{pe^t}{1-(1-p)e^t} \end{aligned}$$

This series converges for $|(1-p)e^t| < 1$, or $t < \ln(1-p)$. As a byproduct, we get

$$\mu(X) = g_X'(0) = \frac{1}{p}$$

$$\mathrm{var}(X) = g_X''(0) - g_X'(0)^2 = \frac{1-p}{p^2}$$

as known before.

Example 4.108. If X is exponentially distributed with parameter λ then

$$\begin{aligned}\mu_k(X) &= \int_0^\infty s^k f_X(s)ds \\ &= \int_0^\infty s^k \lambda e^{-\lambda s} ds \\ &= \lambda(-1)^k \frac{\partial^k}{\partial \lambda^k} \int_0^\infty e^{-\lambda s} ds \\ &= \lambda(-1)^k \frac{\partial^n k}{\partial \lambda^k}\left(\frac{1}{\lambda}\right) \\ &= \frac{k!}{\lambda^k}\end{aligned}$$

after some integration by parts. Hence

$$\begin{aligned}g_X(t) &= \sum_{k=0}^\infty \frac{\mu_k(X)t^k}{k!} \\ &= \sum_{k=0}^\infty \frac{t^k}{\lambda^k} \\ &= \frac{\lambda}{\lambda - t}\end{aligned}$$

converges for $\left|\frac{t}{\lambda}\right| < 1$, or $|t| < \lambda$. Alternatively

$$\begin{aligned}g_X(t) &= \int_0^\infty e^{ts} \lambda e^{-\lambda s} ds \\ &= \frac{\lambda e^{(t-\lambda)s}}{t-\lambda}\Big|_0^\infty \\ &= \frac{\lambda}{\lambda - t}\end{aligned}$$

and from there

$$\begin{aligned}\mu_k(X) &= g_X^{(k)}(0) \\ &= \frac{\lambda k!}{(\lambda - t)^{k+1}}\Big|_{t=0} \\ &= \frac{k!}{\lambda^n}\end{aligned}$$

Exercise 4.109. Show that a Poisson random variable X with parameter λ has moment generating function $g_X(t) = e^{\lambda(e^t - 1)}$ and deduce that $E(X) = \text{var}(X) = \lambda$.

This bookkeeping device $g_X(t)$ completely describes the behavior of X, as we would suspect from the relation between $g_X(t)$ and pmf $p_X(x)$ or the relation between $g_X(t)$ and pdf $f_X(s)$. Surely, $p_X(x)$ determines $g_X(t)$ uniquely in the discrete case and $f_X(s)$

determines $g_X(t)$ uniquely in the continuous case. The converses also hold under appropriate hypotheses.

Theorem 4.110. *The moment generating function $g_X(t)$ of X determines its CDF $F_X(x)$ uniquely and vice versa. If X is a discrete with finite range then its moment generating function $g_X(t)$ determines its pmf $p_X(x)$ uniquely. If X is a continuous with bounded range then its moment generating function $g_X(t)$ determines its pdf $f_X(s)$ uniquely.*

Proof. see literature such as [?GrinsteadSnellItoP, 10.2] and [?GrinsteadSnellItoP, 10.5]. □

This theorem is false if discrete X has infinite range or if continuous X has unbounded range. Beside encapsulating the behavior of a random variable, moment generating function is very useful in handling sum of independent random variables.

Theorem 4.111. *If X and Y are independent random variables then $g_{X+Y}(t) = g_X(t)g_Y(t)$. Generally, if $X_1, \dots, X_n$ are independent random variables and $a_1, \dots, a_n$ are constants then $g_{a_1X_1+\cdots+a_nX_n}(t) = g_{X_1}(a_1t)\cdots g_{X_n}(a_nt)$.*

Proof. Since X and Y are independent, so are t^X and t^Y by exercise 4.105. Hence $g_{X+Y}(t) = E(t^{X+Y}) = E(t^X)E(t^Y) = g_X(t)g_Y(t)$ by proposition 4.93. Proof for the general case is similar. □

Example 4.112. If X is standard normally distributed then $\mu_0(X) = 1, \mu_1(X) = 0$ and $\mu_k(X) = (k-1)\mu_{k-2}(X)$ via integration by parts. So

$$\mu_k(X) = \begin{cases} \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} s^k e^{s^2/2} ds = \frac{(2l)!}{2^l l!} \text{ if } k = 2l \text{ even} \\ 0 \text{ if } k = 2l+1 \text{ odd} \end{cases}$$

Hence

$$\begin{aligned} g_X(t) &= \sum_{k=0}^{\infty} \frac{\mu_k(X)t^k}{k!} \\ &= \sum_{l=0}^{\infty} \frac{t^{2l}}{2^l l!} \\ &= e^{t^2/2} \end{aligned}$$

This series converges for all values of t. For the general normal random variable X with mean μ and standard deviation σ, one can show $g_X(t) = e^{t\mu+(\sigma^2/2)t^2}$. Furthermore, if X with μ_1, σ_1 and Y with μ_2, σ_2 are independent then

$$g_{X+Y}(t) = e^{t(\mu_1+\mu_2)+((\sigma_1^2+\sigma_2^2)/2)t^2}$$

by theorem 4.111. By theorem 4.110, $X+Y$ is normal with mean $\mu_1+\mu_2$ and standard deviation $\sqrt{\sigma_1^2+\sigma_2^2}$. Sum of two independent normal random variables is again a normal random variable.

Exercise 4.113. Show that sum of two independent Poisson random variables is again a Poisson random variable.

4.16.2. *Characteristic Function.* Sometimes the moments $\mu_k(X)$ are not finite or the series $g_X(t) = \sum\limits_{k=0}^{\infty} \frac{\mu_k(X)t^k}{k!}$ does not converge at t and so $g_X(t)$ is not defined. There is a way to circumvent these problems.

Definition 4.114. For a random variable X, we define its characteristic function $\mathbb{R} \xrightarrow{k_X} \mathbb{C}, t \mapsto E(e^{itX}) = \int_{-\infty}^{\infty} e^{its} f_X(s)ds$.

Note that characteristic function takes complex values. Moreover

$$\begin{aligned} ||k_X(t)|| &= || \int_{-\infty}^{\infty} e^{its} f_X(s)ds|| \\ &\leq \int_{-\infty}^{\infty} ||e^{its} f_X(s)||ds \\ &\leq \int\limits_{-\infty}^{\infty} f_X(s)ds \\ &= 1 \end{aligned}$$

so characteristic function always exists, unlike moment generating function. It is differentiable k times on $\mathbb{R}$ iff X has moments up to k^{th} order, and $\mu_k(X) = (-i)^k k_X^{(k)}(0)$. If the moment generating function for X exists then $k_X(t) = g_X(it)$.

When X is discrete with pmf $p_X(x)$ and values x_k for $k \geq 1$ then

$$\begin{aligned} k_X(t) &= E(e^{itX}) \\ &= \sum_{k=1}^{\infty} e^{itx_k} P(X = x_k) \\ &= \sum_{k=1}^{\infty} e^{itx_k} p_X(x_k) \end{aligned}$$

Example 4.115. The Bernoulli random variable X with success rate p has characteristic function

$$\begin{aligned} k_X(t) &= \sum_{k=1}^{2} e^{itx_k} p_X(x_k) \\ &= e^{it0} p_X(0) + e^{it1} p_X(1) \\ &= 1 - p + e^{it} p \end{aligned}$$

Hence

$$\begin{aligned} \mu_1(X) &= (-i)^1 k_X^{(1)}(0) = p \\ \mu_2(X) &= (-i)^2 k_X^{(2)}(0) = p \\ \text{var}(X) &= \mu_2(X) - \mu_1(X)^2 = p - p^2 = p(1-p) \end{aligned}$$

Exercise 4.116. Calculate the characteristic function for binomial random variable (X, n, p) and deduce its third moment.

We have the following result.

Theorem 4.117. *The characteristic function $k_X(t)$ of X determines its CDF $F_X(x)$ uniquely and vice verse. If X is continuous random variable then $f_X(s) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-its} k_X(t)dt$.*

Proof. see literature. □

In analysis this means $k_X(t)$ is the Fourier transform of $f_X(s)$. By now, one sees the pattern

$$F_X(x) = E(1_{\{X \le x\}})$$
$$g_X(t) = E(e^{tX})$$
$$k_X(t) = E(e^{itX})$$

All three, plus $p_X(x)$ when X is discrete or plus $f_X(s)$ when X is continuous, are closely related. Each uniquely determines X, yet each provides a different insight into the behavior of X.

Exercise 4.118. For each of the following continuous random variables X, find its characteristic function and deduce its mean and variance.

a. X with $f_X(s) = \begin{cases} \frac{1}{2} \text{ over } [0,2] \\ 0 \text{ elsewhere} \end{cases}$

b. X with $f_X(s) = \begin{cases} \frac{s}{2} \text{ over } [0,2] \\ 0 \text{ elsewhere} \end{cases}$

c. X with $f_X(s) = \begin{cases} |1-s| \text{ over } [0,2] \\ 0 \text{ elsewhere} \end{cases}$

d. X with $f_X(s) = \begin{cases} \frac{3s^2}{8} \text{ over } [0,2] \\ 0 \text{ elsewhere} \end{cases}$

Like moment generating function, characteristic function is useful in handling sum of independent random variables.

Theorem 4.119. *If X and Y are independent random variables then $k_{X+Y}(t) = k_X(t)k_Y(t)$. Generally, if $X_1, \dots, X_n$ are independent random variables and $a_1, \dots, a_n$ are constants then $k_{a_1X_1+\dots+a_nX_n}(t) = k_{X_1}(a_1t)\cdots k_{X_n}(a_nt)$.*

Proof. similar to the proof of theorem 4.111. □

Example 4.120. If $X_1, \dots, X_n$ are independent identically distributed random variables and $\overline{X}_n = \frac{X_1+\dots+X_n}{n}$ is their average then $k_{\overline{X}_n}(t) = k_{X_1}(\frac{t}{n})\cdots k_{X_n}(\frac{t}{n}) = k_X^n(\frac{t}{n})$ where X is distributed identically to the X_i.

5. Inequalities, Equalities, Convergences and Limits for Random Variables

5.1. **Inequalities.** Beside explicit calculation, we can bound different probabilities for X by its invariants $E(X), \operatorname{var}(X), \sigma(X)$.

5.1.1. *Markov's Inequality.* It bounds the probability that a random variable X is greater than some $a > 0$ by its mean.

Proposition 5.1. *If X is a random variable and $a > 0$ then*

$$P(|X| \ge a) \le \frac{E(|X|)}{a}$$

Proof. We have $|X| = |X|1_{|X|\geq a} + |X|1_{|X|<a} \geq a1_{|X|\geq a}$, so $E(|X|) \geq E(a1_{|X|\geq a}) = aP(|X| \geq a)$. □

Example 5.2. By Markov's inequality, no more than 1/5 of the population has more than 5 times the average income. Let X denote the income of the population with mean μ then $P(X \geq 5\mu) \leq \frac{\mu}{5\mu} = \frac{1}{5}$.

Exercise 5.3. Bound the probability that a driver is speeding over 60km/h by Markov's inequality if you know all drivers average 50km/h.

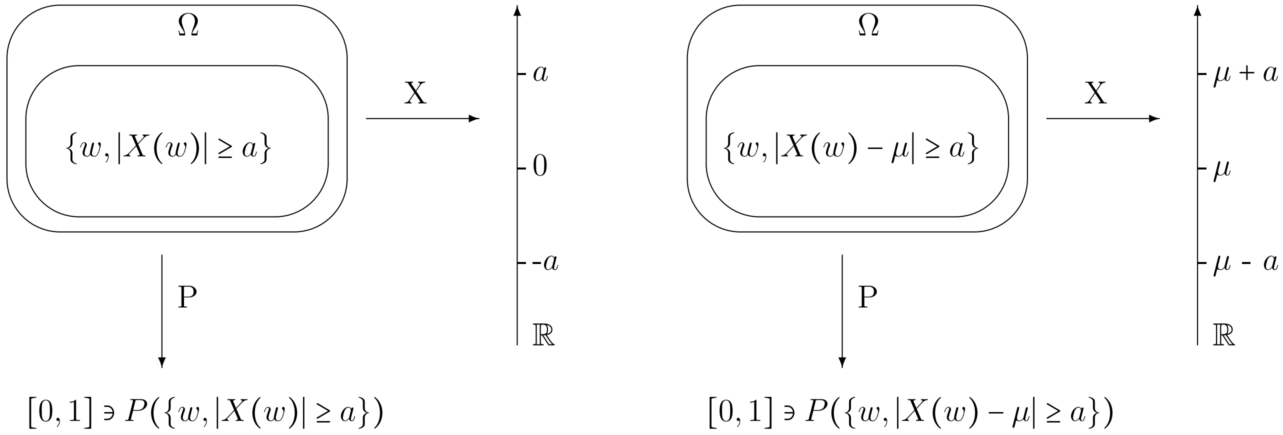
5.1.2. *Chebyshev's Inequality.* How effective is Markov's inequality when $a \leq \mu$? Not much, you can see. However there is a useful corollary of Markov's inequality, called Chebyshev's inequality, which bounds the probability that X can be some distance away from its mean by its variance,

Corollary 5.4. *If X is a random variable and $a > 0$ then*

$$P(|X - \mu| \geq a) \leq \frac{var(X)}{a^2}$$

.

Proof. Let $Y = (X - \mu)^2$ then $P(|X - \mu| \geq a) = P(Y \geq a^2) \leq \frac{E(Y)}{a^2} = \frac{\text{var}(X)}{a^2}$ by Markov's inequality. □


Ω
$\{w, |X(w)| \geq a\}$
X
a
0
-a
ℝ
P
$[0,1] \ni P(\{w, |X(w)| \geq a\})$
Ω
$\{w, |X(w) - \mu| \geq a\}$
X
$\mu + a$
μ
μ - a
ℝ
P
$[0,1] \ni P(\{w, |X(w) - \mu| \geq a\})$

Exercise 5.5. The probability that a random variable X can be twice its standard deviation away from its mean is $P(|X - \mu| \geq 2\sigma)$.

a. Bound $P(|X - \mu| \geq 2\sigma)$ by Chebyshev's inequality.
b. Calculate $P(|X - \mu| \geq 2\sigma)$ when X is uniformly distributed over $[5, 9]$ and compare with (a).
c. Calculate $P(|X - \mu| \geq 2\sigma)$ when X is normally distributed with mean 1, standard deviation 3 and compare with (a).

Example 5.6. Each Bernoulli trial X_i has success rate p. How many trials are needed to be at least 90% sure that our estimate $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$ is within 0.02 of its true value?

So each trial X_i has $\mu(X_i) = p$, $\text{var}(X_i) = p(1-p)$ and $\sigma(X_i) = \sqrt{p(1-p)}$. Our average $\overline{X}_n$ has $\mu(\overline{X}_n) = p$ and $\text{var}(\overline{X}_n) = \frac{p(1-p)}{n}$. We want $P(|\overline{X}_n - \mu(\overline{X}_n)| < 0.02) \geq 0.9$ or equivalently $P(|\overline{X}_n - \mu(\overline{X}_n)| \geq 0.02) < 0.1$. By Chebyshev's inequality

$$\begin{aligned} P(|\overline{X}_n - \mu(\overline{X}_n)| \geq 0.02) &\leq \frac{\mathrm{var}(\overline{X}_n)}{(0.02)^2} \\ &= \frac{\frac{p(1-p)}{n}}{(0.02)^2} \\ &= \frac{10000p(1-p)}{4n} \end{aligned}$$

Thus we need $\frac{10000p(1-p)}{4n} < 0.1$ or $n > \frac{100000p(1-p)}{4} = 25000p(1-p)$

a. If they say $p = 1/5$ for example then $n \geq 4000$ will do.
b. If p is not given then $p(1-p)$ is at most $1/4$ anyway, and $n \geq 6250$ will do.

5.1.3. *Hoeffding's Inequality.* Stronger than both Markov's bound and Chebyshev's bound is Hoeffding's inequality, which bounds the probability that the average of finitely many random variables deviates from its mean by some nonnegative ϵ.

Proposition 5.7. *If the X_i take values in $[a_i, b_i]$ and $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$ with mean μ then*

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{\frac{-2n^2\epsilon^2}{\Sigma(b_i-a_i)^2}}$$

Proof. see literature. □

Example 5.8. If we use Hoeffding's inequality for example 5.6 then we get

$$P(|\overline{X}_n - \mu| \geq 0.02) \leq 2e^{-2n(0.02)^2}$$

So we need $2e^{-2n(0.02)^2} \leq 0.1$, or $n \geq 1873$. This is independent of p, compare this with n in example 5.6.

Exercise 5.9. Define the domain $(\Omega, \mathcal{F}, P)$ for $\overline{X}_n$ in example 5.6. Understand what it means to be at least 90% sure that $\overline{X}_n$ is within 0.02 of its true value. Understand why it depends on n.

5.1.4. *Chernoff's Inequality.* There are many variants of this inequality, here is a version to which we can give examples.

Proposition 5.10. *If X is a random variable and $\delta > 0$ then*

$$P(X \geq \mu(1+\delta)) \leq e^{-\frac{\delta^2\mu}{3}}$$

Proof. see literature. □

Example 5.11. We use Chernoff's inequality to bound the percentage of population with more than 5 times the average income

$$P(X \geq \mu(1+4)) \leq e^{-\frac{4^2\mu}{3}}$$

as in example 5.2. This is much better than 1/5.

Example 5.12. Suppose we flip a fair coin n times and count the number of tails by X.

a. We bound the chances that more than 75% of the flips are tails. Since X has mean $\mu = \frac{n}{2}$

$$\begin{aligned} P(X \geq 0.75n) &= P(X \geq \mu(1+0.5)) \\ &\leq e^{-\frac{0.5^2 n}{6}} \end{aligned}$$

by Chernoff's inequality.

b. We bound the number of tosses needed to be at least 90% sure that our estimate is within 5 of the true value. By Chernoff's inequality

$$\begin{aligned} P\left(|X - \frac{n}{2}| \geq 5\right) &= 2P\left(X \geq \frac{n}{2} + 5\right) \\ &= 2P\left(X \geq \frac{n}{2}\left(1 + \frac{10}{n}\right)\right) \\ &\leq 2e^{-\left(\frac{10}{n}\right)^2 \frac{n}{6}} \end{aligned}$$

Therefore we need $2e^{-\left(\frac{10}{n}\right)^2 \frac{n}{6}} \leq 0.1$ or $n \geq 6$.

5.1.5. *Jensen's Inequality.* Last of our list is Jensen's inequality. It relates the mean of a random variable X with the mean of a function g of X.

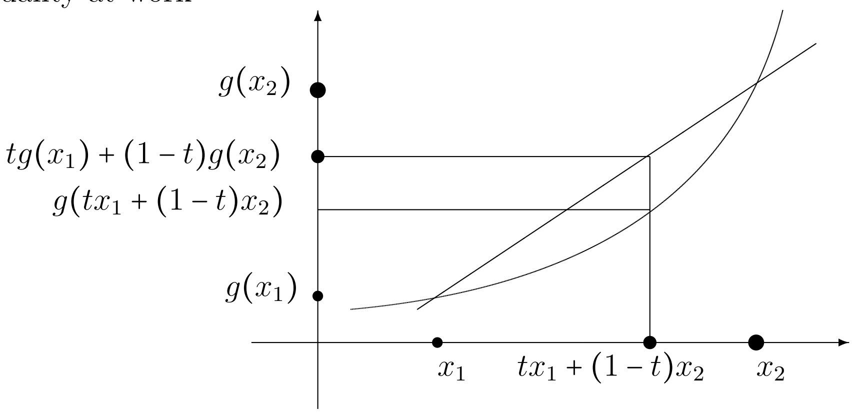
Proposition 5.13. *If X is a random variable and $(\mathbb{R}, \mathcal{B}(\mathbb{R})) \xrightarrow{g(u)} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measurable function then*

$$E(g(X)) \leq g(E(X)) \text{ if } g \text{ is concave.}$$

$$E(g(X)) \geq g(E(X)) \text{ if } g \text{ is convex.}$$

Proof. This is Jensen's inequality in analysis with $E(g(X)) = \int_\Omega g(X)(\omega)\, dP$ and $P(\Omega) = 1$ finite. □

If g is convex and X is discrete with $p_X(x_1) = t$ and $p_X(x_2) = 1 - t$ then we can see Jensen's inequality at work



Example 5.14. Let X be exponentially distributed with $\lambda = 10^2$ and $\mu = \frac{1}{10^2}$. Then

$$\begin{aligned} E(\log_{10}(X)) &\leq \log_{10}(E(X)) \\ &= \log_{10}(\frac{1}{10^2}) \\ &= -2 \end{aligned}$$

where $g(u) = \log_{10}(u)$ is concave. On the other hand

$$\begin{aligned} E(X^3) &= E(g(X)) \\ &\geq g(E(X)) \\ &= \frac{1}{10^6} \end{aligned}$$

where $g(u) = u^3$ is convex.

5.2. **Equalities for Random Variables.** We know what it means for two real numbers to be equal. We also know what it means for them to be close, or for a sequence of them to approach a certain real number.

Definition 5.15. We say a sequence of real numbers $\{x_n\}_{n\geq 1}$ converges to x if any open set $U \ni x$ contains all but finitely many x_n, or equivalently, if given $\epsilon > 0$ there exists a k such that $|x - x_n| \leq \epsilon$ for all $n \geq k$. Else we say $\{x_n\}_{n\geq 1}$ diverges.

Example 5.16. The sequence $\{1/n\}_{n\geq 1}$ converges to 0 while the sequence $0, 1, 0, 0, 1, 0, 0, 0, 1, \ldots$ diverges.

We look to do the same with random variables $\{X_n\}_{n\geq 1}$ on the same probability space $(\Omega, \mathcal{F}, P)$. For each $w \in \Omega$ we have the sequence of real numbers $\{X_n(w)\}_{n\geq 1}$ and this is our starting point.

Definition 5.17. We say two random variables X and Y are equal surely (or everywhere) if $X(w) = Y(w)$ for all $w \in \Omega$.

So X and Y are equal surely if the event $\{w, X(w) = Y(w)\} = \Omega$ and the event $\{w, X(w) \neq Y(w)\} = \varnothing$. However, in the presence of P we can consider equality of random variables in a weaker sense.

Definition 5.18. We say two random variables X and Y are equal almost surely (or almost everywhere) if $P(\{w, X(w) = Y(w)\}) = 1$.

Equivalently, X and Y are equal almost surely if they differ only on a set of measure 0, or $P(\{w, X(w) \neq Y(w)\}) = 0$.

Example 5.19. We can modify any continuous random variable X with pdf f_X into another random variable Y with $Y(w_0) = X(w_0) + 1$ for one w_0 with the same pdf $f_Y = f_X$. Then $\{w, X(w) \neq Y(w)\} = \{w_0\} \neq \varnothing$ has measure 0 so X and Y are not equal surely but they are equal almost surely.

We can think of other modes of equality between random variables. For example, we can say X and Y are equal in mean if $E(X) = E(Y)$ but such definition is not very useful as X and Y in that case can be vastly different. On the other hand, if we only judge equality between two random variables pointwise, we are too strict and missing the point of probability theory. It places emphasis on the distribution of a random variable plus its invariants instead of on what it does pointwise.

Definition 5.20. We say two random variables X and Y are equal in distribution if $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$.

Equality in distribution means the two cumulative distribution functions F_X and F_Y are equal everywhere. It follows that $P(X \in A) = P(Y \in A)$ for all $A \in \mathcal{B}(\mathbb{R})$, or equivalently $X_*(P) = Y_*(P)$ as measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Example 5.21. Consider $\Omega = \{H, T\}$ with $\mathcal{F} = \mathbb{P}(\Omega)$ and $P(H) = P(T) = \frac{1}{2}$. If X maps H to -1 and T to 1 while Y maps H to 1 and T to -1 then $f_X = f_Y$ and X, Y are equal in distribution. However, they are not equal surely or equal almost surely. One can think of the amount of winning of the dealer and the amount of winning of the player when they toss a fair coin.

Next we consider a sequence $(\Omega, \mathcal{F}, P) \xrightarrow{\{X_n\}_{n\geq 1}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ of random variables from the same probability space to the same state space and what it means for them to converge to a certain random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

5.3. **Convergence Surely.** The first type of convergence of random variables is sure convergence, in which we view a sequence of random variables as a sequence of maps, and see when they converge pointwise.

Definition 5.22. We say a sequence of random variables $\{X_n\}_{n\geq 1}$ converges surely (or everywhere, or pointwise) to X, written $X_n \xrightarrow{s} X$, if $\{X_n(w)\}_{n\geq 1}$ converges to $X(w)$ for all $w \in \Omega$.

In sure convergence, the event $\{w \in \Omega, \{X_n(w)\}_{n\geq 1} \text{ converges to } X(w)\}$ is all of Ω, or equivalently the event $\{w, \{X_n(w)\}_{n\geq 1} \text{ does not converge to } X(w)\}$ is empty.

Example 5.23. If we take the sequence $\{\frac{1}{n}\}_{n\geq 1}$ in example 5.16 as a sequence of constant random variables $(\Omega, \mathcal{F}, P) \xrightarrow{1/n} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ then they converge to the constant random variable 0. For any $\omega \in \Omega$, clearly $\frac{1}{n}(\omega) = \frac{1}{n}$ goes to 0 as n goes to infinity.

Example 5.24. Let X be any random variable, then the sequence $\{\frac{1}{n} 1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}\}_{n\geq 1}$ converges surely to the constant random variable 0. For ω such that $X(\omega) \neq 0, 1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}(\omega) = 0$ for n large, hence $\frac{1}{n} 1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}(\omega) = 0$ for n large. For ω such that $X(\omega) = 0, 1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}(\omega) = 1$ for all n, hence $\frac{1}{n} 1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}(\omega)$ goes to 0 as n goes to infinity. A sequence of continuous random variables may converge surely to a discrete random variable.

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "\frac{1}{n} 1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}"] \arrow[dr, "X"'] & (\mathbb{R}, \mathcal{B}(\mathbb{R}) \\
& (\mathbb{R}, \mathcal{B}(\mathbb{R})) \arrow[u, "\frac{1}{n} 1_{[-\frac{1}{n}, \frac{1}{n}]}"']
\end{tikzcd}$$

Exercise 5.25. Let X be any random variable, show that the sequence $\{1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}\}_{n\geq 1}$ converges surely. What is its limit?

Exercise 5.26. Let X be any random variable such that $\{\omega, X(\omega) = 0\} \neq \varnothing$, show that the sequence $\{n\,1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}\}_{n\geq 1}$ does not converge surely.

5.4. **Convergence Almost Surely.** Because $(\Omega, \mathcal{F}, P)$ is a measure space with probability measure P, we can relax our definition of sure convergence and see when a sequence of random variables converges pointwise, give or take a negligible set.

Definition 5.27. We say a sequence of random variables $\{X_n\}_{n\geq 1}$ converges almost surely (or almost everywhere, or strongly) to X, written $X_n \xrightarrow{a.s.} X$, if $P(\{w \text{ such that } X_n(w) \to X(w) \text{ as } n \to \infty\}) = 1$.

In almost sure convergence, $\{X_n\}$ converges to X surely on a set of measure 1, or equivalently they do not converge to X on a set of measure 0. Any sequence of random variables that converges surely certainly converges almost surely. Here are some less trivial cases.

Exercise 5.28. Reconsider the sequence $\{n\,1_{\{-\frac{1}{n} \leq X \leq \frac{1}{n}\}}\}_{n\geq 1}$ in exercise 5.26, show that

a. It converges almost surely if X is continuous.
b. It may not converge almost surely if X is discrete. Hint: define X such that $\{\omega, X(\omega) = 0\}$ has positive measure.

The set $\{\omega, \{X_n(\omega)\}_{n\geq 1} \text{ does not converge}\}$ when it is nonempty yet negligible is the difference between almost sure convergence and sure convergence. Thus we can redefine each X_n on this set to go from almost sure convergence to sure convergence without changing the distribution of X_n.

5.5. **Convergence in Distribution.** As with equalities for random variables, we focus on how a sequence of random variables converges with respect to its distribution and invariants.

Definition 5.29. We say a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in distribution (or in law) to X, written $X_n \xrightarrow{d} X$, if $F_{X_n}(x)$ converges to $\{F_X(x)\}_{n\geq 1}$ for all $x \in \mathbb{R}$ where F_X is continuous.

It follows from the definition that $\{P(X_n \in A)\}_{n\geq 1}$ converges to $P(X \in A)$ for all $A \in \mathcal{B}(\mathbb{R})$, or in other words $\{X_{n*}(P)\}_{n\geq 1}$ converges to $X_*(P)$ as measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. In the continuous case, this means the areas under $f_{X_n}, n \geq 1$ over any set A converges to the area under f_X over the same set. This convergence of probabilities is useful when we need to approximate some probabilities for the discrete X_n when they converge in distribution to continuous X, as we shall see in an exercise.

Here are some nice ways to help us verify convergence in distribution.

Theorem 5.30. *For a sequence of random variables $\{X_n\}_{n\geq 1}$, the following are equivalent,*

1. *$\{X_n\}_{n\geq 1}$ converges in distribution to random variable X.*
2. *$\{E(f(X_n))\}_{n\geq 1}$ converges to $E(f(X))$ for all bounded continuous map $\mathbb{R} \xrightarrow{f} \mathbb{R}$.*
3. *$\{k_{X_n}\}_{n\geq 1}$ converges to k_X pointwise.*
4. *$\limsup_{n\to\infty} P(X_n \in A) \leq P(X \in A)$ for all closed subsets $A \subset \mathcal{E}(\mathbb{R})$.*

Proof. This follows from the portmanteau lemma in measure theory. Between (1) and (2) we have $\lim\limits_{n\to\infty} \int\limits_A f_{X_n}(s)ds = \int\limits_A f_X(s)ds$ for all $A \in \mathcal{B}(\mathbb{R})$ is equivalent to $\lim\limits_{n\to\infty} \int\limits_{\mathbb{R}} f(s)f_{X_n}(s)ds = \int\limits_{\mathbb{R}} f(s)f_X(s)ds$ for all bounded continuous f. □

None of the above criteria is equivalent to $\{p_{X_n}\}_{n\geq 1}$ converges to p_X pointwise in the discrete case or $\{f_{X_n}\}_{n\geq 1}$ converges to f_X pointwise in the continuous case. We can not use $f = id_{\mathbb{R}}$ to deduce $\{E(X_n)\}_{n\geq 1}$ converges to $E(X)$ because $id_{\mathbb{R}}$ is not bounded. It takes more than convergence in distribution to get convergence of means or convergence of variances. For now, let us see some examples.

Example 5.31. In convergence in distribution, we only require that the $\{F_{X_n}\}_{n\geq 1}$ converges to F_X where F_X is continuous. Each constant random variable $X_n = 1 + \frac{1}{n}$ has CDF

$$F_{X_n}(x) = \begin{cases} 0 \text{ if } x < 1 + \frac{1}{n} \\ 1 \text{ if } x \geq 1 + \frac{1}{n} \end{cases}$$

So $\{F_{X_n}\}_{n\geq 1}$ converges to F' where

$$F'(x) = \begin{cases} 0 \text{ if } x \leq 1 \\ 1 \text{ if } x > 1 \end{cases}$$

Notice that this is not a CDF as it is not right continuous at 1. With a small modification, it becomes the CDF of the constant random variable $X = 1$,

$$F_X = \begin{cases} 0 \text{ if } x < 1 \\ 1 \text{ if } x \geq 1 \end{cases}$$

So $\{X_n\}_{n\geq 1}$ converges in distribution to X even though $\{F_{X_n}(1)\}_{n\geq 1}$ does not converge to $F_X(1)$.

Example 5.32. Each uniform random variable $(X_n, 0, \frac{1}{n})$ has CDF

$$\begin{aligned} F_{X_n}(x) &= \int\limits_{\infty}^{x} f_{X_n}(s)ds \\ &= \int\limits_{-\infty}^{x} n\,1_{[0,\frac{1}{n}]}(s)ds \\ &= n\mu_L([0,x] \cap [0, \frac{1}{n}]) \\ &= \begin{cases} 0 \text{ if } x \leq 0 \\ nx \text{ if } 0 < x < \frac{1}{n} \\ 1 \text{ if } x \geq \frac{1}{n} \end{cases} \end{aligned}$$

So $\{F_{X_n}\}_{n\geq 1}$ converges to F' where

$$F'(x) = \begin{cases} 0 \text{ if } x \leq 0 \\ 1 \text{ if } x > 0 \end{cases}$$

Again this is not a CDF as it is not right continuous at 0. However, $\{X_n\}_{n\geq1}$ still converges in distribution to the constant variable $X = 0$ with CDF

$$F_X(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ if } x \geq 0 \end{cases}$$

Exercise 5.33. Show that the sequence of uniform random variables $\{(X_n, 0, n)\}_{n\geq1}$ does not converge in distribution to any X. Hint: compute and graph F_{X_n} to see their limit.

Exercise 5.34. Let $Y_n = \frac{1}{n}X_n$ where X_n is the geometric random variable with parameter $p = \frac{\lambda}{n}$. Note that $E(Y_n) = \frac{1}{\lambda}$ independent of n.

a. Show that $\{Y_n\}_{n\geq1}$ converges in distribution to the exponential random variable $(Y, \frac{1}{\lambda})$. A sequence of discrete random variables may converge in distribution to a continuous random variable. Hint: compute $1 - F_{X_n}(x)$ and take $\log_e$ to see their limit.
b. As a reward, approximate $P(0 \leq Y_n \leq \lambda)$ for large n.

Exercise 5.35. Prove that the sequence of normal random variables $\{(X_n, \frac{1}{n}, 1)\}_{n\geq1}$ converges in distribution to the standard normal random variable $(X, 0, 1)$.

Exercise 5.36. Let X_n be the continuous random variable with pdf $f_{X_n}(s) = (1 - \cos(2\pi ns))\, 1_{(0,1)}(s)$ and $(X, 0, 1)$ be the uniform random variable. Show that

a. $\{X_n\}_{n\geq1}$ converges in distribution to X.
b. $\{f_{X_n}\}_{n\geq1}$ does not converge to f_X. Convergence in distribution does not imply convergence of pdfs.

5.6. **Convergence in Probability.** Stronger than convergence in distribution is convergence in probability. It says the probability of the event $|X_n - X| \geq \epsilon$ will decrease to 0 as the sequence progresses.

Definition 5.37. We say a sequence of random variables $\{X_n\}_{n\geq1}$ converges in probability (or weakly) to X, written $X_n \xrightarrow{P} X$, if for each $\epsilon > 0$ we have $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.

Equivalently, the probability of the event $|X_n - X| < \epsilon$ will increase to 1 as $n \to \infty$.

Example 5.38. Let X_n be the Bernoulli random variable with success rate $\frac{1}{n}$. Then $P(|X_n - 0| \geq \epsilon) = P(X \geq \epsilon) = \frac{1}{n}$ goes to 0 as n goes to ∞. Hence $\{X_n\}_{n\geq1}$ converges in probability to 0. One also sees that $\{E(X_n^k)\}_{n\geq1}$ converges to $E(X^k)$ for all $k \geq 1$. This belongs to a more general result about convergence of k^{th} moments that we will see in proposition 5.50 and the ensuing discussion.

Example 5.39. Let Y_n be the minimum of the sequence of independent uniform random variables $(X_1, 0, 1), \ldots, (X_n, 0, 1)$. Let $C = \bigcup_{n\geq1}\{X_n \notin [0, 1])\}$ then

$$P(C) \leq \sum_{n\geq1} P(X_n \notin [0, 1]) = 0$$

hence $P(C) = 0$. Disregarding all $\omega \in C$, we have $P(|Y_n - 0| \geq \epsilon) = P(Y_n \geq \epsilon) = P(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) = (1 - \epsilon)^n$ goes to 0 as n goes to ∞. Hence $\{Y_n\}_{n\geq1}$ converges in probability to 0.

Example 5.40. Let an archer shoot arrows at a target and let $\{X_n\}_{n\geq 1}$ be the scores of his shots in month n^{th}. Over time he will improve and the probability that he misses will decrease. So $\{X_n\}_{n\geq 1}$ converges in probability to $X = 0$. But fix a k^{th} shot in each month then that shot will miss every few months, however infrequent that is. Thus the measure of the set of shots that always score 0 will be less than 1 and $\{X_n\}_{n\geq 1}$ does not converge almost surely to X. We make this more rigorous. Let $\{X_n\}_{n\geq 1}$ be the distances of the shots to the center in month n^{th}. In the zeroth month the shots could fall anywhere so X_0 is the uniform bivariate random variable over the unit disk with pdf $f_{X_0}(s) = \frac{1}{\pi}$. But $X_n, n \geq 1$ must reflect this shooter's improvement. Thus we may choose $\{X_n\}_{n\geq 1}$ with pdf $\{f_{X_n}\}_{n\geq 1}$ whose graphs are increasingly taller domes over the unit disk and show they converge in probability to 0 but not almost surely. In either cross section, they can be chosen to look like $f_{X_n}(s) = \frac{1}{\sqrt{2\pi\sigma(n)}}e^{-s^2/2\sigma^2(n)}$ where $\sigma(n)$ decreases. This leads us to 2-dimensional normal random variables with mean $\mu = 0$, decreasing standard deviations $\sigma_1(n), \sigma_2(n)$, and pdfs $f_{X_n}(s,t) = \frac{1}{2\pi\sigma_1(n)\sigma_2(n)}e^{-\left(\frac{s^2}{2\sigma_1^2(n)}+\frac{t^2}{2\sigma_2^2(n)}\right)}$ Note that we have $\sigma_1(n), \sigma_2(n)$ to account for the archer's tendency to spread his shots vertically or horizontally. Then $P(\{w \text{ such that } |X_n(w) - 0| \geq \epsilon\}) \to 0$ as $n \to \infty$ while $P(\{w \text{ such that } X_n(w) \to 0 \text{ as } n \to \infty\}) < 1$.

We have the following comparisons between modes of convergence.

Proposition 5.41. *If a sequence of random variables $\{X_n\}_{n\geq 1}$ converges almost surely to X then it converges in probability to X. The converse is not true.*

Proof. If $\{X_n\}_{n\geq 1}$ converges almost surely to X and $\epsilon > 0$ then for any $\omega \in F = \{\omega, \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$, there exists N such that $|X_n(\omega) - X(\omega)| < \epsilon$ for all $n \geq N$. Hence $\omega \notin A_m = \bigcup_{n\geq m} \{|X_n - X| \geq \epsilon\}$ for $m \geq N$. So $A_m \subset F^c$ for $m \geq N$ of measure 0. It follows $P(|X_n - X| \geq \epsilon) \leq P(A_m) = 0$ for $n \geq m \geq N$, or $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$ and $\{X_n\}_{n\geq 1}$ converges in probability to X. □

The next example completes proposition 5.41.

Example 5.42. Reconsider the independent Bernoulli random variables X_n with success rate $\frac{1}{n}$ in example 5.38. There they converge in probability to 0. That they do not converge almost surely to 0 follows from Borel-Cantelli lemma.

Proposition 5.43. *If a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in probability to X then it converges in distribution to X. As a partial converse, if it converges in distribution to constant $X = c$ then it converges in probability to constant $X = c$.*

Proof. For any $\delta, \epsilon > 0$, we have

$$\begin{aligned} P(X_n \leq x) &= P(\{X_n \leq x\} \cap \{|X_n - X| \leq \delta\}) + P(\{X_n \leq x\} \cap \{|X_n - X| > \delta\}) \\ &\leq P(X \leq x + \delta) + P(|X_n - X| > \delta) \end{aligned}$$

By right continuity of F_X, we can choose δ small enough so that $P(X \leq x + \delta) \leq P(X \leq x) + \frac{\epsilon}{2}$. By convergence in probability of $\{X_n\}_{n\geq 1}$ to X, we can choose n large enough so that $P(|X_n - X| > \delta) < \frac{\epsilon}{2}$. Hence, we have $F_{X_n}(x) \leq F_X(x) + \epsilon$ for n large. Similarly, we have $F_{X_n}(x) \geq F_X(x) - \epsilon$ for n large. This shows $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ wherever F_X is continuous.

For the partial converse

$$\begin{aligned}\lim_{n\to\infty} P(|X_n - c| \geq \epsilon) &\leq \limsup_{n\to\infty} P(|X_n - c| \geq \epsilon)\\ &= \limsup_{n\to\infty} P(X \notin [c-\epsilon, c+\epsilon])\\ &\leq P(c \notin [c-\epsilon, c+\epsilon])\\ &= 0\end{aligned}$$

by theorem 5.30[4]. Hence $\{X_n\}_{n\geq 1}$ converges in probability to $X = c$. □

By this proposition, all sequences in the examples in this subsection can serve as examples of convergence in distribution. And all sequences in the examples in the previous subsection converge in distribution to constants, hence they can serve as examples of convergence in probability.

5.7. **Convergence in k^{th} Moment.** We consider one more mode of convergence for random variables.

Definition 5.44. We say a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in k^{th} moment (or in L^k) to X, written $X_n \xrightarrow{L^k} X$, if the k^{th} absolute moments $E(|X_n|^k), E(|X|^k)$ exist for all n and $E(|X_n - X|^k) \to 0$ as $n \to \infty$.

Example 5.45. Let $X_1, X_2, \dots$ be independent identically distributed random variables with mean μ and standard deviation σ. Then $E(|\overline{X}_n - \mu|^2) = E((\overline{X}_n - \mu)^2) = \operatorname{var}(\overline{X}_n) = \operatorname{var}(\frac{X_1+\cdots+X_n}{n}) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$ goes to 0 as n goes to ∞. Hence $\{\overline{X}_n\}_{n\geq 1}$ converges in second moment to the constant random variable μ.

Proposition 5.46. *If a sequence of random variables $\{X_n\}_{n\geq 1}$ converges in l^{th} moment to X and $k < l$ then it converges in k^{th} moment to X.*

Proof. The map $g(u) = u^{\frac{k}{l}}$ is concave because $\frac{k}{l} < 1$. By Jensen's inequality for the continuous case and integration

$$\begin{aligned}E(|X_n - X|^k) &= \int_\Omega |X_n(\omega) - X(\omega)|^k \, dP\\ &= \int_\Omega (|X_n(\omega) - X(\omega)|^l)^{\frac{k}{l}} \, dP\\ &\leq \left(\int_\Omega |X_n(\omega) - X(\omega)|^l \, dP\right)^{\frac{k}{l}}\\ &= E(|X_n - X|^l)^{\frac{k}{l}}\end{aligned}$$

goes to 0 as n goes to ∞. Proof for the discrete case and summation works similarly. □

The converse to this proposition is not true. However, convergence in any k^{th} moment does imply convergence in probability.

Proposition 5.47. *A sequence of random variables $\{X_n\}_{n\geq 1}$ converges in first moment to X iff it is uniformly integrable and converges in probability to X.*

Proof. If $\{X_n\}_{n\geq 1}$ converges in first moment to X then by Markov's inequality, $P(|X_n - X| \geq \epsilon) \leq \frac{E(|X_n-X|)}{\epsilon}$ goes to 0 as n goes to ∞. Hence $\{X_n\}_{n\geq 1}$ converges in probability to X. Proving uniformly integrability of the X_n or proving the converse would take us too far into analysis and measure theory. We are looking at the equivalence between $\{\int\limits_{\Omega} |X_n(\omega) - X(\omega)|\, dP\}_{n\geq 1}$ converges to 0 and $\{\int\limits_{\Omega} 1_{\{|X_n-X|\geq\epsilon\}}\, dP\}_{n\geq 1}$ converges to 0 plus uniform integrability of $\{X_n\}_{n\geq 1}$. □

Exercise 5.48. The hypothesis of uniform integrability is important. Let $Y_n = nX_n$ where X_n is the Bernoulli random variable with success rate $\frac{1}{n}$. Show that

a. $\{Y_n\}_{n\geq 1}$ converges in probability to 0.
b. $\{Y_n\}_{n\geq 1}$ does not converge in any moment to 0.

Exercise 5.49. There is no relationship between convergence in moments and convergence almost surely.

a. The sequence $\{X_n\}_{n\geq 1}$ in example 5.38 is uniformly integrable, hence it converges in first moment to 0 by the proposition. In fact, it converges in all k^{th} moments to 0 because $E(|X_n - 0|^k) = E(X^k) = 0^k\frac{n-1}{n} + 1^k\frac{1}{n} = \frac{1}{n}$ goes to 0 as n goes to ∞. Show that it does not converge almost surely to 0.
b. Let X_n be the discrete random variable with $p_{X_n}(0) = 1 - \frac{1}{n}$ and $p_{X_n}(n) = \frac{1}{n}$. Show that $\{X_n\}_{n\geq 1}$ converges almost surely to 0 but it does not converge in second moment to 0.

Finally, we relate convergence in k^{th} moment with convergence of k^{th} *absolute* moments, convergence of k^{th} moments, convergence of second moments (or convergence of mean squares), convergence of first moments (or convergence of means), and convergence of variances.

Proposition 5.50. *If $\{X_n\}_{n\geq 1}$ converges in k^{th} moment to X then $\{E(|X_n|^k)\}_{n\geq 1}$ converges to $E(|X|^k)$.*

Proof. We apply Minkowski inequality to the continuous case and integration

$$\begin{aligned} E(|X_n|^k)^{\frac{1}{k}} &= \left(\int\limits_{\mathbb{R}} |X_n|^k\, dP\right)^{\frac{1}{k}} \\ &\leq \left(\int\limits_{\mathbb{R}} |X_n - X|^k\, dP\right)^{\frac{1}{k}} + \left(\int\limits_{\mathbb{R}} |X|^k\, dP\right)^{\frac{1}{k}} \\ &= E(|X_n - X|^k)^{\frac{1}{k}} + E(|X|^k)^{\frac{1}{k}} \end{aligned}$$

Hence $|E(|X_n|^k)^{\frac{1}{k}} - E(|X|^k)^{\frac{1}{k}}| \leq E(|X_n - X|^k)^{\frac{1}{k}}$ goes to 0 as n goes to ∞. It follows $E(|X_n|^k)$ goes to $E(|X|^k)$ as n goes to ∞. Proof for the discrete case and summation works similarly. □

One would hope that convergence of $\{E(|X_n|)\}_{n\geq 1}$ to $E(|X|)$ then implies convergence of $\{E(X_n)\}_{n\geq 1}$ to $E(X)$. This is not true in general, just look at $X_n = (-1)^n$ and $X = 1$.

When $k = 1$, convergence of first moments follows from convergence in first moment, as $|E(X_n) - E(X)| = |E(X_n - X)| \leq E(|X_n - X|)$ goes to 0 as n goes to ∞. One can observe this in some examples above. When $k = 2n$ even, convergence of k^{th} moments follows from convergence of k^{th} absolute moments because we can drop the absolute value sign. In particular, for $k = 2$ convergence in second moment implies convergence of second moments, which involve variances.

Corollary 5.51. *If $\{X_n\}_{n\geq1}$ converges in second moment to X then $\{var(X_n)\}_{n\geq1}$ converges to $var(X)$.*

Proof. By the previous discussion, $|\text{var}(X_n) - \text{var}(X)| = |E(X_n^2) - E(X_n)^2 - E(X^2) + E(X)^2| \leq |E(X_n^2) - E(X^2)| + |E(X_n)^2 - E(X)^2|$ goes to 0 as n goes to ∞. □

Example 5.52. Let X_n be the discrete random variable with $p_{X_n}(\frac{1}{n}) = p_{X_n}(\frac{-1}{n}) = \frac{1}{2}$. Then $E(|X_n - 0|^k) = E(|X_n|^k) = \frac{1}{n^k}\frac{1}{2} + \frac{1}{n^k}\frac{1}{2} = \frac{1}{n^k}$ goes to 0 as n goes to ∞. So $\{X_n\}_{n\geq1}$ converges in all k^{th} moment to $X = 0$. Convergence of first moments, convergence of all even moments and convergence of variances follow while all odd moments are 0.

Exercise 5.53. Let X_n be the discrete random variable with $p_{X_n}(0) = \frac{1}{n}$ and $p_{X_n}(\frac{1}{n}) = \frac{n-1}{n}$. Show that $\{X_n\}_{n\geq1}$ converges in all k^{th} moment to $X = 0$. Deduce that $\{\text{var}(X_n)\}_{n\geq1}$ converges to $\text{var}(X)$.

There remains a lot to discuss about convergence of random variables. For example, one can reevaluate the sequences in previous examples for different modes of convergence. Or one can consider what happens when we change the X_n by a function g. Or one can consider convergence for multivariate random variables in the joint and marginal cases. It goes on and on. We now list two important theorems that are well stated in the language of convergence.

5.8. **Laws of Large Numbers.** They say that the average $\overline{X}_n$ of independent identically distributed variables $X_1, \ldots, X_n$ goes to their mean μ as n gets large.

Theorem 5.54. *(Weak Law of Large Numbers) If $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables with expected value μ then $\{\overline{X}_n\}_{n\geq1}$ converges in probability (or converges weakly, hence the name) to μ.*

Proof. By example 5.45, $\{\overline{X}_n\}_{n\geq1}$ converges in second moment to μ. By proposition 5.46, it converges in first moment to μ. By proposition 5.47, it converges in probability to μ. □

This law means $P(|\overline{X}_n - \mu| > \epsilon)$ goes to 0 as n goes to ∞, or equivalently $P(|\overline{X}_n - \mu| \leq \epsilon)$ goes to 1 as n goes to ∞, which confirms our intuition when we play any game of chance multiple times. If ϵ is small then it may take longer. When $\text{var}(X_i)$ is finite, Chebyshev's inequality reveals how fast at least this convergence happens.

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\text{var}(X_i)}{n\epsilon^2}$$

Since the limit $X = \mu$ is constant, the fact that $\{\overline{X}_n\}_{n\geq1}$ converges in probability to constant μ is equivalent to $\{\overline{X}_n\}_{n\geq1}$ converges in distribution to μ and so is equivalent to $\{k_{X_n}\}_{n\geq1}$ converges to $k(t) = e^{i\mu t}$.

Exercise 5.55. Prove the Weak Law of Large Numbers via convergence of characteristic functions.

Theorem 5.56. *(strong law of large numbers) If $X_1, X_2, \dots$ is a sequence of independent identically distributed random variables with expected value μ then $\{\overline{X}_n\}_{n\geq 1}$ converges almost surely (or converges strongly, hence the name) to μ.*

Proof. more difficult. □

The strong law of large numbers means $P(\{w, \overline{X}_n(w)$ converges to $\mu\}) = 1$, or equivalently $P(\{w, \overline{X}_n(w)$ does not converge to $\mu\}) = 0$. It implies the weak law of large numbers by proposition 5.41. The weak law of large numbers leaves open the possibility that the set $\{|\overline{X}_n - \mu| > \epsilon\}$ has positive measure even for large n. On the other hand, the strong law of large numbers says that the set $\{|\overline{X}_n - \mu| > \epsilon\}$ has measure 0 for all n large enough.

5.9. **Central Limit Theorem.** The following theorem has many applications in applied probability and statistics.

Theorem 5.57. *(central limit theorem) Let $X_1, \dots, X_n$ be a sequence of independent identically distributed random variables with expected value μ and variance σ^2. Then $\left\{\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}\right\}_{n\geq 1}$ converges in distribution to the standard normal random variable $Z = (X, 0, 1)$.*

Proof. We show convergence in distribution via convergence of characteristic functions. Let $Y_i = \frac{X_i - \mu}{\sigma}$ with $\mu_1(Y_i) = 0$ and $\sigma(Y_i) = 1$. Then $\mu_2(X) = 1$ and so Y_i has characteristic function $k_{Y_i}(t) = 1 - \frac{t^2}{2} + t^3(...)$. Write

$$Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{n\overline{X}_n - n\mu}{\sqrt{n}\sigma} = \sum_{i=1}^{n} \frac{\frac{X_i - \mu}{\sigma}}{\sqrt{n}} = \sum_{i=1}^{n} \frac{Y_n}{\sqrt{n}}$$

By theorem 4.119 for characteristic function of a sum

$$\begin{aligned} k_{Z_n}(t) &= k_{\frac{Y_1}{\sqrt{n}}}(t) \cdots k_{\frac{Y_n}{\sqrt{n}}}(t) \\ &= k_{Y_1}\left(\frac{t}{\sqrt{n}}\right) \cdots k_{Y_n}\left(\frac{t}{\sqrt{n}}\right) \\ &= \left(k_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 - \frac{t^2}{2n} + t^3(\dots)\right)^n \end{aligned}$$

goes to $e^{\frac{-t^2}{2}} = k_Z(t)$ as n goes to ∞. By theorem 5.30, $\{Z_n\}_{n\geq 1}$ converges in distribution to Z. We used the little fact $\lim_{n\to\infty}(1 + \frac{x}{n})^n = e^x$. □

That is, the sequence of CDFs of the $\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ converges pointwise to the CDF of $(X, 0, 1)$. This provides an approximation for the distribution of $\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ via the distribution of $(X, 0, 1)$ and vice versa.

Example 5.58. (confidence trials) We reconsider example 5.6. How many trials are needed to be more than 90% sure that your average $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$ is within 0.02 of the true value?

Again each trial X_i has mean $\mu(X_i) = p$, variance $\mathrm{var}(X_i) = p(1-p)$, and standard deviation $\sigma(X_i) = \sqrt{p(1-p)}$ while our average $\overline{X}_n$ has $\mu(\overline{X}_n) = p$ and $\sigma(\overline{X}_n) = \frac{\sigma}{\sqrt{n}}$. We want

$$P(|\overline{X}_n - \mu(\overline{X}_n)| \le 0.02) > 0.9$$
$$P\left(\left|\frac{\overline{X}_n - \mu(\overline{X}_n)}{\sigma(\overline{X}_n)}\right| \le \frac{0.02}{\sigma(\overline{X}_n)}\right) > 0.9$$

Central Limit Theorem says that $P\left(\left|\frac{\overline{X}_n-\mu(\overline{X}_n)}{\sigma(\overline{X}_n)}\right| \le \frac{0.02}{\sigma(\overline{X}_n)}\right) \to P\left(|Z| \le \frac{0.02}{\sigma(\overline{X}_n)}\right)$ as $n \to \infty$. So we want

$$P\left(|Z| \le \frac{0.02}{\sigma(\overline{X}_n)}\right) > 0.9$$
$$P\left(|Z| \le \frac{0.02\sqrt{n}}{\sigma}\right) > 0.9$$
$$P\left(Z \le \frac{0.02\sqrt{n}}{\sigma}\right) > 0.95$$

By standard normal table we need $\frac{0.02\sqrt{n}}{\sigma} \ge 1.65$.

a. If they say $p = 1/3$ for instance then $\mathrm{var}(X_i) = 2/9$ and $\sigma(X_i) = \sqrt{2}/3$ so $n \ge 1513$ will do.
b. If p is not given, then again $\mathrm{var}(X_i)$ is at most 1/4 and $\sigma(X_i)$ is at most 1/2 and $n \ge 1702$ will do.

Example 5.59. (confidence interval) We can apply Central Limit Theorem to a more general situation than in the previous example. Instead of taking n Bernoulli random variables, we collect n random samples, say $X_1 = 0.6, X_2 = 0.74, \ldots, X_n = 0.52$ with the same mean μ and standard deviation σ. Now instead of increasing n, we fix it and build an interval $[\mu-c, \mu+c]$ that contains $\overline{X}_n$ 98% of the times (meaning if you repeat sampling for 100 days, this average falls within $[\mu-c, \mu+c]$ 98 days, without 2 days.) The number c is called the accuracy level, the number 98% is called the confidence level. So $\overline{X}_n$ has same mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. We want

$$P(\mu - c \le \overline{X}_n \le \mu + c) = 0.98$$
$$P\left(\frac{\mu-c-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\mu+c-\mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0.98$$

By Central Limit Theorem, $P\left(\frac{\mu-c-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\overline{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} \le \frac{\mu+c-\mu}{\frac{\sigma}{\sqrt{n}}}\right) \to P\left(\frac{-c\sqrt{n}}{\sigma} \le Z \le \frac{c\sqrt{n}}{\sigma}\right)$ as $n \to \infty$ where Z is the standard normal variable. So we want,

$$P\left(\frac{-c\sqrt{n}}{\sigma} \le Z \le \frac{c\sqrt{n}}{\sigma}\right) = 0.98$$

$$P\left(Z \le \frac{c\sqrt{n}}{\sigma}\right) = 0.99$$

From standard normal table we need $\frac{c\sqrt{n}}{\sigma} = 2.33$ or $c = \frac{2.33\sigma}{\sqrt{n}}$. Usually n is given in the problem.

a. If σ is given as well, c is found!
b. In sampling, you perhaps don't know σ. So you must compute its sample error S_n from the samples X_i and use it as a substitute. Do you know how?

Example 5.60. (confidence level) Continuing with the previous two examples, we can fix n and c and compute how certain we are that our sampling $\overline{X}_n$ will fall within $[\mu-c, \mu+c]$.

Example 5.61. (histogram correction) Toss a fair coin 100 times and let S_{100} denote the number of tails. Use Central Limit Theorem to estimate $P(S_{100} = 50)$ and correct it with histogram.

a. So S_{100} has mean $\mu(S_{100}) = 50$ and $\sigma(S_{100}) = \sigma\sqrt{n} = \frac{1}{2}\sqrt{100} = 5$ to adjust.

$$\begin{aligned} P(S_{100} = 50) &= P(\overline{X}_n = 0.5) \\ &= P\left(\frac{\overline{X}_n - 0.5}{\frac{\sigma}{\sqrt{n}}} = \frac{0.5 - 0.5}{\frac{\sigma}{\sqrt{n}}}\right) \\ &\approx P(Z = 0) \\ &= 0 \end{aligned}$$

This is certainly a bad estimate of

$$\begin{aligned} P(X_{100} = 50) &= C(100, 50)\left(\frac{1}{2}\right)^{50}\left(\frac{1}{2}\right)^{50} \\ &= 0.0795892 \end{aligned}$$

b. With correction, $P(S_n = 50) \approx P(49.5 \le S_n \le 50.5)$. Central Limit Theorem says that as $n \to \infty$,

$$\begin{aligned} P(49.5 \le S_n \le 50.5) &\to P\left(\frac{49.5 - 50}{5}\right) \\ &= P(0.495 \le \overline{X}_n \le 0.505) \\ &\approx P\left(\frac{0.5 - 0.495}{\frac{\sigma}{\sqrt{n}}} \le \overline{X}_n \le \frac{0.505 - 0.5}{\frac{\sigma}{\sqrt{n}}}\right) \\ &\approx P(-0.1 \le Z \le 0.1) \end{aligned}$$

where Z is the standard normal variable. The standard normal table gives the final answer 0.079600, which is much closer to 0.0795892.

5.10. **Exercises.** pages 284: 1, 4, 5, 9, 10.

6. Random Processes

6.1. **Basic Concepts.** Let $\Omega = \{w_\alpha, \alpha \in A\}$ be a portfolio of stocks and

$$\begin{aligned}(\Omega, \mathcal{F}, P) &\xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))\\ w_\alpha &\mapsto X(w_\alpha)\end{aligned}$$

is the valuation. For one thing, the σ-algebra $\mathcal{F}$ of events is not available all at once, nor does it remain constant after initial appearance. For another thing, the valuation X changes as information changes. This calls for temporal notions of $\mathcal{F}$ and of X.

Definition 6.1. A random process X with index time T from a probability space $(\Omega, \mathcal{F}, P)$ to a state space $(S, \mathcal{S})$ is a map

$$\begin{aligned}(\Omega, \mathcal{F}, P) \times T &\xrightarrow{X} (S, \mathcal{S})\\ (w,t) &\mapsto X(\omega, t)\end{aligned}$$

such that for each finite collection $\{t_1, \dots, t_n\}$ the map

$$\begin{aligned}(\Omega, \mathcal{F}, P) &\xrightarrow{(X_{t_1},\dots,X_{t_n})} (S^m, \mathcal{S}^{\otimes m})\\ w &\mapsto (X_{t_1}(w), \dots, X_{t_n}(\omega)) = (X(\omega, t_1), \dots, X(\omega, t_n))\end{aligned}$$

is a multivariate random variable.

When $n = 1$, it follows that each X_t is a random variable. When Ω or S is discrete, the X_t are discrete random variables. We only consider $T \subset \mathbb{R}$, and most often T equals $\mathbb{N}^+, \mathbb{N}, [0, \infty)$ or $\mathbb{R}$. When T equals $\mathbb{N}^+$ or $\mathbb{N}$, the process X is called a *time chain.* The state space $(S, \mathcal{S})$ often equals $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Example 6.2. If $(\Omega, \mathcal{F}, P) \xrightarrow{Y} (S, \mathcal{S})$ is a random variable then for any T we have the random process

$$\begin{aligned}(\Omega, \mathcal{F}, P) \times T &\xrightarrow{X} (S, \mathcal{S})\\ (w,t) &\mapsto X(w,t) = Y(w)\end{aligned}$$

This process is essentially the random variable Y, it does not evolve over time.

Example 6.3. From the pricing $(\Omega, \mathcal{F}, P) \xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))$ in our opening discussion, we get the random process

$$\begin{aligned}(\Omega, \mathcal{F}, P) \times [0,\infty) &\xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))\\ (w,t) &\mapsto X(w,t) = X_t(w)\end{aligned}$$

which tells the price of each stock w at time t. If the price of each stock increases \$$t$ then we have $X(w,t) = X_t(w) = X_0(w) + t$ and every X_t has pdf $f_{X_t}(x)$ that is $f_X(x)$ shifted t to the right.

Example 6.4. From example 5.40, we may have a time chain

$$(\Omega, \mathcal{F}, P) \times \mathbb{N} \xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))$$
$$(w,n) \mapsto X(w,n) = X_n(w)$$

where

$$f_{X_0}(s,t) = \frac{1}{\pi}$$

and

$$f_{X_n}(s,t) = \frac{1}{2\pi\sigma_1(n)\sigma_2(n)} e^{-\left(\frac{s^2}{2\sigma_1^2(n)} + \frac{t^2}{2\sigma_2^2(n)}\right)} \text{ for } n \geq 1$$

Example 6.5. If $Y_1, Y_2, \dots$ are a sequence of random variables then we have the time chain

$$(\Omega, \mathcal{F}, P) \times \mathbb{N}^+ \xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))$$
$$(w,n) \mapsto X(w,n) = X_n(w) = Y_1(w) + \cdots + Y_n(w)$$

That is, each $X_n = Y_1 + \cdots + Y_n$ as random variables.

As there is a lot in the definition of a random process, there is a lot of details about it.

6.1.1. *Sample Path.* It traces the state of a single outcome w over time, or in example 6.3 the price of a single stock over time.

Definition 6.6. For a fixed $w \in \Omega$, the function $T \xrightarrow{X_-(w)} S, t \mapsto X_t(w)$ is called the sample path (or trajectory) of X associated with w. The complete set of all such trajectories is called the ensemble of X.

Example 6.7. Graph the sample path for X in example 6.5 where $Y_1, Y_2, \dots$ are independent identically distributed Bernoulli random variables in $\mathbb{N}^+ \times \mathbb{Z}$.

If both domain T and codomain S for sample paths have some notions of convergence, we can think of limits and continuity.

Definition 6.8. A sample path $T \xrightarrow{X_-(w)} S, t \mapsto X_t(w)$ of X is said to have left limits if $\lim\limits_{s\to t} X_s(w)$ exists for all $s < t \in T$. It is said to have right limits if $\lim\limits_{s\to t} X_s(w)$ exists for all $t < s \in T$.

Definition 6.9. A sample path $T \xrightarrow{X_-(w)} S, t \mapsto X_t(w)$ of X is said to be left continuous if $\lim\limits_{s\to t} X_s(w) = X_t(w)$ for all $s < t \in T$. It is said to be right continuous if $\lim\limits_{s\to t} X_s(w) = X_t(w)$ for all $t < s \in T$. It is said to be continuous if $\lim\limits_{s\to t} X_s(w) = X_t(w)$ for all $s, t \in T$.

By definitions, having left limits or right limits is the first step toward being left continuous or right continuous. These limit definitions are the same ones that we have for functions from $\mathbb{R}$ to $\mathbb{R}$ in analysis. Via continuity for sample paths we can define continuity for X.

Definition 6.10. A random process X is said to be right continuous (left continuous, continuous) if all sample paths of X are right continuous (left continuous, continuous).

It is said to be almost sure right continuous (almost sure left continuous, almost sure continuous) if almost all sample paths are right continuous (left continuous, continuous).

If all sample paths of X are right continuous with left limits then we say X is càdlàg. If almost all sample paths of X are right continuous with left limits then we say X is almost càdlàg. These processes form an important class of random processes in stochastic calculus.

Exercise 6.11. Let $\Omega = \{v, w\}$ with σ-algebra $\mathcal{F} = \mathbb{P}(\Omega)$ and probability measure $P(v) = P(w) = \frac{1}{2}$. Show that the following process $(\Omega, \mathcal{F}, P) \times \mathbb{R} \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is càdlàg.

$$X(v,t) = \begin{cases} 1 \text{ if } t < 0 \\ 0 \text{ if } t = 0 \\ \frac{sin^2(t)}{t} \text{ if } t > 0 \end{cases} \qquad X(w,t) = \begin{cases} 0 \text{ if } t < 0 \\ 1 \text{ if } t = 0 \\ e^{-t} \text{ if } t > 0 \end{cases}$$

Exercise 6.12. Give an example of a left continuous process X that is not continuous.

6.1.2. *Event Path.* It traces the probability of a single event A over time.

Definition 6.13. For a fixed $A \in \mathcal{S}$, the function $T \xrightarrow{P(X_-^{-1}(A))}, t \mapsto P(X_t^{-1}(A)) = P(\{w, X_t(w) \in A\})$ is called the event path of X associated with A.

In the discrete case

$$P(X_t^{-1}(A)) = \sum_{s \in A} P(\{w, X_t(w) = s\}) = \sum_{s \in A} p_{X_t}(s)$$

changes over time as each probability mass $p_{X_t}(s)$ changes over time.

In the continuous case, the area

$$P(X_t^{-1}(A)) = \int_A f_{X_t}(s) ds$$

under the curve $f_{X_t}(s)$ changes over time while

$$P(X_t^{-1}(s)) = P(\{w, X_t(w) = s\}) = 0$$

Example 6.14. The probability $P(X_t^{-1}([100, 200])) = P(100 \leq X_t \leq 200)$ that an Apple share prices between \$100 and \$200 changes over time.

Example 6.15. Graph the event path for X in example 6.4 in $\mathbb{N} \times [0,1]$.

6.1.3. *Measurablity.* If we equip $(\Omega, \mathcal{F}, P) \times [0, \infty)$ with the σ-algebra $\mathcal{F} \otimes \mathcal{B}([0, \infty))$ then we can ask more of each random process X.

Definition 6.16. A random process $(\Omega, \mathcal{F}, P) \times [0, \infty) \xrightarrow{X} (S, \mathcal{S})$ is called measurable if $X^{-1}(A) \in \mathcal{F} \otimes \mathcal{B}([0, \infty))$ for all $A \in \mathcal{S}$.

This just means $(\Omega \times [0, \infty), \mathcal{F} \otimes \mathcal{B}([0, \infty))) \xrightarrow{X} (S, \mathcal{S})$ is a measurable map between measurable spaces. When $S = \mathbb{R}$, it is a consequence of Fubini's theorem that each trajectory $([0, \infty), \mathcal{B}([0, \infty))) \xrightarrow{X_-(w)} (S, \mathcal{S})$ is also measurable. If each $E(X_t)$ is finite then each map $([0, \infty), \mathcal{B}([0, \infty))) \xrightarrow{E(X_-)} (S, \mathcal{S})$ is also measurable. And if $I \subset [0, \infty)$ is an interval such that $\int_I E(X_t) dt < \infty$ then $\int_I |X_t| dt < \infty$ almost surely and $\int_I E(X_t) dt = E(\int_I X_t dt)$.

6.2. **Equalities for Random Processes.** As with random variables, we consider when two random processes $(\Omega, \mathcal{F}, P) \times T \overset{X}{\underset{Y}{\rightrightarrows}} (S, \mathcal{S})$ are equal. For this, we can draw from our experience with equalities for random variables. With disregard to P, we would say they are the same if they are the same pointwise

$$X(w,t) = X_t(w) = Y(w,t) = Y_t(w)$$

for all $w \in \Omega, t \in T$. With regard to P and T, we could say they are the same in weaker senses.

Definition 6.17. X and Y are said to be indistinguishable if $P(X_t = Y_t \text{ for all } t \in T) = 1$.

Definition 6.18. X is said to be a modification of Y if $P(X_t = Y_t) = 1$ for each $t \in T$.

Surely X is a modification of Y iff Y is a modification of X. Indistinguishability clearly implies modification. If we write $\{X_t = Y_t \text{ for all } t \in T\} = \bigcap_{t \in T} \{X_t = Y_t\}$ then $P(\bigcap_{t\in T} \{X_t = Y_t\}) = 1$ implies $P(X_t = Y_t) = 1$ for each t. Or if we write $\{X_t \neq Y_t \text{ for some } t \in T\} = \bigcup_{t\in T} \{X_t \neq Y_t\}$ then $P(\bigcup_{t\in T}\{X_t \neq Y_t\}) = 0$ implies $P(X_t \neq Y_t) = 0$ for each $t \in T$. This means at each time t we have $X_t = Y_t$ almost surely with respect to P, or equivalently if almost all sample paths $X_-(\omega) = Y_-(\omega)$ pointwise. The converse is not true by the following example.

Example 6.19. For any positive continuous random variable $(\Omega, \mathcal{F}, P) \xrightarrow{T} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, define

$$\begin{aligned} (\Omega, \mathcal{F}, P) \times \mathbb{R} &\xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ (w,t) &\mapsto 0 \end{aligned}$$

and

$$\begin{aligned} (\Omega, \mathcal{F}, P) \times \mathbb{R} &\xrightarrow{Y} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ (w,t) &\mapsto \begin{cases} 0 \text{ if } T(w) \neq t \\ 1 \text{ if } T(w) = t \end{cases} \end{aligned}$$

Then X is a modification of Y because for each $t \in \mathbb{R}$ we have

$$\begin{aligned} P(X_t = Y_t) &= P(Y_t = 0) \\ &= P(T \neq t) \\ &= 1 - P(T = t) \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

On the other hand X and Y are distinguishable because

$$\begin{aligned} P(X_t = Y_t \text{ for all } t \in \mathbb{R}) &= P(Y_t = 0 \text{ for all } t \in \mathbb{R}) \\ &= P(\{w, T(w) \neq t \text{ for all } t \in \mathbb{R}\}) \\ &= P(\varnothing) \\ &= 0 \end{aligned}$$

Exercise 6.20. Show that the converse is true if almost all the sample paths of X and of Y are right continuous.

Next is the last mode of equality that we consider for random processes. It places emphasis on the distribution of each multivariate random variable $(X_{t_1}, \dots, X_{t_n})$.

Definition 6.21. X and Y are said to be equal in finite-dimensional distribution if for each finite collection of times $t_1 < \cdots < t_n$, the joint random variables $(X_{t_1}, \dots, X_{t_n})$ and $(Y_{t_1}, \dots, Y_{t_n})$ are equal in distribution.

By definition of equality in distribution for random variables, this means $F_{X_{t_1},\dots,X_{t_n}} = F_{Y_{t_1},\dots,Y_{t_n}}$ everywhere in $\mathbb{R}^m$. It follows that $P((X_1, \dots, X_m) \in A) = P((Y_1, \dots, Y_m) \in A)$ for all $A \in \mathcal{B}(\mathbb{R}^m)$. The question is how equality in finite-dimensional distribution relates to modification and indistinguishability.

Exercise 6.22. Show that if X is a modification of Y then X and Y are equal in finite-dimensional distribution.

6.3. **Family of Random Variables.** One can think of a random process X as a family of random variables evolving over time T.

$$T \xrightarrow{X} RV((\Omega, \mathcal{F}, P), (S, \mathcal{S}))$$

$$t \mapsto X_t \text{ where } (\Omega, \mathcal{F}, P) \xrightarrow{X_t} (S, \mathcal{S})$$

The codomain $RV((\Omega, \mathcal{F}, P), (S, \mathcal{S}))$ of X now has convergence in distribution, convergence in probability, convergence in mean, etc. When the domain of X also has convergence, such as when $T = \mathbb{R}$ or $T = [0, \infty)$ then we can think of continuity for X as a map instead of continuity for X via its sample paths in 6.1.1. Continuity is a nice property to have for X since it reveals the behavior of the family of random variables $\{X_t\}_{t \in T}$ as a whole despite the fact they may be independent.

Definition 6.23. A random process $T \xrightarrow{X} RV((\Omega, \mathcal{F}, P), (S, \mathcal{S}))$ is said to be continuous in distribution if $X_s \xrightarrow{d} X_t$ whenever $s \to t$.

We can define sure continuity, almost sure continuity, continuity in probability and continuity in k^{th}-moment for X in a similar fashion. Again each continuity may imply another under some hypotheses in the same way that we saw with random variables.

Exercise 6.24. Define what it means for a random process $T \xrightarrow{X} RV((\Omega, \mathcal{F}, P), (S, \mathcal{S}))$ to be right continuous in probability.

We can go one step further and consider the composition

$$\begin{tikzcd}
T \arrow[r, "X"] \arrow[dr, "\overline{X}"'] & RV((\Omega, \mathcal{F}, P), (S, \mathcal{S})) \arrow[d, "/"] \\
 & \overline{RV}((\Omega, \mathcal{F}, P), (S, \mathcal{S}))
\end{tikzcd}$$

Any mode of convergence for $RV((\Omega,\mathcal{F},P),(S,\mathcal{S}))$ induces a corresponding mode of convergence for $\overline{RV}((\Omega,\mathcal{F},P),(S,\mathcal{S}))$. For example, we say a sequence $\{\overline{Y_t}\}_{t\in T}$ converges in second moment to $\overline{Y}$ in $\overline{RV}((\Omega,\mathcal{F},P),(S,\mathcal{S}))$ if there exist representations Y_t for $\overline{Y_t}$ and Y for $\overline{Y}$ such that $\{Y_t\}_{t\in T}$ converges in second moment to Y in $RV((\Omega,\mathcal{F},P),(S,\mathcal{S}))$. Moreover, $\overline{RV}((\Omega,\mathcal{F},P),(S,\mathcal{S}))$ is equipped with norm $\|\overline{Y}\| = \sqrt{\text{var}(Y)}$ and we say $\{\overline{Y_t}\}_{t\in T}$ converges in norm to $\overline{Y}$ if $\|Y_t - Y\|$ goes to 0 as t goes to ∞. Hence we can define continuity for $\overline{X}$ in the same manner we did for X above.

Exercise 6.25. Can you relate the following notions of continuity for a random process X and its $\overline{X}$?

a. Sure continuity in sample paths and sure continuity for X.
b. Almost sure continuity in sample paths and almost sure continuity for X.
c. Continuity in distribution and continuity in probability for X.
d. Continuity in second moment and continuity in norm for $\overline{X}$.

6.4. **Increments.** When $S = \mathbb{R}$ with the usual distance $d(a,b) = |a-b|$, we are also interested in $X_s - X_t$ for all $s,t \in T$. Such increments may possess telling properties about X. We use $\{X_t\}_{t\in T}$ for X to underline the changes within X over time.

Definition 6.26. A random process $\{X_t\}_{t\in T}$ is said to be homogeneous if the distribution of any increment $X_s - X_t$ depends only on the length $s - t$.

Example 6.27. The pricing process in example 6.3 is homogeneous, as $X_s - X_t = s - t$ is constant

Definition 6.28. A random process $\{X_t\}_{t\in T}$ is said to be stationary if the families $\{X_{t_1},\dots,X_{t_n}\}$ and $\{X_{t_1+h},\dots,X_{t_n+h}\}$ have the same joint distribution for all $t_1,\dots,t_n$ and h.

Exercise 6.29. Give an example of a stationary process and explain why it is stationary.

Definition 6.30. A random process $\{X_t\}_{t\in T}$ is said to have stationary increments if $X_{t+h} - X_t$ and X_h have the same distribution for all t,h.

Example 6.31. If $Y_1, Y_2, \dots$ are identically distributed random variables then the process $X = \{X_n\}_{n\in\mathbb{N}^+}$ where $X_n = Y_1 + \dots + Y_n$ has stationary increments, as $X_{n+k} - X_n = Y_{n+1} + \dots + Y_{n+k}$ has the same distribution as $X_k = Y_1 + \dots + Y_k$.

Example 6.32. If Y is a random variable then the process $X = \{X_t\}_{t\in T}$ where $X_t = tY$ also has stationary increments, as $X_{t+h} - X_t = (t+h)Y - tY = hY$ has the same distribution as $X_h = hY$.

Definition 6.33. A random process $\{X_t\}_{t\in T}$ is said to have independent increments if $X_{s_1} - X_{t_1}, \dots, X_{s_n} - X_{t_n}$ are mutually independent whenever the time intervals $[t_1,s_1],\dots,[t_n,s_n]$ are pairwise disjoint.

In other words, the differences over disjoint intervals $[t_1,s_1],\dots,[t_n,s_n]$ are independent. For example, a large increase over one interval does not mean a subsequent small increase over another disjoint interval.

Example 6.34. If $Y_1, Y_2, \dots$ are independent identically distributed random variables then the process $X = \{X_n\}_{n \in \mathbb{N}^+}$ where $X_n = Y_1 + \cdots + Y_n$ has independent increments. Surely $X_{n_1'} - X_{n_1} = Y_{n_1+1} + \cdots + Y_{n_1'}$ is independent of $X_{n_2'} - X_{n_2} = Y_{n_2+1} + \cdots + Y_{n_2'}$ whenever $[n_1, n_1']$ and $[n_2, n_2']$ are disjoint. The same does not hold if these intervals are not disjoint, as some $Y_k, k \in [n_1, n_1'] \bigcap [n_2, n_2']$ will appear in both sums.

Here is a nice result about processes with independent stationary increments.

Proposition 6.35. *If $T = [0, \infty)$ and X has independent stationary increments then there exist $a, b \in \mathbb{R}$ such that $E(X_t) = at + b$ is an affine function in t.*

Proof. Consider the function $f(t) = E(X_t - X_0) = E(X_t) - E(X_0)$. For $s, t \in T$ we have

$$\begin{aligned} f(s+t) &= E(X_{s+t} - X_0) \\ &= E(X_{s+t} - X_t + X_t - X_0) \\ &= E(X_{s+t} - X_t) + E(X_t - X_0) \\ &= E(X_s - X_0) + E(X_t - X_0) \\ &= f(s) + f(t) \end{aligned}$$

So $f(t) = at$ is linear for some $a \in \mathbb{R}$. Hence $E(X_t) = f(t) + E(X_0) = at + b$. □

6.5. **Filtration, Adaptibility, Predictability.** When we considered the event path

$$T \xrightarrow{P(X_-(A))} [0,1]$$
$$t \mapsto P(X_t^{-1}(A))$$

there was an assumption that all $X_t^{-1}(A), t \in T$ are in $\mathcal{F}$ ready to be measured. The σ-algebra $\mathcal{F}$ seems constant over time. What if it also changes, that is, what if events become available not all at once but along a timeline?

Recall from definition 3.5 that a collection $\{\mathcal{F}_t\}_{t \in T}$ of σ-algebras indexed by T is called a filtration of $\mathcal{F}$ if

$$\cdots \mathcal{F}_s \subset \mathcal{F}_t \subset \cdots \subset \mathcal{F} \text{ for all } s < t \in T$$

We will now use this aid to develop the theory of random processes.

Definition 6.36. A probability space $(\Omega, \mathcal{F}, P)$ is called a filtered probability space if $\mathcal{F}$ has a filtration $\{\mathcal{F}_t\}_{t \in T}$.

Sometimes we must write $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in T}, P)$ to denote this whole structure. One can think of $\{\mathcal{F}_t\}_{t \in T}$ as a track of information we knew, know and will know over time. Moreover, we can define two new filtrations

$$\{\mathcal{F}_{t^-}\}_{t \in T} \text{ where } \mathcal{F}_{t^-} = \sigma(\bigcup_{s<t} \mathcal{F}_s)$$

is the σ-algebra of events strictly before t and

$$\{\mathcal{F}_{t^+}\}_{t \in T} \text{ where } \mathcal{F}_{t^+} = \bigcap_{t<s} \mathcal{F}_s$$

is the σ-algebra of events immediately after t.

Exercise 6.37. Show that $\mathcal{F}_{t^+} = \bigcap\limits_{t<s} \mathcal{F}_s$ is indeed a σ-algebra and find an example to show $\bigcup\limits_{s<t} \mathcal{F}_s$ may not be a σ-algebra.

Example 6.38. Let $\Omega = \{1,2,3,4,5,6\}$ be the space of all possible outcomes when we toss a die. Next let $\mathcal{F}_1 = \{\varnothing, \Omega\}$, $\mathcal{F}_2 = \{\{1\},\{2,3,4,5,6\},\varnothing,\Omega\}$, and $\mathcal{F} = \mathbb{P}(\Omega)$ then they form a filtration $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}$. If P is the usual probability measure that maps each face to $1/6$ then P still measures events in these σ-algebras consistently.

$$\begin{tikzcd}
\{\varnothing,\Omega\} \arrow[r] \arrow[drr, "P|_{\mathcal{F}_1}"] & \{\{1\},\{2,3,4,5,6\},\varnothing,\Omega\} \arrow[r] \arrow[dr, "P|_{\mathcal{F}_2}"] & \mathbb{P}(\Omega) \arrow[d, "P"'] & \varnothing \arrow[d, "P|_{\mathcal{F}_2}"'] & \{1\} \arrow[d, "P|_{\mathcal{F}_2}"'] & \{2,3,4,5,6\} \arrow[d, "P|_{\mathcal{F}_2}"'] & \mathbb{P}(\Omega) \arrow[d, "P|_{\mathcal{F}_2}"'] \\
 & & {[0,1]} & 0 & 1/6 & 5/6 & 1
\end{tikzcd}$$

Example 6.39. (natural filtration) Given any random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$, we can define a filtration $\{\mathcal{F}_t^X\}_{t\in T}$ of $\mathcal{F}$ as $\mathcal{F}_t^X = \sigma(X_s, s \le t)$ the σ-algebra generated by all events about all $X_s, s \le t$. To say A is in $\mathcal{F}_t^X$ means that by time t, an observer of X knows whether A has occurred or not. By definition, each $\mathcal{F}_t^X$ is the smallest σ-algebra with respect to which $X_s, s \le t$ is measurable. This filtration $\{\mathcal{F}_t^X\}_{t\in T}$ is called the natural filtration of X.

Example 6.40. Given any random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R})))$, we can define another filtration $\{\mathcal{F}_t\}_{t\in T}$ for $\mathcal{F}$ as $\mathcal{F}_t = \langle X_s^{-1}((-\infty, s]), s \le t\rangle$ the σ-algebra generated by all $X_s^{-1}((-\infty, s]), s \le t$.

We introduce some common classes of filtrations.

Definition 6.41. A filtration $\{\mathcal{F}_t\}_{t\in T}$ of $\mathcal{F}$ where $T = \mathbb{N}$ or $T = [0,\infty)$ is said to be complete if $A \in \mathcal{F}_0$ whenever $A \in \mathcal{F}$ and $P(A) = 0$.

This means in a complete filtration, all the negligible events and hence all almost sure events are available at the beginning. Since $\mathcal{F}_0 \subset \mathcal{F}_t, t > 0$, it follows that such events are always available.

Exercise 6.42. Let $\Omega = \mathbb{Z}/12\mathbb{Z} = \{0,\dots,11\}$ with $\mathcal{F} = \mathbb{P}(\Omega)$ and $P(0) = P(1) = 0, P(2) = \cdots = P(11) = \frac{1}{10}$. Give a complete filtration $\mathcal{F}_0 \subsetneq \mathcal{F}_1 \subsetneq \mathcal{F}_2 \subsetneq \mathcal{F}_3$ of $\mathcal{F}$.

Definition 6.43. A filtration $\{\mathcal{F}_t\}_{t\in T}$ is said to be right continuous if $\mathcal{F}_{t^+} = \mathcal{F}_t$ for all $t \in T$. It is said to be left continuous if $\mathcal{F}_{t^-} = \mathcal{F}_t$ for all $t \in T$.

Right continuity means there is no event that is detectable at all time $s > t$ yet it is undetectable at time t, or equivalently there is no sudden increase in information after t. Left continuity means there is no event that is undetectable at all time $s < t$ yet it is detectable at time t, or equivalently there is no sudden increase in information at t. A filtered probability space is said to satisfy the usual conditions if it is complete and right continuous. Such spaces are prevalent in stochastic calculus.

Exercise 6.44. Show that $\{\mathcal{F}_{t^+}\}_{t\in T}$ is the smallest right continuous filtration containing $\{\mathcal{F}_t\}_{t\in T}$ and they are equal if $\{\mathcal{F}_t\}_{t\in T}$ is already right continuous. Similarly, $\{\mathcal{F}_{t^-}\}_{t\in T}$ is the smallest left continuous filtration contained in $\{\mathcal{F}_t\}_{t\in T}$ and they are equal if $\{\mathcal{F}_t\}_{t\in T}$ is already left continuous.

Next we consider how the family $\{X_t\}_{t\in T}$ connects with the filtration $\{\mathcal{F}_t\}_{t\in T}$.

Definition 6.45. (Adapted process) A random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ is said to be adapted to a filtration $\{\mathcal{F}_t\}_{t\in T}$ of $\mathcal{F}$ if each $(\Omega, \mathcal{F}_t, P) \xrightarrow{X_t} (S, \mathcal{S})$ is measurable with respect to $(\mathcal{F}_t, \mathcal{S})$, i.e. if it is a random variable with respect to $(\mathcal{F}_t, \mathcal{S})$.

Example 6.46. Suppose we toss a fair coin three times. Let $\Omega = \{\text{all sequences of } H \text{ and } T \text{ of length 3}\}$ with $\mathcal{F} = \mathbb{P}(\Omega)$ and $P(\omega) = \frac{1}{6}$ for all $\omega \in \Omega$. We define a filtration of $\mathcal{F}$ as follows

$$\begin{aligned}
\mathcal{F}_0 &= \{\varnothing, \Omega\} \\
\mathcal{F}_1 &= \langle\{\text{all sequences starting with } H\}\rangle \\
\mathcal{F}_2 &= \langle\{\text{all sequences starting with } HH\}, \{\text{all sequences starting with } HT\}, \\
&\qquad \{\text{all sequences starting with } TH\}, \{\text{all sequences starting with } TT\}\rangle \\
\mathcal{F}_3 &= \mathbb{P}(\Omega)
\end{aligned}$$

Here $T = \{0, 1, 2, 3\}$ finite time. Now we define X_n to count the numbers of heads after n tosses. Then $\{X_n\}_{n\in T}$ is adapted to $\{\mathcal{F}_n\}_{n\in T}$.

Example 6.47. Every random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ is adapted to its natural filtration $\{\mathcal{F}_t^X\}_{t\in T}$ in example 6.39. People use $\{\mathcal{F}_t^X\}_{t\in T}$ very often to gain adaptability.

An adapted process in a sense does not use future information. At time t, events in $\mathcal{F}_t$ are known to us while some events $X_s^{-1}(A), t < s, A \in \mathcal{S}$ may be unavailable until time s. Given an event A in $\mathcal{S}$, $X_t^{-1}(A)$ is in $\mathcal{F}_t$ because X_t is measurable with respect to $(\mathcal{F}_t, \mathcal{S})$. However $X_t^{-1}(A)$ may not be in $\mathcal{F}_s, s < t$ because $\mathcal{F}_s$ may be smaller than $\mathcal{F}_t$. When X_t is measurable with respect to $(\mathcal{F}_s, \mathcal{S})$ for all $s < t$, we call X a predictable process.

Definition 6.48. (Predictable process)

i. A random process $(\Omega, \mathcal{F}, P) \times \mathbb{N} \xrightarrow{X} (S, \mathcal{S})$ over discrete time $\mathbb{N}$ is said to be predictable with respect to a filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ of $\mathcal{F}$ if each $(\Omega, \mathcal{F}_{n-1}, P) \xrightarrow{X_n} (S, \mathcal{S})$ is a random variable.
ii. A random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ over continuous time T is said to be predictable with respect to a filtration $\{\mathcal{F}_t\}_{t\in T}$ of $\mathcal{F}$ if each $(\Omega, \mathcal{F}_{t^-}, P) \xrightarrow{X_t} (S, \mathcal{S})$ is a random variable.

This means in a predictable process one can predict the values X_n using the information $\mathcal{F}_{n-1}$ available at time $n-1$ when T is discrete, or the values X_t using the information $\mathcal{F}_{t^-}$ available before time t when T is continuous.

Exercise 6.49. Can you modify example 6.46 to construct a predictable process over finite time?

Exercise 6.50. For $T = [0, \infty)$, show that every left continuous process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ adapted to a filtration $\{\mathcal{F}_t\}_{t\in T}$ of $\mathcal{F}$ is predictable with respect to that filtration.

6.6. **Random Times.** Looking at the random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))$ in example 6.3, we may also ask when a stock reaches a certain price for the first time, or second, or third, etc. More generally, we want a function to record the moments an outcome takes a certain state.

Definition 6.51. A random time is a measurable function $(\Omega, \mathcal{F}, P) \xrightarrow{\mathcal{T}} (T, \mathcal{B}(T))$.

So a random time is just a random variable that gives us values in time. Infinity is included in T to account for the scenario that $\mathcal{T}$ never achieves a certain state.

Example 6.52. We look at what happens during and after an earthquake. Let $\Omega = (0,10)$ with σ-algebra $\mathcal{F} = \mathcal{B}((0,10))$ and probability measure P. Let $(\Omega, \mathcal{F}, P) \xrightarrow{\mathcal{T}} ([0,\infty], \mathcal{B}([0,\infty]))$ record when an earthquake of magnitude between 0 and 10 occurs for the first time then it is a random time. Let $(\Omega, \mathcal{F}, P) \times [0,\infty) \xrightarrow{X} ([0,\infty), \mathcal{B}([0,\infty)))$ record the damages in dollars then X is a random process. Surely $\mathcal{T}$ and X are related.

Exercise 6.53. Suppose $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ is measurable random process and $(\Omega, \mathcal{F}, P) \xrightarrow{\mathcal{T}} (T, \mathcal{B}(T))$ is a random time. Set $\Omega' = \{\mathcal{T} < \infty\} \subset \Omega$. Show that

$$\begin{aligned} (\Omega', \mathcal{F}, P) &\xrightarrow{X_{\mathcal{T}}} (S, \mathcal{S}) \\ w &\mapsto X_{\mathcal{T}}(w) = X(w, \mathcal{T}(w)) \end{aligned}$$

is a random variable.

$$\begin{tikzcd}
(\Omega', \mathcal{F}, P) \arrow[r, "{(i,\mathcal{T})}"] \arrow[dr, "X_{\mathcal{T}}"] & (\Omega, \mathcal{F}, P) \times [0,\infty) \arrow[d, "X"] \\
& (S, \mathcal{S})
\end{tikzcd}$$

$$\begin{tikzcd}
w \arrow[r] \arrow[dr] & (w, \mathcal{T}(w)) \arrow[d] \\
& X_{\mathcal{T}}(w) = X_{\mathcal{T}(w)}(w)
\end{tikzcd}$$

It is reasonable to expect that the event $\{\mathcal{T} \leq t\}$ of outcomes occurring before or at time t to be part of the information available at time t.

Definition 6.54. A random time $(\Omega, \mathcal{F}, P) \xrightarrow{\mathcal{T}} (T, \mathcal{B}(T))$ is called a stopping time with respected the filtration $\{\mathcal{F}_t\}_{t \in T}$ of $\mathcal{F}$ if $\{\mathcal{T} \leq t\} \in \mathcal{F}_t$ for all $t \in T$. It is called an optional time with respect to the filtration if $\{\mathcal{T} < t\} \in \mathcal{F}_t$ for all $t \in T$.

If $T = \mathbb{N}$ or $T = \mathbb{N}^+$ countable then one can show that $\mathcal{T}$ is a stopping time iff $\{\mathcal{T} = t\} \in \mathcal{F}_t$ for all $t \in T$.

Example 6.55. Any constant map $(\Omega, \mathcal{F}, P) \xrightarrow{c} ([0,\infty], \mathcal{B}([0,\infty]))$ is a stopping time with respected any filtration $\{\mathcal{F}_t\}_{t \in T}$ of $\mathcal{F}$. For any $t \in T$ we have $\{c \leq t\}$ is either $\varnothing$ or Ω, hence it belongs to $\mathcal{F}_t$.

Proposition 6.56. *Every stopping time $\mathcal{T}$ with respect to a filtration $\{\mathcal{F}_t\}_{t \in T}$ is optional, and the two concepts coincide if the filtration is right continuous.*

Proof. If $\mathcal{T}$ is a stopping time then $\{\mathcal{T} \le t - \frac{1}{n}\} \in \mathcal{F}_{t-\frac{1}{n}} \subset \mathcal{F}_t$ for all $n \ge 1$, so $\{\mathcal{T} < t\} = \bigcup_{n\ge1}\{\mathcal{T} \le t - \frac{1}{n}\} \in \mathcal{F}_t$ and $\mathcal{T}$ is optional. If $\mathcal{T}$ is optional and $\mathcal{F}$ is right continuous then $\{\mathcal{T} \le t\} = \bigcap_{\epsilon>0}\{\mathcal{T} < t+\epsilon\} \in \mathcal{F}_{t+\epsilon}$ for every $t \in T$ and every $\epsilon > 0$, so $\{\mathcal{T} \le t\} \in \bigcap_{\epsilon>0}\mathcal{F}_{t+\epsilon} = \mathcal{F}_{t^+} = \mathcal{F}_t$. □

Exercise 6.57. If $\mathcal{T}$ is optional with respect to a filtration $\{\mathcal{F}_t\}_{t\in T}$ and $0 < \theta \in \mathbb{R}$ then show that $\mathcal{T} + \theta$ is a stopping time.

The reason $\mathcal{T}$ has the name "stopping time" is because $\mathcal{T}$ often marks the moment where some process X stops changing. In more practical situations, $\mathcal{T}$ marks the strategic moment when the player stops playing or the investor stops trading. More precisely, let $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ be a random process, let $(\Omega, \mathcal{F}, P) \xrightarrow{\mathcal{T}} (T, \mathcal{B}(T))$ be a random time and define

$$\begin{aligned}(\Omega, \mathcal{F}, P) \times [0, \infty) &\xrightarrow{X^{\mathcal{T}}} (S, \mathcal{S})\\ (w, t) &\mapsto X(w, \min\{t, \mathcal{T}(w)\})\end{aligned}$$

Then $X^{\mathcal{T}}$ is the stopped process of X by $\mathcal{T}$. For each $w \in \Omega$ its state remains constant after time $\mathcal{T}(w)$. Compare the trajectories of X and of $X^{\mathcal{T}}$. This $X^{\mathcal{T}}$ is very different from $X_{\mathcal{T}}$ in exercise 6.53.

Exercise 6.58. If X is predictable process and $\mathcal{T}$ is a stopping time with respected the filtration for X, show that the stopped process $X^{\mathcal{T}}$ is also predictable.

Now we can record when the first, or second or third time something happens. Given a random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ and an event $A \subset S$, consider the random times

$$\begin{aligned}(\Omega, \mathcal{F}, P) &\xrightarrow{T_1^A} (T, \mathcal{B}(T))\\ w &\mapsto \inf\{t, X_t(w) \in A\}\end{aligned}$$

and for $n \ge 2$,

$$\begin{aligned}(\Omega, \mathcal{F}, P) &\xrightarrow{T_n^A} (T, \mathcal{B}(T))\\ w &\mapsto \inf\{t > T_{n-1}^A(w), X_t(w) \in A\} - T_{n-1}^A(\omega)\end{aligned}$$

Then T_1^A records the first time X is in A while each T_n^A is the time between the $(n-1)^{\text{th}}$ and n^{th} visits to A. Each T_n^A is called the n^{th} hitting time and $T^A = \{T_n^A\}_{n\in\mathbb{N}^+}$ is called the hitting process associated to X and A. One can also define the last hitting time as

$$\begin{aligned}(\Omega, \mathcal{F}, P) &\xrightarrow{L^A} (T, \mathcal{B}(T))\\ w &\mapsto \sup\{t, X_t \in A)\end{aligned}$$

If the process X is adapted to filtration $\{\mathcal{F}_t\}_{t\in T}$ then it is true that each n^{th} hitting time T_n^A is a stopping time with respected the filtration $\{\mathcal{F}_t\}_{t\in T}$, while the last hitting time L^A in general is not.

6.7. **Common Random Processes.** We survey some common random processes.

6.7.1. *Bernoulli Process.* A random process $(\Omega, \mathcal{F}, P) \times \mathbb{N}^+ \xrightarrow{X} (S, \mathcal{S})$ is called a Bernoulli process if each $(\Omega, \mathcal{F}, P) \xrightarrow{X_n} (S, \mathcal{S})$ is a Bernoulli random variable. Note that the X_n need not be independent or identically distributed. We may even let them evolve by allowing the success rate p to change over time

$$\begin{aligned} \mathbb{N}^+ &\xrightarrow{p} [0,1] \\ n &\mapsto p(n) \end{aligned}$$

Conversely, if $\{X_n\}_{n \in \mathbb{N}^+}$ are Bernoulli random variables then we can string them together into a random process $X = \{X_n\}_{n \in \mathbb{N}^+}$. Moreover, let $Y_n = X_1 + \cdots + X_n$ then $Y = \{Y_n\}_{n \in \mathbb{N}^+}$ is called a binomial process.

6.7.2. *Normal Process.* Following along this line, we can replace the Bernoulli X_n with any type of random variables in 4.5 and get a random process of the same name, though some additional properties may be required and T may be continuous. For example, $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ is called a normal process if for each finite collection $t_1 < \cdots < t_n$ of times, the tuple $(X_{t_1}, \ldots, X_{t_n})$ is a multivariate normal random variable.

6.7.3. *Poisson Process.* As we have seen with random variables, the continuous analogue of a binomial random process would be a Poisson random process. A random process $(\Omega, \mathcal{F}, P) \times \mathbb{R} \xrightarrow{X} (S, \mathcal{S})$ is called Poisson process if it satisfies four conditions

1. $X_0 = 0$
2. X has independent increments
3. X has stationary increments
4. $P(X_{s+t} - X_s = k) = P(\{w, X_{s+t}(w) - X_s(w) = k\}) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$ for $k = 0, 1, 2, \ldots$

The last property means that the number of occurrences in any time interval of length t is Poisson distributed with mean λt. This process is homogeneous iff $P(X_{s+t} - X_s = k)$ depends only on t and k iff λ is constant over time. Other properties include the waiting time until the next occurrence has exponential distribution and the occurrences are uniformly distributed over any time interval. Poisson processes best model the number of customers in line, or the number of webpage requests to a server, or the number of emitted particles via radioactive decay and so on along a timeline.

6.7.4. *Markov Process.* A random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ adapted to a filtration $\{\mathcal{F}_t\}_{t \in T}$ is called a Markov process if it satisfies the Markov property

$$P(X_t \in A \,|\, \mathcal{F}_s) = P(X_t \in A \,|\, X_s), \text{ for all } A \in \mathcal{S} \text{ and } s < t \in T$$

On the left side we mean $E(1_{X_t \in A} \,|\, \mathcal{F}_s)$ and on the right side we mean $E(1_{X_t \in A} \,|\, \sigma(X_s))$ where $\mathcal{F}_s \supseteq \sigma(X_s)$ by measurability of X_s. This Markov property essentially means a Markov process is memoryless, its future state depends as much on its present state (given by all events in $\sigma(X_s)$ about X_s) as it does on its full history (given by $\mathcal{F}_t'$). A random process X without a given filtration is tested for Markov property with respect to its natural filtration $\{\mathcal{F}_t^X\}_{t \in T}$.

Example 6.59. A Poisson process satisfies Markov property so it is a continuous-time Markov process.

When $T = \mathbb{N}$ and S are discrete, Markov property has the simple form

$$P(X_n = s_n \,|\, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) = P(X_n = s_n \,|\, X_{n-1} = s_{n-1})$$

and any X satisfying this property is called a Markov chain. Each $P(X_{n_1} = s_1 | X_{n_2} = s_2)$ is denoted by $p_{s_1,s_2}(n_1, n_2)$ and called transitional probability. Together, all $p_{s_1,s_2}(n_1, n_2), s_i \in S, n_i \in \mathbb{N}$ completely determine X. It is homogenous iff $p_{s_1,s_2}(n_1, n_2) = p_{s_1,s_2}(n_1 - n_2, 0)$ for all $s_1, s_2 \in S$ and $n_1 \geq n_2 \in \mathbb{N}$. In particular, $p_{s_1,s_2}(n+1, n) = p_{s_1,s_2}(1, 0)$ for all $n \in \mathbb{N}$, in which case we simply write it as p_{s_1,s_2}.

Example 6.60. Liangbiang never has two nice days in a row. If it has a nice day, it is just as likely to have snow or rain the next day. If it has snow or rain, it has an even chance of having the same the next day. If there is a change from snow or rain, only half of the time is this a change to a nice day. We model this weather pattern as a Markov chain. Let $T = \mathbb{N}$ and $S = \{n, r, s\}$. Then we can form what is called a transition matrix of all transitional probabilities

$$\left(\begin{array}{cccc} & r & n & s \\ r & 1/2 & 1/4 & 1/4 \\ n & 1/2 & 0 & 1/2 \\ s & 1/4 & 1/4 & 1/2 \end{array}\right)$$

What are $\Omega, \mathcal{F}$, and P?

Example 6.61. Every homogeneous process is Markov. For example, if T is discrete then we see that

$$\begin{aligned} P(X_n = s_n \,|\, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) &= P(X_n - X_{n-1} = s_n - s_{n-1} \,|\, X_{n-1} = s_{n-1}, \ldots, X_0 = s_0) \\ &= P(X_n - X_{n-1} = s_n - s_{n-1} \,|\, X_{n-1} = s_{n-1}) \\ &= P(X_n = s_n \,|\, X_{n-1} = s_{n-1}) \end{aligned}$$

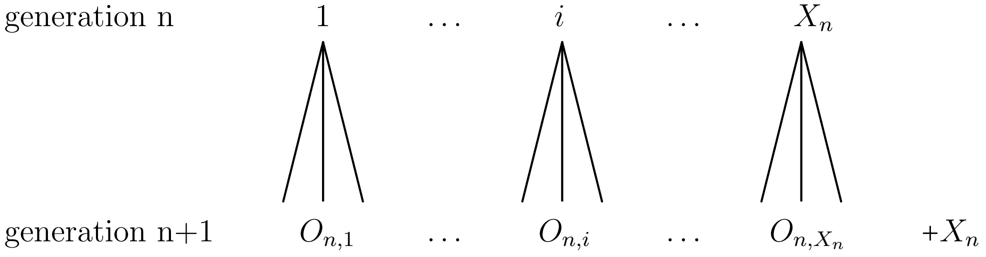
since the distribution of $X_n - X_{n-1}$ depends only on $n - (n-1)$. In particular, the random process in example 6.3 is Markov.

Exercise 6.62. Suppose you randomly draw from a collection of five \$1 bills, five \$2 bills, five \$5 bills, five \$10 bills in your pocket and $X_n, 1 \leq n \leq 20$ is the sum of money after n draws.

a. Define $\Omega, \mathcal{F}, P, N, X, S, \mathcal{S}$ so that $(\Omega, \mathcal{F}, P) \times N \xrightarrow{X = \{X_n\}_{n \in N}} (S, \mathcal{S})$ is a random process.
b. Calculate $P(X_6 = 12 \,|\, X_5 = 10, X_4 = 8, X_3 = 6, X_2, = 4, X_1 = 2)$.
c. Calculate $P(X_6 = 12 \,|\, X_5 = 10)$.
d. Conclude that X is not a Markov chain.

Example 6.63. (Branching Process) Another Markov chain is the branching process $\{X_n\}_{n \in \mathbb{N}}$ that models a population in which each individual i in generation n^{th} produces a random number of offspring $O_{n,i}$ with the same distribution for all n, i

$$\begin{aligned} X_1 &= 1 \\ X_2 &= 1 + O_{1,1} \\ &\vdots \\ X_{n+1} &= X_n + \sum_{i=1}^{X_n} O_{n,i} \end{aligned}$$



Example 6.64. (Simple Random Walk) Simple random walk takes many forms, one of which is the game where the gambler wins \$1 every time the dealer tosses head with probability p and loses \$1 every time the dealer tosses tail with probability $1-p$. If he begins with S_0 dollars then his fortune after n tosses would be

$$S_n = S_{n-1} + B_n = \cdots = S_0 + \sum_{k=1}^{n} B_k$$

where $\{B_k\}_{k\in\mathbb{N}^+}$ is a sequence of Bernoulli variables taking value 1 with probability p and -1 with probability $1-p$. One has some choices to describe simple random walk in the language of random processes. Here are two candidates.

i. Let $\Omega = \{H, T\}, \mathcal{F} = \mathbb{P}(\Omega), P(H) = p, P(T) = 1-p$. For $n \geq 1$, let $(\Omega, \mathcal{F}, P) \xrightarrow{B_n} (\mathbb{Z}, \mathbb{P}(\mathbb{Z}))$ be independent identically distributed Bernoulli random variables. Now let $X_0 = S_0$ and $X_n = X_{n-1} + B_n$ for $n \geq 1$. Then $(\Omega, \mathcal{F}, P) \times \mathbb{N} \xrightarrow{X=\{X_n\}_{n\in\mathbb{N}}} (\mathbb{Z}, \mathbb{P}(\mathbb{Z}))$ is the process we want.

ii. Or let Ω be the set of all sequences of 1 and -1 of finite lengths then $(\Omega, \mathbb{P}(\Omega))$ is a measurable space. Let $(\Omega, \mathbb{P}(\Omega)) \xrightarrow{P} [0,1], w \mapsto \frac{p^k(1-p)^{n-k}}{2^n}$ be a probability measure where k is the number of 1 in each sequence w of length n. One can verify that $P(\Omega) = 1$ (compare it with P in 4.35.) Now let $X_0 = S_0$ and

$$X_n = \begin{cases} X_{n-1} + n^{\text{th}} \text{ digit in } w, & \text{if } w \text{ has length at least } n. \\ X_{n-1} \text{ otherwise.} \end{cases}$$

for $n \geq 1$. Then $(\Omega, \mathcal{F}, P) \times \mathbb{N} \xrightarrow{X=\{X_n\}_{n\in\mathbb{N}}} (\mathbb{Z}, \mathbb{P}(\mathbb{Z}))$ is another description.

You may even define simple random walk in your own way. Whichever choice you make will affect how you solve the following exercises. When $p = \frac{1}{2}$ this simple random walk is called symmetric, else it is called asymmetric.

Exercise 6.65. Let X be the random process in example 6.64 above.

a. Calculate the probability $P(X_n = m)$ that X_n takes state m. Hint: if a is the number of 1's and b is the number of -1's then $a + b = n$ and $a - b = m$.
b. Calculate the transition probability $P(X_{n_1} = m_1 \,|\, X_{n_2} = m_2)$.
c. Is X homogeneous?
d. Is X stationary?
e. Does it have stationary increments?
f. Does it have independent increments?

If X is a Markov chain and $A \subset S$ is an event then the hitting process $\{T_n^A\}_{n\in\mathbb{N}^+}$ associated to X and A in subsection 6.6 are independent identically distributed random variables since X is memoryless. We can use them to define transiency and recurrency for a state $s \in S$.

Definition 6.66. A state s is said to be transient if, given that we start at s, there is a positive chance that we will leave state s forever, that is $P(T_n^s = \infty) > 0$ for some n and hence $P(T_{n'}^s = \infty) > 0$ for all $n' > n$. A state s is said to be recurrent if it is not transient.

Equivalently, the probability that we will keep returning to a transient state s is less than 1, that is $P(T_n^s < \infty \text{ for all } n) < 1$, and the probability that we will return to a recurrent state s infinitely many times is 1, that is $P(T_n^s < \infty \text{ for all } n) = 1$.

Exercise 6.67. Show that any state s in an asymmetric simple random walk is transient while any state in a symmetric simple random walk is recurrent. This means the gambler in example 6.64 with limited bankroll will always lose if he plays a fair game against a dealer with infinite pocket, his fortune will perform a symmetric simple random walk and reach 0 at some point, at which time the game is over.

6.7.5. *Martingale Process.* A random process $(\Omega, \mathcal{F}, P) \times T \xrightarrow{X} (S, \mathcal{S})$ adapted to a filtration $\{\mathcal{F}_t\}_{t\in T}$ of $\mathcal{F}$ is called martingale if it satisfies the following two conditions

1. $-\infty < E(X_t) < \infty$ for all $t \in T$
2. $E(X_t \,|\, \mathcal{F}_s) = X_s$ for all $s \le t \in T$

The second condition is often called martingale property. It means the expectation of the next value X_t given current knowledge $\mathcal{F}_s$ at current time s is equal to the present observed value X_s. One can not design a winning strategy based on game history in a martingale model. When $T = \mathbb{N}$ is discrete, this property has the simple form

$$E(X_n \,|\, X_{n-1}, \dots, X_0) = X_{n-1}$$

It follows from the law of iterated expectations 4.67 and induction that

$$\begin{aligned} E(X_n) &= E(E(X_n \,|\, X_{n-1}, \dots, X_0)) \\ &= E(X_{n-1}) \\ &\vdots \\ &= E(X_0) \end{aligned}$$

If $E(X_t \,|\, \mathcal{F}_s) \ge X_s$ for all $s \le t \in T$ then X is called submartingale. If $E(X_t \,|\, \mathcal{F}_s) \le X_s$ for all $s \le t \in T$ then X is called supermartingale. Clearly, X is martingale iff it is both submartingale and supermartingale. Compare all this with Markov property.

Exercise 6.68. Show that X is submartingale iff $-X$ is supermartingale.

Example 6.69. If $(\Omega, \mathcal{F}, P) \xrightarrow{X} (S, \mathcal{S})$ is a random variable with finite mean and $\{\mathcal{F}_t\}_{t\in T}$ is a filtration of $\mathcal{F}$ then $Y = \{Y_t\}_{t\in T}$ where $Y_t = E(X \,|\, \mathcal{F}_t)$ is martingale since $E(Y_t \,|\, \mathcal{F}_s) = E(E(X \,|\, \mathcal{F}_t) \,|\, \mathcal{F}_s) = E(X \,|\, \mathcal{F}_s) = Y_s$ for all $s < t$. The law of iterated expectations was used for the second last equality. It also guided us in creating a martingale process from a random variable in a natural way.

Example 6.70. If $X = \{X_t\}_{t\in T}$ is martingale and g is a convex function such that $E(g(X)) < \infty$ for all $t \in T$ then $g(X) = \{g(X)_t\}_{t\in T} = \{g(X_t)\}_{t\in T}$ is submartingale. By Jensen's inequality

$$\begin{aligned} E(g(X)_t \mid \mathcal{F}_s) &= E(g(X_t) \mid \mathcal{F}_s) \\ &\geq g(E(X_t \mid \mathcal{F}_s)) \\ &= g(X_s) \\ &= g(X)_s \end{aligned}$$

Example 6.71. Reconsider the gambler's fortune X in example 6.64. We have

$$\begin{aligned} E(X_n \,|\, X_{n-1}, \dots, X_0) - X_{n-1} &= E(X_n - X_{n-1} \,|\, X_{n-1}, \dots, X_0) \\ &= 1p + (-1)(1-p) \\ &= 2p - 1 \end{aligned}$$

If $p > \frac{1}{2}$ then X is submartingale and you expect to win some money. If $p = \frac{1}{2}$ then X is martingale and you expect to break even. If $p < \frac{1}{2}$ then X is supermartingale and you expect to lose.

Exercise 6.72. One way to get a martingale process Y from X in example 6.64 is to let $Y_n = X_n^2 - n$ where X_n is the gambler's fortune at time n.

a. Verify that $E(Y_n) < \infty$.
b. Show that $E(Y_n \,|\, Y_{n-1}, \dots, Y_0) = Y_{n-1}$ and conclude that Y is martingale.
c. Use (b) and the law of iterated expectations to show that the gambler's total gain or loss varies roughly within $(-\sqrt{n}, \sqrt{n})$.

6.7.6. *Wiener Process.* Random walk is a discrete-time process that enjoys homogeneity and increment independence. Its continuous-time analogue is the Wiener process $(\Omega, \mathcal{F}, P) \times [0, \infty) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that satisfies three conditions

1. $X_0 = 0$.
2. The map $[0, \infty) \xrightarrow{X} RV((\Omega, \mathcal{F}, P), (S, \mathcal{S})), t \mapsto X_t$ is almost surely continuous.
3. X has independent increments with $X_t - X_s \sim X(0, \sigma^2(t-s))$ for all $0 < s < t$ and some constant σ.

It follows from the the first and third conditions that $X_t \sim X(0, \sigma^2 t)$ with density function $f(s) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{s^2}{2\sigma^2 t}}$, which then implies that X has stationary increments.

Example 6.73. (Brownian motion) The Wiener process X with $\sigma = 1$ is called Brownian motion (the process, to describe Brownian motion the phenomenon.)

6.7.7. *Lévy Process.* Both Poisson process and Wiener process belong to a more general class called Lévy processes. A Lévy process $(\Omega, \mathcal{F}, P) \times [0, \infty) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies three conditions

1. $X_0 = 0$ almost surely.
2. X has independent increments.
3. X has stationary increments.

If X is a Lévy process then it has a modification $T \xrightarrow{X'} RV((\Omega, \mathcal{F}, P), (S, \mathcal{S}))$ that is almost càdlàg.

7. Appendix

7.1. **limsup and liminf of a Sequence of Numbers.** Given a sequence $\{u_n, n \geq 1\}$ in a partially order set X, we can discuss $\max\{u_n, n \geq 1\}, \min\{u_n, n \geq 1\}, \sup\{u_n, n \geq 1\}, \inf\{u_n, n \geq 1\}, \limsup\limits_{n\to\infty} u_n$, and $\liminf\limits_{n\to\infty} u_n$.

Definition 7.1. We define $\max\{u_n, n \geq 1\}$ to be the greatest u_n and $\min\{u_n, n \geq 1\}$ to be the smallest u_n if they exist.

Example 7.2. Find max and min of

(1) $\{u_n, n \geq 1\} = \{1, 5, 25, 125, 125, 125, \dots\}$.
(2) $\{u_n, n \geq 1\} = \{\sin(\frac{n\pi}{4}), n \geq 1\}$.

Definition 7.3. We define $\sup\{u_n, n \geq 1\}$ to be the least upper bound of the u_n and $\inf\{u_n, n \geq 1\}$ to be the greatest lower bound of the u_n if they exist.

They always exist if $\{u_n, n \geq 1\}$ is a bounded sequence in $\mathbb{R}$.

Example 7.4. Find $\max, \min, \sup, \inf$ of the following sequences.

a. $\{u_n, n = 1, 2, 3\} = \{1, 3, 5\}$.
b. $\{u_n, n \geq 1\} = \{u_n \in \mathbb{Q}, u_n^2 < 2\}$.
c. $\{u_n, n \geq 1\} = \{e^{-n}, n \geq 1\}$.
d. $\{u_n, n \geq 1\} = \{(-1)^n + \frac{1}{n}, n \geq 1\}$.
e. Can you relate $\inf\{u_n, n \geq 1\}$ and $\sup\{u_n, n \geq 1\}$?

Given a sequence $\{u_n, n \geq 1\}$, consider the decreasing sequence

$$\{a_k, k \geq 1\} \text{ where } a_k = \sup\{u_n, n \geq k\}$$

and the increasing sequence

$$\{b_k, k \geq 1\} \text{ where } b_k = \inf\{u_n, n \geq k\}$$

Also consider the set of accumulation points of $\{u_n, n \geq 1\}$

$$A(u_n, n \geq 1) = \{\text{all } a \text{ such that } a = \lim_{k\to\infty} u_{n_k} \text{ for some subsequence } u_{n_k}\}$$

This set contains its own supremum and infimum.

Definition 7.5. We define $\limsup\limits_{n\to\infty} u_n = \inf\{a_k, k \geq 1\}$ and $\liminf\limits_{n\to\infty} u_n = \sup\{b_k, k \geq 1\}$.

One can show that $\limsup\limits_{n\to\infty} u_n = \sup A(u_n, n \geq 1)$ and $\liminf\limits_{n\to\infty} u_n = \inf A(u_n, n \geq 1)$.

Example 7.6. If

$$u_n = \begin{cases} \frac{1}{n} \text{ if } n = 3k \\ 1 - 2^{-n} \text{ if } n = 3k+1 \\ (1+\frac{1}{n})^n \text{ if } n = 3k+2 \end{cases}$$

then $A(u_n, n \geq 1) = \{0, 1, e\}$ and $\limsup_{n\to\infty} u_n = e$ while $\liminf_{n\to\infty} u_n = 0$.

Example 7.7. Let $u_n = \sin(n)$ for $n = 1, 2, 3, \ldots$. Note that $1, 2, 3, \ldots$ are equidistributed mod π.

a. Find $\sup\{u_n, n \geq 1\}, \inf\{u_n, n \geq 1\}, \limsup_{n\to\infty} u_n$ and $\liminf_{n\to\infty} u_n$.

b. Modify the u_n so that $\sup\{u_n, n \geq 1\} \neq \limsup_{n\to\infty} u_n$.

7.2. **limsup and liminf of a Sequence of Sets.** Given a sequence $\{U_n, n \geq 1\}$ of subsets of X, consider the decreasing sequence

$$\{A_k, k \geq 1\} \text{ where } A_k = \bigcup_{n \geq k} U_n$$

and the increasing sequence

$$\{B_k, k \geq 1\} \text{ where } B_k = \bigcap_{n \geq k} U_n$$

Definition 7.8. We define $\limsup_{n\to\infty} U_n = \bigcap_k A_k$ and $\liminf_{n\to\infty} U_n = \bigcup_k B_k$.

One can show that $u \in \limsup_{n\to\infty} U_n$ iff $u \in U_n$ for infinitely many n and $u \in \liminf_{n\to\infty} U_n$ iff $u \in U_n$ for all but finitely many n.

Example 7.9. If $\{U_n, n \geq 1\}$ is the sequence $\{0\}, \{1\}, \{0\}, \{1\}, \{0\}, \{1\}, \ldots$ then $\limsup_{n\to\infty} U_n = \{0, 1\}$ and $\liminf_{n\to\infty} U_n = \varnothing$.

Example 7.10. If $\{U_n, n \geq 1\}$ is the sequence $\{0,5\}, \{1,5\}, \{0,5\}, \{1\}, \{0\}, \{1\}, \{0\}, \ldots$ then $\limsup_{n\to\infty} U_n = \{0, 1\}$ and $\liminf_{n\to\infty} U_n = \varnothing$.

Example 7.11. If $\{U_n, n \geq 1\}$ is the sequence $\{0\}, \{1\}, \{0\}, \{1,5\}, \{0,5\}, \{1,5\}, \{0,5\}, \ldots$ then $\limsup_{n\to\infty} U_n = \{0, 1, 5\}$ and $\liminf_{n\to\infty} U_n = \{5\}$.

7.3. **Information.**

Definition 7.12. For a discrete random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we define its information to be

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*(P)) \xrightarrow{I_X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$$

$$x \mapsto \begin{cases} 0 \text{ if } X_*(P)(x) = p_X(x) = 0 \\ \log_2 \frac{1}{X_*(P)(x)} = \log_2 \frac{1}{p_X(x)} \text{ otherwise} \end{cases}$$

Then I_X is a discrete random variable with probability mass function

$$\begin{aligned} p_{I_X}(y) &= I_{X*}(X_*(P))(y) \\ &= X_*(P)(I_X^{-1}(y)) \\ &= X_*(P)(\{x_i, \log_2 \frac{1}{p_X(x_i)} = y\}) \\ &= \sum_{x_i, \log_2 \frac{1}{p_X(x_i)} = y} p_X(x_i) \end{aligned}$$

Example 7.13. When X is the Bernoulli random variable with $p_X(1) = p$ and $p_X(0) = 1-p$ then I_X has $p_{I_X}(\log_2 \frac{1}{p}) = p$ and $p_{I_X}(\log_2 \frac{1}{1-p}) = 1-p$.

Example 7.14. When X represents an event $F \in \mathcal{F}$. This should be a constant, perhaps a constant random variable.

Example 7.15. Or when $X = X \mid Y = y$.

Definition 7.16. Given a continuous random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we define its information to be

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*(P)) &\xrightarrow{I_X} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ x &\mapsto \log_2 \frac{1}{f_X(x)} \end{aligned}$$

Then I_X is a continuous random variable with probability density function f_{I_X}?

Here the chosen base is 2 and the unit of I_X is called bits, which are popular in information theory. Other choices are e or 10, and the units are *nat* or *hartley* respectively.

Note that our definitions of information work for multivariate random variables. It also works for conditional probability. If $G \in \mathcal{F}$ is an event of nonzero measure then we can use the conditional probability mass function $p_{X|G}$ and the probability density function $f_{X|G}$ in 4.7.

Example 7.17. If (X, Y) is a discrete bivariate random variable then

$$\begin{aligned} p_{X,Y}(z) &= X_*(P)(I_{X,Y}^{-1}(z)) \\ &= (X,Y)_*(P)(\{(x_i, y_j), \log_2 \frac{1}{p_{X,Y}(x_i, y_j)} = z\}) \\ &= \sum_{(x_i, y_j), \log_2 \frac{1}{p_{X,Y}(x_i, y_j)} = z} p_{X,Y}(x_i, y_j) \end{aligned}$$

Example 7.18. If X is discrete then it has $p_{X|G}$ and its conditional information

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*(P(\mid G)) &\xrightarrow{I_{X|G}} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ x &\mapsto \begin{cases} 0 \text{ if } p_{X|G}(x) = 0 \\ \log_2 \frac{1}{p_{X|G}(x)} \text{ otherwise} \end{cases} \end{aligned}$$

has $p_{I_{X|G}}(y) = \sum_{x_i, \log_2 \frac{1}{p_{X|G}(x)} = y} p_{X|G}(x)$.

Example 7.19. If X is continuous then it has $f_{X\,|\,G}$ and its conditional information

$$\begin{aligned}(\mathbb{R}, \mathcal{B}(\mathbb{R}), X_*(P(\,|G)) &\xrightarrow{I_{X\,|\,G}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))\\ x &\mapsto \log_2 \frac{1}{f_{X\,|\,Y=y}(x)}\end{aligned}$$

has probability density function $f_{I_{X\,|\,G}}$?

Question 7.20. *What is $I_{X\,|\,Y}$ in general? This may lead to $H(X\,|\,Y)$ in the continuous case. How does this compare to $H(I_X\,|\,I_Y)$?*

7.4. **Entropy.**

Definition 7.21. For each random variable $(\Omega, \mathcal{F}, P) \xrightarrow{X} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we define its entropy $H(X)$ to be $E(I_X)$.

The explicit formulas are

$$\begin{aligned}H(X) &= E(I_X)\\ &= \sum_{y_j} p_{I_X}(y_j) y_j\\ &= \sum_{x_i} p_X(x_i) \log_2 \frac{1}{p_X(x_i)}\end{aligned}$$

when X is discrete and

$$\begin{aligned}H(X) &= E(I_X)\\ &= \int_{\mathbb{R}} I_X(x) dI_{X*}(X_*(P))\\ &= \int_{\mathbb{R}} I_X(x) dX_*(P)\\ &= \int_{\mathbb{R}} f_X(x) \log_2 \frac{1}{f_X(x)} dx\end{aligned}$$

when X is continuous.

Example 7.22. The Bernoulli random variable (X, p) has entropy

$$\begin{aligned}H(X) &= p_X(0) \log_2 \frac{1}{p_X(0)} + p_X(1) \log_2 \frac{1}{p_X(1)}\\ &= p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}\end{aligned}$$

It is maximal when $p = \frac{1}{2}$.

Example 7.23. Among the class of all discrete random variables $(\Omega, \mathcal{F}, P) \xrightarrow{X_\alpha} \{x_1, \ldots, x_n\}$ of the same mean $\mu = \frac{x_1 + \cdots + x_n}{n}$, the uniform random variable X with $P(X = x_i) = \frac{1}{n}$ has maximal entropy.

Example 7.24. The exponential random variable (X,λ) with $f_X(x) = \lambda e^{-\lambda x}$ has entropy

$$\begin{aligned}
H(X) &= \int_0^\infty \lambda e^{-\lambda x}\log_2 \frac{1}{\lambda e^{-\lambda x}}dx \\
&= -\int_0^\infty \lambda e^{-\lambda x}\log_2(\lambda e^{-\lambda x})dx \\
&= -\int_0^\infty \lambda e^{-\lambda x}\log_2 e\log_e(\lambda e^{-\lambda x})dx \\
&= -\log_2 e\left(\int_0^\infty \lambda e^{-\lambda x}\log_e\lambda + \int_0^\infty \lambda e^{-\lambda x}(-\lambda x)dx\right) \\
&= -\log_2 e\left(\log_e\lambda\int_0^\infty \lambda e^{-\lambda x}dx - \lambda\int_0^\infty x\lambda e^{-\lambda x}dx\right) \\
&= -\log_2 e\left(\log_e\lambda - \lambda E(X)\right) \\
&= -\log_2 e(\log_e\lambda - 1) \\
&= -\log_2\lambda + \log_2 e
\end{aligned}$$

Example 7.25. Among the class of all continuous random variables $(\Omega,\mathcal{F},P)\xrightarrow{X_\alpha}(\mathbb{R},\mathcal{B}(\mathbb{R}))$ of the same standard deviation σ, the normal random variables X with that standard deviation (and any mean) have maximal entropy.

Example 7.26. The entropy of a discrete bivariate random variable (X,Y) is

$$\begin{aligned}
H(X,Y) &= E(I_{X,Y}) \\
&= \sum_{z_k} p_{X,Y}(z_k)z_k \\
&= \sum_{x_i,y_j} p_{X,Y}(x_i,y_j)\log_2\frac{1}{p_{X,Y}(x_i,y_j)}
\end{aligned}$$

Example 7.27. The entropy of a continuous bivariate random variable (X,Y) is

$$\begin{aligned}
H(X,Y) &= E(I_{X,Y}) \\
&= \iint_{\mathbb{R}^2} f_{X,Y}(x,y)log_2\frac{1}{f_{X,Y}(x,y)}dxdy
\end{aligned}$$

Example 7.28. The conditional entropy of a discrete random variable X given $Y = y$ is

$$\begin{aligned}
H(X\,|\,Y=y) &= E(I_{X\,|\,Y=y}) \\
&= \sum_{x_i} p_{X\,|\,Y=y}(x_i)\log_2\frac{1}{p_{X\,|\,Y=y}(x)}
\end{aligned}$$

The map

$$\begin{aligned} (\mathbb{R}, \mathcal{B}(\mathbb{R}), Y_*(P)) &\xrightarrow{H(X\,|\,Y=-)} (\mathbb{R}, \mathcal{B}(\mathbb{R})) \\ y &\mapsto H(X\,|\,Y=y) \end{aligned}$$

is indeed a random variable, which means it has expectation, variance, etc.

Definition 7.29. We define conditional entropy $H(X\,|\,Y)$ of X given Y to be $E(H(X\,|\,Y = -))$.

Explicitly, we have

$$\begin{aligned} H(X\,|\,Y) &= \sum_{y_j} Y_*(P)(y_j) H(X\,|\,Y=y) \\ &= \sum_{y_j} p_Y(y_j) \sum_{x_i} p_{X\,|\,Y=y_j}(x_i) \log_2 \frac{1}{p_{X\,|\,Y=y_j}(x_i)} \end{aligned}$$

When X, Y are joint,

$$\begin{aligned} H(X\,|\,Y) &= \sum_{x_i,y_j} p_Y(y_j) p_{X\,|\,Y=y_j}(x_i) \log_2 \frac{1}{p_{X\,|\,Y=y_j}(x_i)} \\ &= \sum_{x_i,y_j} p_{X,Y}(x_i,y_j) \log_2 \frac{1}{p_{X\,|\,Y=y_j}(x_i)} \end{aligned}$$

And when X, Y are independent,

$$\begin{aligned} H(X,Y) &= \sum_{x_i,y_j} p_{X,Y}(x_i,y_j) \log_2 \frac{1}{p_{X,Y}(x_i,y_j)} \\ &= \sum_{x_i,y_j} p_{X,Y}(x_i,y_j) \log_2 \frac{p_X(x_i)p_Y(y_j)}{p_{X,Y}(x_i,y_j)p_{X,Y}(x_i,y_j)} \\ &= \sum_{x_i,y_j} p_{X,Y}(x_i,y_j) \left(\log_2 \frac{1}{p_{X\,|\,Y=y_j}(x_i)} + \log_2 \frac{1}{p_{Y\,|\,X=x_i}(y_j)} \right) \\ &= \sum_{x_i,y_j} p_{X,Y}(x_i,y_j) \log_2 \frac{1}{p_{X\,|\,Y=y_j}(x_i)} + \sum_{x_i,y_j} p_{X,Y}(x_i,y_j) \log_2 \frac{1}{p_{Y\,|\,X=x_i}(y_j)} \\ &= H(X\,|\,Y) + H(Y\,|\,X) \end{aligned}$$

The case for continuous X, Y runs parallel. Explicitly, we have

$$\begin{aligned} H(X\,|\,Y) &= \int_{\mathbb{R}} f_Y(y) \int_{\mathbb{R}} f_{X\,|\,Y}(x) \log_2 \frac{1}{f_{X\,|\,Y}(x)} dx \\ &= \iint_{\mathbb{R}^2} f_Y(y) f_{X\,|\,Y}(x) \log_2 \frac{1}{f_{X\,|\,Y}(x)} dxdy \end{aligned}$$

When X, Y are joint,

$$\begin{aligned} H(X \mid Y) &= \iint\limits_{\mathbb{R}^2} f_Y(y) f_{X \mid Y)}(x) \log_2 \frac{1}{f_{X \mid Y}(x)} dxdy \\ &= \iint\limits_{\mathbb{R}^2} f_{X,Y}(x,y) \log_2 \frac{1}{f_{X \mid Y}(x)} dxdy \end{aligned}$$

And when X, Y are independent,

$$\begin{aligned} H(X,Y) &= \iint\limits_{\mathbb{R}^2} f_{X,Y}(x,y) \log_2 \frac{1}{f_{X,Y}(x,y)} dxdy \\ &= \iint\limits_{\mathbb{R}^2} f_{X,Y}(x,y) \log_2 \frac{1}{f_{X \mid Y}(x) f_{Y \mid X}(y)} dxdy \\ &= \iint\limits_{\mathbb{R}^2} f_{X,Y}(x,y) \left(\log_2 \frac{1}{f_{X \mid Y}(x)} + \log_2 \frac{1}{f_{Y \mid X}(y)} \right) dxdy \\ &= \iint\limits_{\mathbb{R}^2} f_{X,Y}(x,y) \log_2 \frac{1}{f_{X \mid Y}(x)} dxdy + \iint\limits_{\mathbb{R}^2} f_{X,Y}(x,y) \log_2 \frac{1}{f_{Y \mid X}(y)} dxdy \\ &= H(X \mid Y) + H(Y \mid X) \end{aligned}$$

Question 7.30. *If* $\mathbb{R}^2 \xrightarrow{I_{X \mid Y=-}} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *is a bivariate random variable while* $(\mathbb{R}, \mathcal{B}(\mathbb{R}), Y_*(P)) \xrightarrow{H(X \mid Y=-)} (\mathbb{R},$ *and* $(\Omega, \mathcal{F}, P) \xrightarrow{EX \mid Y)} (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *are random variable then* $E(I_{X \mid Y=-}) = E(H(X \mid Y = -)) = H(E(X \mid Y))$*? Do we have* $EH = HE$*, and what does that mean?*

Question 7.31. *How to define* $I(X,Y) \neq I_{X,Y}$*?*

Work out the difference between $I_X, I_{X,Y}$ (random variables) and $I(X), I(X,Y)$ (numbers). How does each relate to entropy H?

Go through $X_*(P(\,|G))$ to $p_{X \mid G}$ and $f_{X \mid G}$.

Conditional entropy for continuous $X, Y, H(X,Y), H(X|Y), H(Y|X)$ with identity.

References