MATH 3235: Probability Theory

Frank Qiang
Instructor: Christian Houdre

Georgia Institute of Technology
Fall 2024

Contents

Chapter 1

Events and Probabilities

1.1 Probability Spaces

Definition 1.1. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

- Ω is called the *sample space* (the set of all possible outcomes of a random experiment);
- $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, called the *event space*,[1] is nonempty and must satisfy:
  - (i) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,
  - (ii) if $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$;
- $\mathbb{P}$ is a probabilty measure on $(\Omega, \mathcal{F})$ (to be defined later).

Remark. In general, when Ω is finite or countably infinite, one takes $\mathcal{F} = \mathcal{P}(\Omega)$.

Proposition 1.1. *We always have* $\varnothing, \Omega \in \mathcal{F}$.

Proof. Since $\mathcal{F} \neq \varnothing$, there exists some event $A \in \mathcal{F}$. Then we get $A^c \in \mathcal{F}$ and $\Omega = A \cup A^c \in \mathcal{F}$ by the complement and union properties of $\mathcal{F}$. Finally $\varnothing = \Omega^c \in \mathcal{F}$ by the complement property. □

Definition 1.2. A *probability measure* on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, \infty)$ such that

- (i) $\mathbb{P}(\Omega) = 1$,
- (ii) and $\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$ whenever $A_1, A_2, \cdots \in \mathcal{F}$ are pairwise disjoint.[2]

Proposition 1.2. *The following properties hold for any probability measure* $\mathbb{P}$ *on* $(\Omega, \mathcal{F})$*:*

(1) For any $A \in \mathcal{F}$*, we have* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(2) Let $A, B \in \mathcal{F}$ *with* $A \subseteq B$*. Then* $\mathbb{P}(A) \leq \mathbb{P}(B)$.

(3) Let $A, B, C \in \mathcal{F}$*. Then*

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

This is the principle of inclusion-exclusion.

Proof. (1) Observe that $A \cup A^c = \Omega$ and $A \cap A^c = \varnothing$, so $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$.

[1] The elements of $\mathcal{F}$ are called *events*. Events with cardinality 1 are called *elementary*.

[2] i.e. $A_i \cap A_j = \varnothing$ whenever $i \neq j$.

(2) Write $B = A \cup (B \setminus A)$.[3] Since $A \cap (B \setminus A) = \varnothing$, we have $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A)$.[4]

(3) Left as an exercise. Follow similar ideas as in (2). □

Remark. Observe that property (2) implies $\mathbb{P}(A) \leq \mathbb{P}(\Omega) = 1$ since any $A \subseteq \Omega$.

Example 1.2.1. Pick a point uniformly at random from the unit square $\Omega = [0,1] \times [0,1]$ and record its coordinates. Then the probability of the point being inside a fixed shape $S \subseteq \Omega$ is $|S|$, the area of S.

Remark. Note that $\mathbb{P}$ only satisfies *countable* additivity. For instance let $\Omega = [0,1]$ and $\mathbb{P}$ be the uniform measure on Ω. Then $\Omega = \bigcup_{x \in [0,1]} \{x\}$ and $\mathbb{P}(\{x\}) = 0$ for every $x \in [0,1]$, but $\mathbb{P}(\Omega) = 1$. This is because the union $\bigcup_{x \in [0,1]} \{x\}$ is uncountable.

Definition 1.3. Let Ω be finite and $\mathcal{F} = \mathcal{P}(\Omega)$. The uniform probability on $(\Omega, \mathcal{F})$ is the one such that

$$\mathbb{P}(\{\omega\}) = \frac{1}{\operatorname{card} \Omega} \quad \text{for all } \omega \in \Omega.$$

Proposition 1.3. *Let $\mathbb{P}$ be the uniform probability on a finite set Ω and let $A \in \mathcal{F}$. Then*

$$\mathbb{P}(A) = \frac{\operatorname{card} A}{\operatorname{card} \Omega}.$$

Proof. Note that A is finite since Ω is and so we may enumerate its elements as $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$, where $n = \operatorname{card} A$. Then the sets $\{\omega_i\}_{i=1}^n$ are pairwise disjoint and thus we have

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i=1}^{n} \{\omega_i\}\right) = \sum_{i=1}^{n} \mathbb{P}(\{\omega_i\}) = \sum_{i=1}^{n} \frac{1}{\operatorname{card} \Omega} = \frac{n}{\operatorname{card} \Omega} = \frac{\operatorname{card} A}{\operatorname{card} \Omega},$$

which is the desired result. □

1.2 Conditional Probability

Definition 1.4. Let $B \in \mathcal{F}$ such that $\mathbb{P}(B) > 0$. Then the *conditional probability* of A given B, written $\mathbb{P}(A|B)$, is given by

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

Remark. The intuition is that the extra information gained by knowing the occurrence of B should update our computation of the probability of A.

Remark. Another way to think about conditional probability is a restriction of the sample space to B.

Example 1.4.1. Suppose a family has two children, one of which is a girl. What is the probability the other is a girl? Define the sample space to be

$$\Omega = \{(B,G), (B,B), (G,G), (G,B)\}.$$

[3] Note that $B \setminus A \in \mathcal{F}$ since $B \setminus A = B \cap A^c = (B^c \cup A)^c$.

[4] Since $\mathbb{P} : \mathcal{F} \to [0, \infty)$, we have $\mathbb{P}(B \setminus A) \geq 0$.

Note that each elementary event is equally likely, i.e.

$$\mathbb{P}(\{(B,G)\}) = \mathbb{P}(\{(G,B)\}) = \mathbb{P}(\{(B,B)\}) = \mathbb{P}(\{(G,G)\}) = \frac{1}{4}.$$

Let $A = \{\text{both of them are } G\} = \{(G,G)\}$ and $B = \{\text{one is a girl}\} = \{(G,B),(B,G),(G,G)\}$. Then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{(G,G)\})}{\mathbb{P}(\{(G,B),(B,G),(G,G)\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Remark. If we instead condition on the event that one of them is a girl born on a Monday, then the probability changes! Carry out the calculation, and it should be 13/27 for a 7-day week.

Example 1.4.2. A drug test is 98% accurate, i.e. a drug user tests positive 98% of the time and a non-drug user tests negative 98% of the time. Among a given population, it is known 2% of people use drugs. Suppose I pick a person at random in the population and this person tests positive. What is the probability that the person is a drug user? Define the events

$$A = \text{the person is a drug user} \quad \text{and} \quad B = \text{the person tests positive.}$$

Then the goal is to compute $\mathbb{P}(A|B)$. The 98% accuracy assumption implies that $\mathbb{P}(B|A) = 0.98$. Now

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

Note that $B = (B \cap A) \cup (B \cap A^c)$ and this is a disjoint union, so

$$\begin{aligned}\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) &= \mathbb{P}(B|A)\mathbb{P}(A) \\ &+\mathbb{P}(B|A^c)\mathbb{P}(A^c) = 0.98(0.02) + 0.02(0.98) = 2(0.98)(0.02).\end{aligned}$$

Here we noted that the 98% accuracy of the test also implies that $\mathbb{P}(B|A^c) = 0.02$. Thus we get

$$\mathbb{P}(A|B) = \frac{0.98(0.02)}{2(0.98)(0.02)} = \frac{1}{2}.$$

Compute as an exercise that $\mathbb{P}(A^c|B^c) = 0.996$.

Remark. This test is designed clear non-drug users, not to identify drug users.

1.3 Bayes' Theorem

Definition 1.5. A *partition* of Ω is a collection or sequence of events $\{B_k\}_{k=1}^{\infty}$ such that

$$B_i \cap B_j = \varnothing \text{ for } i \neq j \quad \text{and} \quad \Omega = \bigcup_{k=1}^{\infty} B_k.$$

Remark. For any event A, observe that

$$A = A \cap \Omega = A \cap \left(\bigcup_{k=1}^{\infty} B_k \right) = \bigcup_{k=1}^{\infty} (A \cap B_k).$$

Then we get

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{k=1}^{\infty}(A \cap B_k)\right) = \sum_{k=1}^{\infty}\mathbb{P}(A \cap B_k) = \sum_{k=1}^{\infty}\mathbb{P}(A|B_k)\mathbb{P}(B_k).$$

This is the *partition theorem* in the book (Grimmett and Welsh). Now observe that

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B_i \cap A)}{\sum_{k=1}^{\infty}\mathbb{P}(A|B_k)\mathbb{P}(B_k)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{k=1}^{\infty}\mathbb{P}(A|B_k)\mathbb{P}(B_k)}$$

for each $i = 1, 2, \ldots$. This is *Bayes' theorem*, which relates posterior probabilities to prior probabilities.

1.4 Conditional Independence

Proposition 1.4. *Let B be such that $\mathbb{P}(B) > 0$. Then $Q : \mathcal{F} \to [0,1]$ given by $A \mapsto Q(A) = \mathbb{P}(A|B)$ is a probability measure.*

Proof. (i) Observe that

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

(ii) Let $\{A_k\}_{k=1}^{\infty} \subseteq \mathcal{F}$ be pairwise disjoint. Then observe that we have

$$\begin{aligned} Q\left(\bigcup_{k=1}^{\infty}A_k\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty}A_k|B\right) &= \frac{\mathbb{P}\left(\bigcup_{k=1}^{\infty}(A_k \cap B)\right)}{\mathbb{P}(B)} \\ &= \frac{\sum_{k=1}^{\infty}\mathbb{P}(A_k \cap B)}{\mathbb{P}(B)} = \sum_{k=1}^{\infty}\mathbb{P}(A_k|B) = \sum_{k=1}^{\infty}Q(A_k). \end{aligned}$$

Thus Q is indeed a probability measure. □

Definition 1.6. Two events A and B are called *independent*, written $A \perp\!\!\!\perp B$, if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Example 1.6.1. Let A, B be events with $\mathbb{P}(A) = 0.6$ and $\mathbb{P}(B) = 0.8$. Then $0.4 \leq \mathbb{P}(A \cap B) \leq 0.6$. This is because $A \cap B \subseteq A$ implies $\mathbb{P}(A \cap B) \leq \mathbb{P}(A) = 0.6$. Also noting that $A \cap B \subseteq \Omega$ and so $\mathbb{P}(A \cap B) \leq \mathbb{P}(\Omega) = 1$ implies that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B) \geq 0.6 + 0.8 - 1 = 0.4.$$

Note that if A and B were independent, then we can immediately conclude $\mathbb{P}(A \cap B) = 0.6(0.8) = 0.48$.

Proposition 1.5. *Assume A and B are events with $0 < \mathbb{P}(A), \mathbb{P}(B) < 1$. The following are equivalent:*

1. *A and B are independent,*
2. *$\mathbb{P}(A|B) = \mathbb{P}(A)$,*
3. *$\mathbb{P}(B|A) = \mathbb{P}(B)$,*
4. *$\mathbb{P}(A^c|B) = \mathbb{P}(A^c)$,*
5. *and $\mathbb{P}(B^c|A) = \mathbb{P}(B^c)$.*

Proof. Note that $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B)$ and $A \perp\!\!\!\perp B$ if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Then use cancellation since $\mathbb{P}(B) \neq 0$. Work out the rest as an exercise. □

Definition 1.7. We say that three events A, B, C are *independent* if A, B, C are pairwise independent[5] and $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

Remark. Pairwise independence does not imply independence. Consider flipping a fair coin twice. Let

$$A = \{\text{first flip is } T\}, \quad B = \{\text{second flip is } H\}, \quad C = \{\text{both flips are the same}\}.$$

Then A, B, C are pairwise independent but $\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = 1/8$.

1.5 Continuity of Probability Measures

Remark. We want a system of probability that can say if I flip a fair coin infinitely many times, then

$$\mathbb{P}(\text{never get heads}) = 0.$$

For this experiment, we can set the sample space to be

$$\Omega = \{\text{all sequences like } (H, T, T, \dots)\}.$$

The event space $\mathcal{F}$ is a little complicated, but it includes events like $\{$heads on the nth throw$\}$ and their complements and countable unions. We would like to show that $\mathbb{P}(\{(T, T, T, \dots)\}) = 0$. We know that

$$\mathbb{P}(A_n) = \mathbb{P}(\text{no heads in first } n \text{ tosses}) = \frac{1}{2^n}.$$

As $n \to \infty$, we can see that $2^{-n} \to 0$. If we have $\mathbb{P}(A_n) \to \mathbb{P}(\{(T, T, T, \dots)\})$, then we can conclude that

$$\mathbb{P}(\{(T, T, T, \dots)\}) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \frac{1}{2^n} = 0.$$

Notice that $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$ and $\bigcap_{n=1}^{\infty} A_n = \{(T, T, T, \dots)\}$. For sake of convenience, we will take complements and work with unions: We set

$$B_n = A_n^c = \{\text{at least one heads in first } n \text{ tosses}\},$$

then $B_1 \subseteq B_2 \subseteq B_3 \subseteq \dots$ and $\bigcup_{n=1}^{\infty} B_n = \Omega \setminus \{(T, T, T, \dots)\}$. But this union is not disjoint. To fix this, set $C_i = B_i \setminus B_{i-1}$, then

$$B_1 \cup \bigcup_{n=2}^{\infty} C_n = \Omega \setminus \{(T, T, T, \dots)\},$$

which is now a disjoint union. Taking probabilities, we can use countable additivity to get

$$\mathbb{P}(B_1 \cup C_2 \cup C_3 \cup \dots) = \mathbb{P}(B_1) + \mathbb{P}(C_2) + \mathbb{P}(C_3) + \dots. \tag{$*$}$$

First, note that $B_1 \cup C_2 \cup C_3 \cup \dots = B_1 \cup B_2 \cup B_3 \cup \dots = \Omega \setminus \{(T, T, T, \dots)\}$. Thus in $(*)$, the LHS is $1 - \mathbb{P}(\{(T, T, T, \dots)\})$. Now observe that in the RHS, we have $\mathbb{P}(C_i) = \mathbb{P}(B_i) - \mathbb{P}(B_{i-1})$. Then

$$\begin{aligned}
\mathbb{P}(B_1) + \mathbb{P}(C_2) + \mathbb{P}(C_3) + \dots &= \lim_{n\to\infty} [\mathbb{P}(B_1) + \mathbb{P}(C_2) + \dots + \mathbb{P}(B_n)] \\
&= \lim_{n\to\infty} [\mathbb{P}(B_1) + (\mathbb{P}(B_2) - \mathbb{P}(B_1)) + \dots + (\mathbb{P}(B_n) - \mathbb{P}(B_{n-1}))] \\
&= \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} [1 - \mathbb{P}(A_n)] = \lim_{n\to\infty} \left[1 - \frac{1}{2^n}\right] = 1.
\end{aligned}$$

Thus matching the LHS and RHS, we see that $1 - \mathbb{P}(\{(T, T, T, \dots)\}) = 1$ and so $\mathbb{P}(\{(T, T, T, \dots)\}) = 0$.

[5] i.e. $\mathbb{P}(X \cap Y) = \mathbb{P}(X)\mathbb{P}(Y)$ for any $X, Y \in \{A, B, C\}$.

Theorem 1.1 (Continuity of probability measures)**.** *If* $\{B_i\}_{i=1}^{\infty}$ *are nested events,*[6] *then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

Proof. This was essentially the argument in the previous remark. □

1.6 Homework Problems

Problems #1, 2, 9, 10, 14, 16, 17, 19 from Grimmett and Welsh.

[6] i.e. $B_1 \subseteq B_2 \subseteq B_3 \subseteq \dots$.

Chapter 2

Discrete Random Variables

2.1 Probability Mass Functions

Example 2.0.1. Consider the following game: Flip a fair coin 10 times and roll a fair die. I give you

$$(\text{number of heads}) \times (\text{number on die}) \text{ dollars.}$$

This is a simple game, but it is kind of painful to write in terms of events (e.g. $\mathbb{P}(\text{win} \geq \$10)$)). We would have to set

$$\Omega = \{\text{all sequences like } (H, T, H, H, T, T, T, T, T, H, 4)\}$$

and $\mathcal{F} = \mathcal{P}(\Omega)$. It is also not immediately obviously which sequences are in $\{\text{win} \geq \$10\}$. Instead, we would prefer something like

> "Let H be the number of heads in 10 fair coin tosses and let R be the outcome of a roll of a fair die. Then you get HR dollars."

How do we do this in our axiomatic framework? What are H, R? Here are some observations:

- H, R are real numbers,
- and they are determined by the outcome of the experiment.

Thus we should think of H, R as functions from Ω to $\mathbb{R}$. These are examples of *discrete random variables*.

Remark. The name "random variable" is just historic. Really, H, R are non-random functions.

Remark. Can every function $X : \Omega \to \mathbb{R}$ be a discrete random variable? Note that we want to talk about probabilities like $\mathbb{P}(X = 17)$. This indicates that the event

$$\{X = 17\} = \{\omega \in \Omega : X(\omega) = 17\}$$

has to be in $\mathcal{F}$. So we require that X is *measurable*, i.e. for every $x \in \mathbb{R}$, we have $\{x \in \Omega : X(\omega) = x\} \in \mathcal{F}$. Also H, R must have special properties, for instance they can only take on finitely many values.

Definition 2.1. A function $X : \Omega \to \mathbb{R}$ is a *discrete random variable* if

(i) for every $x \in \mathbb{R}$, we have $\{X = x\} \in \mathcal{F}$,

(ii) and $X(\Omega) = \{x \in \mathbb{R} : x = X(\omega) \text{ for some } \omega\}$ is finite or countably infinite.

Remark. Often, we only care about what values X can take and with what probabilities. We store this data in a special function called the *probability mass function*.

Definition 2.2. Let X be a discrete random variable. Then its *probability mass function (pmf)* is

$$p_X : \mathbb{R} \to [0,1] \quad \text{defined by} \quad p_X(s) = \mathbb{P}(X = s).$$

Example 2.2.1. Let X be the outcome of the roll of a fair die. Then

$$p_X(s) = \begin{cases} 1/6 & \text{if } s \in \{1,2,3,4,5,6\}, \\ 0 & \text{otherwise.} \end{cases}$$

Remark. Another sentence we want to say is:

> "A discrete random variable X takes values $\{1,7,9\}$ with probabilities $1/2, 1/3, 1/6$, respectively if and only if
>
> $$p_X(s) = \begin{cases} 1/2 & \text{if } s = 1, \\ 1/3 & \text{if } s = 7, \\ 1/6 & \text{if } s = 9, \\ 0 & \text{otherwise.} \end{cases}$$

How do we know this exists? In other words, does there exist $(\Omega, \mathcal{F}, \mathbb{P})$ and $X : \Omega \to \mathbb{R}$ with this pmf?

Theorem 2.1. *Let $S = \{s_i : i \in I\}$ be a countable subset of $\mathbb{R}$ and let $\{\pi_i : i \in I\}$ be a collection of numbers such that $\pi_i \geq 0$ and*

$$\sum_{i \in I} \pi_i = 1.$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a discrete random variable $X : \Omega \to \mathbb{R}$ such that

$$p_X(s) = \begin{cases} \pi_i & \text{if } s = s_i, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Take $\Omega = S$ and $\mathcal{F} = \mathcal{P}(S)$. Set

$$\mathbb{P}(A) = \sum_{i : s_i \in A} \pi_i$$

and define $X : \Omega \to \mathbb{R}$ given by $X(\omega) = \omega$. Then one can check that X has the desired pmf. □

Remark. This allows us to just say

> "Let X be a discrete random variable taking these values with these probabilities"

without worrying about the underlying $(\Omega, \mathcal{F}, \mathbb{P})$.

2.2 Common Discrete Random Variables

Example 2.2.2. Some common examples of discrete random variables are:

1. *Constant random variables*: Define $X : \Omega \to \mathbb{R}$ by $\omega \mapsto X(\omega) = C$.

2. *Bernoulli random variables*: For $0 < p < 1$, we say that $X \sim \text{Ber}(p)$ if

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } q = 1 - p. \end{cases}$$

This models a possibly unfair coin flip. The Bernoulli random variable X has pmf

$$p_X(s) = \begin{cases} p & \text{if } s = 1, \\ 1 - p & \text{if } s = 0, \\ 0 & \text{otherwise.} \end{cases}$$

3. *Binomial random variables*: For $n \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and $0 < p < 1$, we say that $X \sim \text{Bin}(n, p)$ if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

for $k = 0, 1, \ldots, n$ and $\mathbb{P}(X = k) = 0$ otherwise. To that this is indeed a pmf, observe that

$$\sum_{k=0}^{n} \mathbb{P}(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1^n = 1.$$

The $n = 1$ case reduces to a Bernoulli random variable.

4. *Geometric random variables*: For $0 < p < 1$, we say that $X \sim \text{Geo}(p)$ if

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

for $k = 1, 2, 3, \ldots$ and $\mathbb{P}(X = k) = 0$ otherwise. The above function is clearly nonnegative and

$$\sum_{k=1}^{\infty} p(1 - p)^{k-1} = \frac{p}{1 - (1 - p)} = 1,$$

so this is indeed a pmf. The geometric random variable models the number of independent Bernoulli trials needed to obtain the first success.

Example 2.2.3. Consider the random variable X which counts the number of independent Bernoulli trials needed to get the 4th success. Note that the range of X is $\{4, 5, 6, \ldots\}$. Then

$$\mathbb{P}(X = k) = \binom{k-1}{3} p^3 (1 - p)^{k-4} p = \binom{k-1}{3} p^4 (1 - p)^{k-4}$$

for $k = 4, 5, 6, \ldots$ and $\mathbb{P}(X = k) = 0$ otherwise. This is because the last trial must be a success and the previous $k - 1$ trials need to contain 3 successes. Here $X = \text{NBin}(n = 4, p)$, the *negative binomial random variable*. In general, $X \sim \text{NBin}(n, p)$ takes on values $n, n + 1, n + 2, \ldots$ and

$$\mathbb{P}(X = k) = \binom{k-1}{n-1} p^n (1 - p)^{k-n}$$

for $k = n, n + 1, n + 2, \ldots$. Note that the $n = 1$ case reduces to a geometric random variable. The name comes from the binomial theorem with negative exponents.

Example 2.2.4. We say that X is a *Poisson random variable* with parameter $\lambda > 0$, written $X \sim \text{Poi}(\lambda)$, if X takes the values $k = 0, 1, 2, \ldots$ with probability mass function

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

Note that p_X is clearly nonnegative and

$$\sum_{k=0}^{\infty} p_X(k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1,$$

so p_X is indeed a pmf. One can view the Poisson random variable in the following manner: Suppose $X \sim \text{Bin}(n, p)$ with $n \gg 1$ and $p \ll 1$, e.g. $n = 10^5$ and $p = 10^{-4}$. Then

$$\mathbb{P}(X = 100) = \binom{10^5}{100} \left(\frac{1}{10^4}\right)^{100} \left(1 - \frac{1}{10^4}\right)^{10^5 - 100}.$$

This is very difficult to compute. Instead, we approximate this via the Poisson random variable.

Proposition 2.1. *Let $n \to \infty$ and $p = p(n) \to 0$ in such a way that $np(n) \to \lambda > 0$ as $n \to \infty$. Then*

$$\binom{n}{k} p^k (1-p)^{n-k} \xrightarrow[n\to\infty]{} \frac{e^{-\lambda}\lambda^k}{k!},$$

i.e. $p_X(k) \to p_Y(k)$ pointwise for $k = 0, 1, 2, \ldots$, where $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Poi}(\lambda)$.

Proof. Observe that

$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{1}{k!}\left[n(n-1)\ldots(n-k+1)p^k(1-p)^{-k}(1-p)^n\right] \\ &= \frac{1}{k!}\left[\frac{n(n-1)\ldots(n-k+1)}{n^k} n^k p^k (1-p)^{-k}(1-p)^n\right].\end{aligned}$$

Now notice that $n^k p^k = (np)^k \to \lambda^k$ since $np \to \lambda$, and

$$\lim_{n\to\infty} \frac{n(n-1)\ldots(n-k+1)}{n^k} = 1 \quad \text{and} \quad \lim_{p\to 0}(1-p)^{-k} = 1.$$

Finally, setting $\lambda = np$,

$$(1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}.$$

Putting all of this together, we see that

$$\binom{n}{k} p^k (1-p)^{n-k} \xrightarrow[n\to\infty]{} \frac{e^{-\lambda}\lambda^k}{k!},$$

which is the desired result. □

2.3 Expectation of Random Variables

Remark. Suppose $X : \Omega \to \mathbb{R}$ is a discrete random variable and $h : \mathbb{R} \to \mathbb{R}$. Then we have:

$$\begin{array}{ccc} \Omega & \xrightarrow{X} & \mathbb{R} \\ & \searrow^{h(X)} & \downarrow h \\ & & \mathbb{R} \end{array}$$

In particular, $h \circ X : \Omega \to \mathbb{R}$ is also a random variable.

Definition 2.3. Let X be a discrete random variable. The *(mathematical) expectation* of X is[1]

$$\mathbb{E}[X] = \sum_{x \in \mathcal{R}(X)} x p_X(x)$$

if the above sum exists and converges absolutely,[2] where p_X is the probability mass function of X.

Remark. When X is discrete, the expectation coincides with the usual notion of a mean. In general, the expectation is some kind of weighted mean.

Remark. Observe that the sum in the definition of $\mathbb{E}[X]$ need not converge. Even worse, if it only converges conditionally, then by the Riemann rearrangement theorem we may get any real value we wish by reordering the sum. This is why we require absolute convergence.

Example 2.3.1. Set $Y = X^2$. Then we have

$$\mathbb{E}[X^2] = \mathbb{E}[Y] = \sum_{y \in \mathcal{R}(Y)} y\mathbb{P}(Y = y) = \sum_{y \in \mathcal{R}(X^2)} y\mathbb{P}(X^2 = y).$$

If we explicitly let

$$X = \begin{cases} 1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2, \end{cases}$$

we see that $\mathbb{E}[X] = 0$. We can also see that $\mathbb{E}[X^2] = 1$ since $X^2 = 1$ with probability 1. Equivalently, we can compute that

$$\mathbb{E}[X^2] = \sum_{y \in \mathcal{R}(X^2)} y\mathbb{P}(X^2 = y) = 1 \cdot \mathbb{P}(X^2 = 1) = 1.$$

Proposition 2.2 (Law of the unconcious statistician)**.** *For any $h : \mathbb{R} \to \mathbb{R}$ and $X : \Omega \to \mathbb{R}$ discrete,*

$$\mathbb{E}[h(X)] = \sum_{x \in \mathcal{R}(X)} h(x) p_X(x)$$

where p_X is the pmf of X, provided these sums exist and converge absolutely.

Proof. Let $Y = h(X)$. Then we have

$$\mathbb{E}[h(X)] = \mathbb{E}[Y] = \sum_{y \in \mathcal{R}(Y)} y p_Y(y) = \sum_{y \in \mathcal{R}(Y)} y\mathbb{P}(h(X) = y) = \sum_{x \in \mathcal{R}(X)} h(x)\mathbb{P}(h(X) = y).$$

[1] We write $\mathcal{R}(X)$ to denote the range of X.
[2] i.e. $\sum_{x \in \mathcal{R}(X)} |x| p_X(x) < \infty$.

Note that that y in the last term is $h(x)$, and thus $\mathbb{P}(h(X) = y) = \mathbb{P}(h(X) = h(x)) = p_X(x)$. Then

$$\mathbb{E}[h(X)] = \sum_{x \in \mathcal{R}(X)} h(x) p_X(x),$$

which is precisely the desired result. □

Remark. In the discrete case, we do not require that h be measurable since $\mathcal{R}(X)$ is at most countable.

Proposition 2.3. *We have the following properties of expectation:*

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.

(ii) If $X = C$ is constant, then $\mathbb{E}[X] = C$.

(iii) If $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Proof. (i) Since $X \geq 0$, we have $\mathcal{R}(X) \subseteq [0, \infty)$ and thus

$$\mathbb{E}[X] = \sum_{x \in \mathcal{R}(X)} x p_X(x) \geq 0$$

since every term in the sum is nonnegative.

(ii) Since $\mathcal{R}(X) = \{C\}$, we have $\mathbb{P}(X = C) = 1$ and thus

$$\mathbb{E}[X] = \sum_{x \in \mathcal{R}(X)} x p_X(x) = C \cdot \mathbb{P}(X = C) = C.$$

This is the desired result.

(iii) We compute that

$$\begin{aligned}\mathbb{E}[aX + bY] &= \sum_{ax+by \in \mathcal{R}(aX+bY)} (ax + by) p_{aX+bY}(ax + by) = \sum_{x \in \mathcal{R}(X)} axp_X(x) + \sum_{y \in \mathcal{R}(Y)} byp_Y(y) \\ &= a \sum_{x \in \mathcal{R}(X)} x p_X(x) + b \sum_{y \in \mathcal{R}(Y)} y p_Y(y) = a\mathbb{E}[X] + b\mathbb{E}[Y],\end{aligned}$$

which is the desired equality. □

Example 2.3.2. We compute the following:

1. Let $X \sim \text{Ber}(p)$. Then $\mathbb{E}[X] = 0(1 - p) + 1p = p$.
2. Let $X \sim \text{Bin}(n, p)$. Then

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} = n \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= n \sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-1-(k-1)} = np.\end{aligned}$$

In the last step we re-index with $j = k-1$, and then recognize the terms as the pmf of a $\text{Bin}(n-1, p)$ random variable, which must sum to 1 over $0 \leq j \leq n-1$.

3. Let $X \sim \text{Geo}(p)$. Then

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = p\sum_{k=1}^{\infty} \frac{d}{dx}\,(1-x)^k\Big|_{x=p} = p\frac{d}{dx}\sum_{k=1}^{\infty}(1-x)^k\Bigg|_{x=p}$$
$$= p\frac{d}{dx}\,\frac{1-x}{1-(1-x)}\Bigg|_{x=p} = p\frac{d}{dx}\,\frac{1-x}{x}\Bigg|_{x=p} = p\frac{d}{dx}\left(1-\frac{1}{x}\right)\Bigg|_{x=p} = p\cdot\frac{1}{p^2} = \frac{1}{p}.$$

The exchange of the sum and derivative is justified since $0 < p < 1$, so we are in the region of uniform convergence of the power series.

2.4 Moments

Recall that by the law of the unconscious statistician, we have $\mathbb{E}[X^2] = \sum_{x\in\mathcal{R}(X)} x^2 p_X(x)$. More generally,

$$\mathbb{E}[X^k] = \sum_{x\in\mathcal{R}(X)} x^k p_X(x)$$

for $k \geq 1$, provided this series converges absolutely.

Definition 2.4. For a random variable X,

- $\mathbb{E}[X^k]$ is called the *moment of order k* of X,
- $\mathbb{E}[|X|^k]$ is called the *absolute moment of order k* of X,
- $\mathbb{E}[(X-\mathbb{E}[X])^k]$ is called the *centered moment of order k* of X.
- and $\mathbb{E}[|X-\mathbb{E}[X]|^k]$ is called the *centered absolute moment of order k* of X.

2.5 Variance

Definition 2.5. Let X be a random variable with finite 2nd moment, i.e. we have $\mathbb{E}[X^2] < \infty$. Then the *variance* of X is given by

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

Example 2.5.1. Define the random variables $X = 0$,

$$Y = \begin{cases} 1 & \text{with probability } 1/2, \\ -1 & \text{with probability } 1/2, \end{cases} \quad \text{and} \quad Z = \begin{cases} 10 & \text{with probability } 1/2, \\ -10 & \text{with probability } 1/2. \end{cases}$$

Then $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z] = 0$. But observe that we have

$$\text{Var}[X] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] = 0,$$
$$\text{Var}[Y] = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[Y^2] = 1,$$
$$\text{Var}[Z] = \mathbb{E}[(Z-\mathbb{E}[Z])^2] = \mathbb{E}[Z^2] = 100,$$

which are not the same.

Remark. The variance is a measure of the spread of a random variable about its mean.

Definition 2.6. The positive square root of $\mathrm{Var}[X]$ is called the *standard deviation* of X.

Proposition 2.4. *We have the following properties of variance:*

(i) $\mathrm{Var}[\alpha X] = \alpha^2 \mathrm{Var}[X]$ *for all* $\alpha \in \mathbb{R}$,

(ii) $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$,

(iii) $\mathrm{Var}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2$,

(iv) and $\mathrm{Var}[X] = 0$ *if and only if* X *is constant.*

Proof. (i) We can compute that

$$\mathrm{Var}[\alpha X] = \mathbb{E}[(\alpha X - \mathbb{E}[\alpha X])^2] = \mathbb{E}[\alpha^2 (X - \mathbb{E}[X])^2] = \alpha^2 \mathbb{E}[(X - \mathbb{E}[X])^2] = \alpha^2 \mathrm{Var}[X],$$

by the linearity of expectation.

(ii) Again by the linearity of expectation, we have

$$\begin{aligned} \mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[(\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \end{aligned}$$

since $\mathbb{E}[X]$ and $(\mathbb{E}[X])^2$ are constants.

(iii) Simply write $\mathbb{E}[X(X-1)] = \mathbb{E}[X^2 - X] = \mathbb{E}[X^2] - \mathbb{E}[X]$ and apply (ii).

(iv) $(\Leftarrow)$ If X is constant, then $X = \mathbb{E}[X]$, so

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(\mathbb{E}[X] - \mathbb{E}[X])^2] = \mathbb{E}[0] = 0.$$

$(\Rightarrow)$ Suppose $\mathrm{Var}[X] = 0$. Then we find that

$$0 = \mathrm{Var}[X] = \sum_{x \in \mathcal{R}(X)} (x - \mathbb{E}[X])^2 p_X(x).$$

This is a sum of nonnegative terms, so each term must be zero, i.e. $(x - \mathbb{E}[X])^2 p_X(x) = 0$. Since $x \in \mathcal{R}(X)$, we must have $p_X(x) > 0$, and thus $(x - \mathbb{E}[X])^2 = 0$. This gives $x = \mathbb{E}[X]$ for every $x \in \mathcal{R}(X)$, so we conclude that $X = \mathbb{E}[X]$ must be constant. □

Exercise 2.1. Compute the variance for the following random variables:

(i) $X \sim \mathrm{Ber}(p)$. The answer should be $\mathrm{Var}[X] = p(1-p)$.

(ii) $X \sim \mathrm{Bin}(n, p)$. The answer should be $\mathrm{Var}[X] = np(1-p)$.

(iii) $X \sim \mathrm{Poi}(\lambda)$. The answer should be $\mathrm{Var}[X] = \lambda$.

(iv) $X \sim \mathrm{Geo}(p)$. We know $\mathbb{E}[X] = 1/p$, what is $\mathrm{Var}[X]$?

(v) $X \sim \mathrm{NBin}(r, p)$. We know $\mathbb{E}[X] = r/p$, what is $\mathrm{Var}[X]$?

2.6 Conditional Expectation

Definition 2.7. Let X be a random variable and let B be an event such that $\mathbb{P}(B) > 0$. Then the *conditional expectation* of X given B is defined by

$$\mathbb{E}[X|B] = \sum_{x \in \mathcal{R}(X)} x\mathbb{P}(X = x|B) = \sum_{x \in \mathcal{R}(X)} \frac{x\mathbb{P}(\{X = x\} \cap B)}{\mathbb{P}(B)}.$$

Remark. Recall that $\{B_i\}_{i=1}^{\infty}$ is a partition of Ω if the B_i are pairwise disjoint events and $\Omega = \bigcup_{i=1}^{\infty} B_i$.

Theorem 2.2 (Partition theorem in expectation)**.** *Let X be a discrete random variable and $\{B_k\}_{k=1}^{\infty}$ be a partition of Ω with $\mathbb{P}(B_k) > 0$ for each $k \geq 1$. Then*

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{E}[X|B_k]\mathbb{P}(B_k).$$

Proof. Use the definition of conditional expectation to write

$$\begin{aligned}\sum_{k=1}^{\infty} \mathbb{E}[X|B_k]\mathbb{P}(B_k) = \sum_{k=1}^{\infty} \sum_{x \in \mathcal{R}(X)} x\mathbb{P}(X = x|B_k)\mathbb{P}(B_k) &= \sum_{x \in \mathcal{R}(X)} x \sum_{k=1}^{\infty} \mathbb{P}(X = x|B_k)\mathbb{P}(B_k) \\ &= \sum_{x \in \mathcal{R}(X)} x\mathbb{P}(X = x) = \mathbb{E}[X]\end{aligned}$$

by the usual partition theorem. Exchanging the sums is permissible by absolute convergence of $\mathbb{E}[X]$. □

Remark. One can see this as saying "the expectation of the conditional expectation is the expectation."

Example 2.7.1. Suppose a coin flips heads with probability p and tails with probability $1 - p$. What is the expected length of the initial run (of consecutive heads if the first flip is heads, or of consecutive tails if the first flip is tails)? Let X be the length of the initial run and H be the event that the first flip is heads. Then

$$\mathbb{P}(X = k|H) = p^{k-1}(1 - p) \quad \text{for } k = 1, 2, \ldots.$$

Similarly we find

$$\mathbb{P}(X = k|H^c) = (1 - p)^{k-1}p \quad \text{for } k = 1, 2, \ldots.$$

Since $\{H, H^c\}$ is a partition of Ω, we can use the partition theorem in expectation to write

$$\mathbb{E}[X] = \mathbb{E}[X|H]\mathbb{P}(H) + \mathbb{E}[X|H^c]\mathbb{P}(H^c). \tag{$*$}$$

We can compute that

$$\mathbb{E}[X|H] = \sum_{x \in \mathcal{R}(X)} x\mathbb{P}(X = x|H) = \sum_{k=1}^{\infty} kp^{k-1}(1 - p) = (1 - p)\sum_{k=1}^{\infty} kp^{k-1} = \frac{1 - p}{(1 - p)^2} = \frac{1}{1 - p}.$$

Similarly we can find

$$\mathbb{E}[X|H^c] = \sum_{x \in \mathcal{R}(X)} x\mathbb{P}(X = x|H^c) = \sum_{k=1}^{\infty} k(1 - p)^{k-1}p = p\sum_{k=1}^{\infty} k(1 - p)^{k-1} = \frac{p}{p^2} = \frac{1}{p}.$$

Thus by substituting these values into $(*)$ we obtain

$$\mathbb{E}[X] = \frac{1}{1-p} \cdot p + \frac{1}{p} \cdot (1-p) = \frac{1}{p(1-p)} - 2.$$

In particular, if $p = 1/2$, then we find that $\mathbb{E}[X] = 2$.

2.7 Homework Problems

Problems #1, 2, 4, 5, 6, 7, 9, 10 from Grimmett and Welsh.

Chapter 3

Multivariate Discrete Random Variables

3.1 Discrete Random Vectors

Definition 3.1. A *random vector* is a function from Ω to $\mathbb{R}^d$, where $d \geq 2$. We say that the random vector is *bivariate* if $d = 2$, *trivariate* if $d = 3$, etc.

Definition 3.2. A random vector is said to be *discrete* if its range is at most countably infinite.

Example 3.2.1. Let $d = 2$ and X be a 2-dimensional random vector. Then $X : \Omega \to \mathbb{R}^2$ is given by

$$\omega \mapsto X(\omega) = (X_1(\omega), X_2(\omega)).$$

In particular, each coordinate X_1 and X_2 is a function from Ω to $\mathbb{R}$ and is thus itself a random variable.

Proposition 3.1. *A function $X : \Omega \to \mathbb{R}^d$ with $d \geq 2$ is a random vector if and only if each of its coordinates $X_1, X_2, \ldots, X_d$ are random variables.*

Proof. Most of this is immediate. More justification is necessary to ensure measurability (i.e. preimages of points are events), but that is the subject of a later course in probability. □

Proposition 3.2. *A random vector $X : \Omega \to \mathbb{R}^d$ is discrete if and only if each of its component random variables $X_1, X_2, \ldots, X_d$ are discrete.*

Proof. ($\Rightarrow$) Observe that there is a surjection $\mathcal{R}(X) \to \mathcal{R}(X_i)$ by projecting onto the ith coordinate, so $\mathcal{R}(X_i)$ can only be at most countable since $\mathcal{R}(X)$ is.

($\Leftarrow$) Notice that $\mathcal{R}(X) \subseteq \mathcal{R}(X_1) \times \mathcal{R}(X_2) \times \cdots \times \mathcal{R}(X_d)$. Finite products of countable sets are countable, so $\mathcal{R}(X)$ is a subset of a countable set and thus countable. □

Definition 3.3. Let $X : \Omega \to \mathbb{R}^d$ be a discrete random vector. The *probability mass function* of X is

$$p_X(x_1, x_2, \ldots, x_d) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d)$$

for all $(x_1, x_2, \ldots, x_d) \in \mathcal{R}(X)$. This probability mass function must satisfy:

(i) $p_x(x_1, \ldots, x_d) \geq 0$,

(ii) $\displaystyle\sum_{x_1 \in \mathcal{R}(X_1)} \sum_{x_2 \in \mathcal{R}(X_2)} \cdots \sum_{x_d \in \mathcal{R}(X_d)} p_X(x_1, \ldots, x_d) = 1$,

(iii) and for all $A \subseteq \mathbb{R}^d$,

$$\mathbb{P}(X \in A) = \sum_{x \in A \cap \mathcal{R}(X)} p_X(x_1, \ldots, x_d).$$

Note that $A \cap \mathcal{R}(X)$ is countable even when A might not be.

3.2 Marginal Distributions

Definition 3.4. The *one-dimensional marginal pmf* p_{X_k} of a discrete random vector

$$X = (X_1, \ldots, X_d) : \Omega \to \mathbb{R}^d$$

with joint pmf $p_X(x_1, \ldots, x_d)$ is given by

$$p_{X_k}(x) = \sum_{x_1 \in \mathcal{R}(X_1)} \cdots \sum_{x_{k-1} \in \mathcal{R}(X_{k-1})} \sum_{x_{k+1} \in \mathcal{R}(X_{k+1})} \cdots \sum_{x_d \in \mathcal{R}(X_d)} p_X(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots x_d).$$

One can similarly define an *n-dimensional marginal pmf* by summing over all but n of the variables. Note that there are a total of $\binom{d}{n}$ n-dimensional marginal pmfs for X.

Remark. Note that we indeed have

$$\sum_{x \in \mathcal{R}(X_k)} p_{X_k}(x) = \sum_{x_1 \in \mathcal{R}(X_1)} \cdots \sum_{x_d \in \mathcal{R}(X_d)} p_X(x_1, \ldots, x_d) = 1$$

since we can sum in any order by absolute convergence. A similar thing works for the other marginals.

3.3 Revisiting Expectation

Definition 3.5. Let $h : \mathbb{R}^d \to \mathbb{R}$ and X be a d-dimensional discrete random vector. Then the *expectation* of $h(X)$ is

$$\mathbb{E}[h(X)] = \sum_{x_1 \in \mathcal{R}(X_1)} \cdots \sum_{x_d \in \mathcal{R}(X_d)} h(x_1, \ldots, x_d) p_X(x_1, \ldots, x_d),$$

provided that this sum exists and converges absolutely. Here p_X is the joint pmf of X.

Remark. Observe that we have:

$$\begin{array}{ccc} \Omega & \xrightarrow{X} & \mathbb{R}^d \\ & \searrow^{h(X)} & \downarrow h \\ & & \mathbb{R} \end{array}$$

In particular, we see that $h(X)$ is a random variable.

Proposition 3.3. *Let X_1, X_2 be two discrete random variables and let $a, b \in \mathbb{R}$. Then*

$$\mathbb{E}[aX_1 + bX_2] = a\mathbb{E}[X_1] + b\mathbb{E}[X_2],$$

provided the expectations exist.

Proof. Note that (X_1, X_2) is a random vector and thus has a joint pmf $p_{(X_1,X_2)}$. Define $h : \mathbb{R}^2 \to \mathbb{R}$ by

$$(x_1, x_2) \mapsto h(x_1, x_2) = ax_1 + bx_2.$$

Then we find that

$$\begin{aligned}
\mathbb{E}[h(X_1, X_2)] &= \sum_{x_1 \in \mathcal{R}(X_1)} \sum_{x_2 \in \mathcal{R}(X_2)} h(x_1, x_2) p_{(X_1,X_2)}(x_1, x_2) \\
&= \sum_{x_1 \in \mathcal{R}(X_1)} \sum_{x_2 \in \mathcal{R}(X_2)} (ax_1 + bx_2) p_{(X_1,X_2)}(x_1, x_2) \\
&= \sum_{x_1 \in \mathcal{R}(X_1)} \sum_{x_2 \in \mathcal{R}(X_2)} ax_1 p_{(X_1,X_2)}(x_1, x_2) + \sum_{x_1 \in \mathcal{R}(X_1)} \sum_{x_2 \in \mathcal{R}(X_2)} bx_2 p_{(X_1,X_2)}(x_1, x_2) \\
&= a \sum_{x_1 \in \mathcal{R}(X_1)} x_1 \sum_{x_2 \in \mathcal{R}(X_2)} p_{(X_1,X_2)}(x_1, x_2) + b \sum_{x_1 \in \mathcal{R}(X_1)} x_2 \sum_{x_2 \in \mathcal{R}(X_2)} p_{(X_1,X_2)}(x_1, x_2) \\
&= a \sum_{x_1 \in \mathcal{R}(X_1)} x_1 p_{X_1}(x_1) + b \sum_{x_2 \in \mathcal{R}(X_2)} x_2 p_{X_2}(x_2) = a\mathbb{E}[X_1] + b\mathbb{E}[X_2],
\end{aligned}$$

which is the desired result. Manipulating the sums above is justified by absolute convergence. □

3.4 Independence of Random Variables

Definition 3.6. Two discrete random variables X and Y are said to be *independent*, written $X \perp\!\!\!\perp Y$, if for any $x \in \mathcal{R}(X)$ and $y \in \mathcal{R}(Y)$, the events $\{X = x\}$ and $\{Y = y\}$ are independent, i.e.

$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

Remark. We will often write $\mathbb{P}(X = x, Y = y)$ instead of $\mathbb{P}(\{X = x\} \cap \{Y = y\})$.

Theorem 3.1. *The discrete random variables X and Y are independent if and only if there exist functions $f, g : \mathbb{R} \to \mathbb{R}$ such that*

$$p_{(X,Y)}(x, y) = f(x)g(y)$$

for all $x, y \in \mathbb{R}$. Here $p_{(X,Y)}$ is the joint pmf of (X, Y).

Proof. ($\Leftarrow$) Assume that $p_{(X,Y)}(x, y) = f(x)g(y)$ for all $x, y \in \mathbb{R}$. Then

$$p_X(x) = \sum_{y \in \mathcal{R}(Y)} p_{(X,Y)}(x, y) = \sum_{y \in \mathcal{R}(Y)} f(x)g(y) = f(x) \sum_{y \in \mathcal{R}(Y)} g(y),$$

and similarly by symmetry we find that

$$p_Y(x) = \sum_{x \in \mathcal{R}(X)} p_{(X,Y)}(x, y) = \sum_{x \in \mathcal{R}(X)} f(x)g(y) = g(y) \sum_{x \in \mathcal{R}(X)} f(x).$$

But $p_{(X,Y)}$ is the joint pmf, so we must have

$$1 = \sum_{x \in \mathcal{R}(X)} \sum_{y \in \mathcal{R}(Y)} f(x)g(y) = \sum_{x \in \mathcal{R}(X)} f(x) \sum_{y \in \mathcal{R}(Y)} g(y).$$

But then we can use this to write

$$p_{(X,Y)}(x,y) = f(x)g(y) = f(x)g(y)\sum_{x\in\mathcal{R}(X)} f(x)\sum_{y\in\mathcal{R}(Y)} g(y)$$
$$= \left(f(x)\sum_{y\in\mathcal{R}(Y)} g(y)\right)\left(g(y)\sum_{x\in\mathcal{R}(X)} f(x)\right) = p_X(x)p_Y(y),$$

where the last line follows from the previous computations. This is the desired result.

($\Leftarrow$) This is clear. By independence $p_{(X,Y)}(x,y) = p_X(x)p_Y(y)$, so we can set $f = p_X$ and $g = p_Y$. □

Proposition 3.4. *Let X and Y be two independent discrete random variables. Then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. We can write

$$\mathbb{E}[XY] = \sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} xyp_{(X,Y)}(x,y) = \sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} xyp_X(x)p_Y(y)$$
$$= \left(\sum_{x\in\mathcal{R}(X)} xp_X(x)\right)\left(\sum_{y\in\mathcal{R}(Y)} yp_Y(y)\right) = \mathbb{E}[X]\mathbb{E}[Y]$$

where the second step follows by independence. □

Remark. More generally, the same argument shows that if X, Y are independent, then

$$\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$$

for any functions $h_1, h_2 : \mathbb{R} \to \mathbb{R}$.

Definition 3.7. Let X and Y be two random variables. Then

- X, Y are *uncorrelated* if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, i.e. $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$,
- X, Y are *positively correlated* if $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] > 0$,
- and X, Y are *negatively correlated* if $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] < 0$.

Remark. The previous result shows that if X, Y are independent, then they are uncorrelated.

Example 3.7.1. However, the converse is not true in general. Let X take the values $-1, 0, 1$ with probability $1/3$. Clearly $\mathbb{E}[X] = 0$. Now set

$$Y = \begin{cases} 0 & \text{if } X = 0, \\ 1 & \text{if } X \neq 1. \end{cases}$$

First observe that X and Y are dependent since

$$\mathbb{P}(X = 0, Y = 1) = 0 \neq \frac{1}{3}\cdot\frac{2}{3} = \mathbb{P}(X = 0)\mathbb{P}(Y = 1).$$

Now since $\mathbb{E}[X] = 0$, we clearly have $\mathbb{E}[X]\mathbb{E}[Y] = 0$. Also we can compute that

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{x\in\mathcal{R}(X)}\sum_{y\in\mathcal{R}(Y)} xy\mathbb{P}(X=x, Y=y)\\ &= -1\cdot 1 p_{(X,Y)}(-1,1) + 1\cdot 1 p_{(X,Y)}(1,1) = -\frac{1}{3}+\frac{1}{3} = 0.\end{aligned}$$

In particular, this means that $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 = 0$, so X and Y are uncorrelated. However, as we previously computed, X and Y are not independent.

Remark. If we instead demand that $\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$ for all functions $h_1, h_2 : \mathbb{R}\to\mathbb{R}$, then we are indeed able to conclude that X and Y are independent.

Proposition 3.5. *If $\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$ for all functions $h_1, h_2 : \mathbb{R}\to\mathbb{R}$, then X and Y are independent.*

Proof. For each $x_1 \in \mathcal{R}(X)$ and $y_1 \in \mathcal{R}(Y)$, choose $h_1(x) = \mathbb{1}_{\{x_1\}}$ and $h_2(y) = \mathbb{1}_{\{y_1\}}$. Then

$$\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[\mathbb{1}_{\{x_1\}}(X)\mathbb{1}_{\{y_1\}}(Y)] = \mathbb{P}(X=x_1, Y=y_1)$$

and also

$$\mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)] = \mathbb{E}[\mathbb{1}_{\{x_1\}}(X)]\mathbb{E}[\mathbb{1}_{\{y_1\}}(Y)] = \mathbb{P}(X=x_1)\mathbb{P}(Y=y_1),$$

which implies the desired result since $\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$. □

3.5 Convolution and Random Variables

Definition 3.8. The *convolution* of two pmfs p_X, p_Y is given by

$$(p_X * p_Y)(z) = \sum_{x\in\mathcal{R}(X)} p_X(x)p_Y(z-x) = \sum_{y\in\mathcal{R}(Y)} p_X(z-y)p_Y(y).$$

Proposition 3.6. *Let X and Y be two independent discrete random variables with pmfs p_X and p_Y, respectively. Then $X+Y$ is discrete and has pmf given by the convolution $p_X * p_Y$.*

Proof. It is clear that $X+Y$ is discrete, so it suffices to find its pmf. Let X take the values $x \in \mathcal{R}(X)$ and Y take the values $y \in \mathcal{R}(Y)$. Let $Z = X+Y$ take the values $z \in \mathcal{R}(X+Y)$. Then we find that

$$p_Z(z) = \mathbb{P}(Z=z) = \mathbb{P}(X+Y=z) = \mathbb{P}\left(\bigcup_{x\in\mathcal{R}(X)} (\{X=x\}\cap\{Y=z-x\})\right).$$

These events are pairwise disjoint, so we have

$$\begin{aligned}p_Z(z) &= \sum_{x\in\mathcal{R}(X)} \mathbb{P}(\{X=x\}\cap\{Y=z-x\})\\ &= \sum_{x\in\mathcal{R}(X)} \mathbb{P}(X=x)\mathbb{P}(Y=z-x) = \sum_{x\in\mathcal{R}(X)} p_X(x)p_Y(z-x)\end{aligned}$$

by independence. This is precisely the convolution of p_X and p_Y. □

Proposition 3.7. *Let $X_1, \ldots, X_n$ be n independent discrete random variables with pmfs $p_{X_1}, \ldots, p_{X_n}$. Then $X_1 + \cdots + X_n$ is discrete and has pmf $p_{X_1} * p_{X_2} * \cdots * p_{X_{n-1}} * p_{X_n}$.*

Proof. This follows from induction using the previous result and convolution being associative.[1] □

Example 3.8.1. Let $X \sim \text{Ber}(p)$ and $Y \sim \text{Ber}(p)$, and let X and Y be independent. The pmf of $X+Y$ is the convolution of the pmfs of X and Y. Now $X+Y$ takes the values $z = 0, 1, 2$, and

$$p_{X+Y}(z) = \sum_{x \in \mathcal{R}(X)} p_X(x) p_Y(z-x).$$

For $z = 0$, we have

$$p_{X+Y}(0) = \sum_{x=0}^{1} p_X(x) p_Y(-x) = p_X(0)p_Y(0) + p_X(1)p_Y(-1) = (1-p)(1-p) + 0 = (1-p)^2$$

since $p_Y(-1) = 0$. For $z = 1$, we have

$$p_{X+Y}(1) = \sum_{x=0}^{1} p_X(x) p_Y(1-x) = p_X(0)p_Y(1) + p_X(1)p_Y(0) = (1-p)p + p(1-p) = 2p(1-p).$$

Finally, when $z = 2$, we see that

$$p_{X+Y}(2) = \sum_{x=0}^{1} p_X(x) p_Y(2-x) = p_X(0)p_Y(2) + p_X(1)p_Y(1) = 0 + p(p) = p^2$$

since $p_Y(2) = 0$. In particular, this shows us that $X + Y \sim \text{Bin}(2, p)$. Induct to get the general case.

3.6 Indicator Functions

Definition 3.9. The *indicator function* of a set A is

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

Remark. We have the following properties of indicator functions:

(i) $\mathbb{1}_{A \cap B} = \mathbb{1}_A \mathbb{1}_B$. To see this, we can write

$$(\mathbb{1}_A \mathbb{1}_B)(x) = \mathbb{1}_A(x)\mathbb{1}_B(x) = \begin{cases} 1 & \text{if } x \in A \text{ and } x \in B, \\ 0 & \text{otherwise.} \end{cases}$$

(ii) $\mathbb{1}_A + \mathbb{1}_{A^c} = 1$.

(iii) $\mathbb{1}_{A \cup B} = 1 - \mathbb{1}_{A^c \cap B^c}$.

[1] One way to see that convolution is associative is to note that the addition of random variables is associative.

(iv) $\mathbb{1}_{A\Delta B} = \mathbb{1}_A + \mathbb{1}_B \pmod 2$. Here $A\Delta B = (A \setminus B) \cup (B \setminus A)$ is the *symmetric difference* of A and B.

Remark. Observe that $\mathbb{1}_A$ is a function $\mathbb{1}_A : \Omega \to \mathbb{R}$. In particular, if A is an event, then $\mathbb{1}_A$ is a random variable and

$$\mathbb{E}[\mathbb{1}_A] = 0 \cdot \mathbb{P}(\mathbb{1}_A = 0) + 1 \cdot \mathbb{P}(\mathbb{1}_A = 1) = \mathbb{P}(\mathbb{1}_A = 1) = \mathbb{P}(A).$$

Using this, we can take expectations on property (ii) above to get

$$\mathbb{E}[\mathbb{1}_A + \mathbb{1}_{A^c}] = \mathbb{E}[\mathbb{1}_A] + \mathbb{E}[\mathbb{1}_{A^c} = \mathbb{P}(A) + \mathbb{P}(A^c) = 1.$$

Now observe that property (iii) says that

$$\begin{aligned}\mathbb{1}_{A\cup B} &= 1 - \mathbb{1}_{A^c \cap B^c} = 1 - \mathbb{1}_{A^c}\mathbb{1}_{B^c} = 1 - (1 - \mathbb{1}_A)(1 - \mathbb{1}_B) \\ &= 1 - (1 - \mathbb{1}_A - \mathbb{1}_B + \mathbb{1}_A\mathbb{1}_B) = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A\mathbb{1}_B.\end{aligned}$$

From here taking expectations gives

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{E}[\mathbb{1}_{A\cup B}] = \mathbb{E}[\mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A\mathbb{1}_B] \\ &= \mathbb{E}[\mathbb{1}_A] + \mathbb{E}[\mathbb{1}_B] - \mathbb{E}[\mathbb{1}_A\mathbb{1}_B] = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).\end{aligned}$$

More generally, let $A_1, \ldots, A_n$ be arbitrary events with indicators $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_n}$. Let $A = \bigcup_{i=1}^n A_i$. Then

$$\begin{aligned}\mathbb{1}_A = \mathbb{1}_{\bigcup_{i=1}^n A_i} &= 1 - \mathbb{1}_{\bigcap_{i=1}^n A_i^c} = 1 - \prod_{i=1}^n \mathbb{1}_{A_i^c} = 1 - \prod_{i=1}^n (1 - \mathbb{1}_{A_i}) \\ &= \sum_{i=1}^n \mathbb{1}_{A_i} - \sum_{1\le i<j\le n} \mathbb{1}_{A_i}\mathbb{1}_{A_j} + \sum_{1\le i<j<k\le n} \mathbb{1}_{A_i}\mathbb{1}_{A_j}\mathbb{1}_{A_k} - \cdots + (-1)^{n+1}\mathbb{1}_{A_1}\ldots\mathbb{1}_{A_n}.\end{aligned}$$

At this point, taking expectations precisely recovers the general inclusion-exclusion formula.

3.7 Homework Problems

Problems #2, 3, 5, 7, 12, 13, 14 from Grimmett and Welsh.

Chapter 4

Probability Generating Functions

4.1 Probability Generating Functions

The probability generating function is a concept for discrete random variables, and more precisely for those X taking values in $\{0, 1, 2, \dots\}$. Such random variables have a pmfs $p_X(k) = p_k = \mathbb{P}(X = k)$ for $k = 0, 1, 2, \dots$. Then $\{p_k\}_{k\geq 0} \subseteq \mathbb{R}$ is a sequence such that $p_k \geq 0$ for each k and $\sum_{k=0}^{\infty} p_k = 1$.

Definition 4.1. Let X be a discrete random variable taking values in $\{0, 1, 2, \dots\}$ with pmf given by $\mathbb{P}(X = k) = p_k$ for $k = 0, 1, 2, \dots$. Then the *probability generating function (pgf)* of X is given by

$$G_X(s) = p_0 + p_1 s + p_2 s^2 + \cdots = \mathbb{E}[s^X],$$

for the values of s for which $\mathbb{E}[|s^X|] = \sum_{k=0}^{\infty} p_k |s|^k < \infty$.

Proposition 4.1. *The series $G_X(s) = \sum_{k=0}^{\infty} p_k s^k$ converges absolutely for $s \in \mathbb{R}$ where $|s| \leq 1$.*

Proof. Observe that $G_X(s) = \mathbb{E}[s^X]$ and for $|s| \leq 1$,

$$\mathbb{E}[|s^X|] = \sum_{k=0}^{\infty} p_k |s|^k \leq \sum_{k=0}^{\infty} p_k = 1$$

since $p_k = p_X(k)$ is a pmf. Thus the interval of convergence is guaranteed to contain $[-1, 1]$. □

Example 4.1.1. Consider the following probability generating functions:

- Let $X \sim \text{Ber}(p)$. Then

  $$G_X(s) = s^0 \mathbb{P}(X = 0) + s^1 \mathbb{P}(X = 1) = 1(1 - p) + ps = 1 - p + ps.$$

  This exists for all $s \in \mathbb{R}$.

- Let $X \sim \text{Bin}(n, p)$. Then

  $$G_X(s) = \sum_{k=0}^{n} s^k \binom{n}{k} p^k (1 - p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (ps)^k (1 - p)^{n-k} = (1 - p + ps)^n.$$

  This again exists for all $s \in \mathbb{R}$.

- Let $X \sim \text{Geo}(p)$. Then

$$G_X(s) = \sum_{k=1}^{\infty} p(1-p)^{k-1}s^k = ps\sum_{k=1}^{\infty}(1-p)^{k-1}s^{k-1} = ps\sum_{k=0}^{\infty}((1-p)s)^k.$$

  When $|(1-p)s| < 1$, we get

$$G_X(s) = \frac{ps}{1-(1-p)s}$$

  Thus G_X exists for $|s| < 1/(1-p)$.

- Let $X \sim \text{Poi}(\lambda)$. Then

$$G_X(s) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!}s^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda s)^k}{k!} = e^{-\lambda}e^{\lambda s} = e^{\lambda(s-1)}.$$

  This series converges for all $s \in \mathbb{R}$.

4.2 Properties of PGFs

Theorem 4.1. *Let X and Y have probability generating functions G_X and G_Y, respectively. Then*

$$G_X(s) = G_Y(s) \quad \textit{for all } s \textit{ in their interval of convergence}$$

if and only if $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for all $k = 0, 1, 2, \ldots$.

Proof. ($\Leftarrow$) If $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for each k, then

$$G_X(s) = \sum_{k=0}^{\infty}\mathbb{P}(X = k)s^k = \sum_{k=0}^{\infty}\mathbb{P}(Y = k)s^k = G_Y(s).$$

This holds for every s for which the above series converge.

($\Rightarrow$) Assume that $G_X(s) = G_Y(s)$ for all $|s| \leq 1$. First observe that plugging in $s = 0$ in $G_X(s)$ and $G_Y(s)$ immediately gives $\mathbb{P}(X = 0) = \mathbb{P}(Y = 0)$. Then because we have absolute convergence for $|s| \leq 1$ and thus uniform convergence on $|s| \leq 1$, we can differentiate term by term to get

$$\frac{d}{ds}G_X(s) = \frac{d}{ds}\sum_{k=0}^{\infty}\mathbb{P}(X = k)s^k = \sum_{k=0}^{\infty}\frac{d}{ds}\left[\mathbb{P}(X = k)s^k\right] = \sum_{k=0}^{\infty}\mathbb{P}(X = k)ks^{k-1}$$

and

$$\frac{d}{ds}G_Y(s) = \frac{d}{ds}\sum_{k=0}^{\infty}\mathbb{P}(Y = k)s^k = \sum_{k=0}^{\infty}\frac{d}{ds}\left[\mathbb{P}(Y = k)s^k\right] = \sum_{k=0}^{\infty}\mathbb{P}(Y = k)ks^{k-1}.$$

These are equal since $G_X(s) = G_Y(s)$, so we can again evaluate at $s = 0$ to get

$$1 \cdot \mathbb{P}(X = 1) = 1 \cdot \mathbb{P}(Y = 1),$$

i.e. $\mathbb{P}(X = 1) = \mathbb{P}(Y = 1)$. Continue by induction to get $k! \cdot \mathbb{P}(X = k) = k! \cdot \mathbb{P}(Y = k)$ for every $k \geq 1$, so $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for each k. This is the desired result. □

Theorem 4.2. *Let X be an integer-valued random variable with pgf G_X. Then for every $\ell \geq 1$,*

$$\frac{d^\ell}{ds^\ell} G_X(s)\bigg|_{s=1} = \mathbb{E}[X(X-1)\dots(X-\ell+1)]$$

Proof. As before,

$$\frac{d^\ell}{ds^\ell} G_X(s) = \sum_{k=0}^{\infty} k(k-1)\dots(k-\ell+1)\mathbb{P}(X=k)s^{k-\ell}$$

for all $|s| \leq 1$. Evaluating at $s = 1$ gives

$$\frac{d^\ell}{ds^\ell} G_X(s) = \sum_{k=\ell}^{\infty} k(k-1)\dots(k-\ell+1)\mathbb{P}(X=k) = \mathbb{E}[X(X-1)\dots(X-\ell+1)],$$

as desired. In particular, this expectation is finite if and only if the series on the left converges. □

Remark. Recall from before that $\mathrm{Var}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2$. We can compute this via

$$\mathrm{Var}[X] = G_X''(1) + G_X'(1) - (G_X'(1))^2$$

using the probability generating function if X is integer-valued.

Remark. The value $\mathbb{E}[X(X-1)\dots(X-\ell+1)]$ is called the *factorial moment of order ℓ* of X.

Theorem 4.3. *Let X, Y be independent random variables. Then $G_{X+Y}(s) = G_X(s)G_Y(s)$ for all $|s| \leq 1$.*

Proof. We have $G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X s^Y] = \mathbb{E}[s^X]\mathbb{E}[s^Y] = G_X(s)G_Y(s)$ by independence. □

Remark. Another way to show this is to use the convolution formula for the pmf of $X + Y$.

Corollary 4.3.1. *For independent random variables $X_1, \dots, X_n$, we have*

$$G_{X_1+\dots+X_n}(s) = \prod_{k=1}^{n} G_{X_k}(s)$$

Remark. We actually also have the converse: If $G_{X+Y}(s) = G_X(s)G_Y(s)$ for all $|s| \leq 1$, then X and Y are independent. One can show this using the convolution formula.

4.3 The Random Sum Formula

Let $S = X_1 + \dots + X_N$, where $X_1, X_2, \dots$ are independent random variables and N is an integer-valued random variable independent of the X_i. If $N = 0$, then the sum is empty and by convention we say that $X_1 + \dots + X_N = 0$. If $N \geq 1$, then we define

$$(X_1 + \dots + X_N)(\omega) = X_1(\omega) + X_2(\omega) + \dots + X_{N(\omega)}(\omega).$$

Theorem 4.4 (Random sum formula)**.** *For integer-valued independent random variables $N, X_1, X_2, \dots$,*

$$G_{X_1+\dots+X_N}(s) = G_N(G_X(s)),$$

given that $X_1, X_2, \dots$ are identically distributed with $p_X = p_{X_1} = p_{X_2} = \dots$.

Proof. Since the events $\{N = n\}$ partition the sample space Ω, by the partition theorem we can write

$$G_{X_1+\cdots+X_N}(s) = \mathbb{E}[s^{X_1+\cdots+X_N}] = \sum_{n=0}^{\infty} \mathbb{E}[s^{X_1+\cdots+X_N}|N = n]\mathbb{P}(N = n)$$
$$= \sum_{n=0}^{\infty} \mathbb{E}[s^{X_1+\cdots+X_n}]\mathbb{P}(N = n).$$

Now by independence of the X_i, we can split the expectation and get

$$G_{X_1+\cdots+X_N}(s) = \sum_{n=0}^{\infty} G_{X_1}(s)\ldots G_{X_n}(s)\mathbb{P}(N = n) = \sum_{n=0}^{\infty}(G_X(s)^n)\mathbb{P}(N = n)$$

since the X_i are identically distributed. Now the value on the right is precisely $G_N(G_X(s))$. □

Remark. We will use "i.i.d." to denote "independent and identically distributed".

Remark. Set $S = X_1 + \cdots + X_N$, so that $G_S(s) = G_N(G_X(s))$. Now differentiating $G_S(s)$ gives

$$G'_S(s) = G'_N(G_X(s))G'_X(s).$$

At $s = 1$, we get $G'_S(1) = G'_N(G_X(1))G'_X(1) = G'_N(1)G'_X(1)$. But $G'_Y(1) = \mathbb{E}[Y]$, so this says that

$$\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[X]$$

One can continue this process. For instance, differentiating again yields

$$G''_S(s) = \frac{d}{ds}\left[G'_N(G_X(s))G'_X(s)\right] = G''_N(G_X(s))G'_X(s)G'_X(s) + G'_N(G_X(s))G''_X(s).$$

Evaluating at $s = 1$, we obtain

$$\mathbb{E}[S(S-1)] = G''_S(1) = G''_N(1)(G'_X(1))^2 + G'_N(1)G''_X(1)$$
$$= \mathbb{E}[N(N-1)](\mathbb{E}[X])^2 + \mathbb{E}[N]\mathbb{E}[X(X-1)].$$

This allows us to find $\mathrm{Var}[S]$ with some more work.

Example 4.1.2. Suppose that we have 20 rabbits, and each has probability 1/2 of escaping. The next morning, each remaining rabbit gives birth to a number of offspring following a Poisson distribution with parameter $\lambda = 3$, and we want to determine the pgf of the number of offspring. To do this, let N be the number of rabbits not escaping, and observe that $N \sim \mathrm{Bin}(20, 1/2)$. Then

$$G_S(s) = G_{X_1+\cdots+X_N}(s) = G_N(G_X(s))$$

by the random sum formula, where $X \sim \mathrm{Poi}(3)$. Then

$$G_N(s) = \left(\frac{1}{2} + \frac{1}{2}s\right)^{20} \quad \text{and} \quad G_X(s) = e^{3(s-1)},$$

so we find that

$$G_S(s) = G_N(G_X(s)) = \left(\frac{1}{2} + \frac{1}{2}e^{3(s-1)}\right)^{20}.$$

In particular, $\mathbb{E}[S] = \mathbb{E}[N]\mathbb{E}[X] = 10(3) = 30$.

4.4 Homework Problems

Problems #2, 5, 6, 8, 9, 10, 11 from Grimmett and Welsh.

Chapter 5

Continuous Random Variables

5.1 Distribution Functions

Remark. When X is discrete, we have the pmf $p_X(x) = \mathbb{P}(X = x)$ for $x \in \mathbb{R}$. Note that $p_X(x) > 0$ for only countably many x, and outside of $\mathcal{R}(X)$ we "set $p_X(x) = 0$" (really p_X is only defined on $\mathcal{R}(X)$). We will now study the situation where we have $p_X(x) = 0$ everywhere.

Definition 5.1. Let X be a random variable. Then the *(cumulative) distribution function (cdf)* of X is

$$F_X(x) = \mathbb{P}(X \leq x) \quad \text{for } x \in \mathbb{R}.$$

Remark. Note that we can write

$$\mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in (-\infty, x]\}) = \mathbb{P}(X^{-1}((-\infty, x])).$$

This hints that $X^{-1}((-\infty, x])$ must be in the event space $\mathcal{F}$.

Example 5.1.1. Suppose that X is discrete. Then

$$F_X(x) = \sum_{y \in \mathcal{R}(X), y \leq x} \mathbb{P}(X = y).$$

- If $X \sim \text{Ber}(p)$, then
$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1. \end{cases}$$

- If $X \sim \text{Bin}(n, p)$, then
$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{\ell=0}^{k-1} \binom{n}{\ell} p^\ell (1-p)^{n-\ell} & \text{if } k - 1 \leq x < k, \text{ for } k = 1, 2, \ldots, n, \\ 1 & \text{if } x \geq n. \end{cases}$$

- If $X \sim \text{Geo}(p)$, then
$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - (1-p)^k & \text{if } k \leq x < k + 1, \text{ for } k = 0, 1, 2, \ldots. \end{cases}$$

This is because for $k \le x < k+1$, we have

$$\mathbb{P}(X \le x) = \sum_{\ell=1}^{k} p(1-p)^{\ell-1} = p\sum_{\ell=0}^{k-1} p(1-p)^{\ell} = p\frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k$$

- If $X \sim \text{Poi}(\lambda)$, then

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \sum_{k=0}^{n} e^{-\lambda}\lambda^k/k! & \text{if } n \le x < n+1, \text{ for } n = 0, 1, 2, \ldots. \end{cases}$$

Proposition 5.1. *Let X be a random variable with distribution function F_X. Then*

(i) F_X is a function $F_X : \mathbb{R} \to [0,1]$,

(ii) F_X is non-decreasing, i.e. if $x \le y$, then $F_X(x) \le F_X(y)$,

(iii) $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$,

(iv) F_X is continuous from the right, i.e. $\lim_{y\to x^+} F_X(Y) = F_X(x)$ for every $x \in \mathbb{R}$,

(v) F_X has a left-hand limit at every $x \in \mathbb{R}$, i.e. $F_X(x^-) = \lim_{y\to x^-} F_X(y)$ exists.

(vi) $\mathbb{P}(X = x) = F_X(x) - F_X(x^-)$,

(vii) and $\mathbb{P}(X \in [a,b]) = F_X(b) - F_X(a^-)$.

Proof. (i) Under F, we have $x \mapsto F(x) = \mathbb{P}(X \le x) \in [0,1]$.

(ii) Observe that $\{X \le x\} \subseteq \{X \le y\}$ if $x \le y$, so $F_X(x) = \mathbb{P}(X \le x) \le \mathbb{P}(X \le y) = F_X(y)$.

(iii) Fix any increasing sequence $\{a_n\}$ with $a_n \to \infty$. Then $\{X \le a_n\}$ is an increasing sequence of events and $\bigcup_{n=1}^{\infty}\{X \le a_n\} = \Omega$ since $a_n \to \infty$. Thus by the continuity from below of probability measures,

$$\lim_{n\to\infty} F_X(a_n) = \lim_{n\to\infty} \mathbb{P}(X \le a_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty}\{X \le a_n\}\right) = \mathbb{P}(\Omega) = 1$$

Since this holds for any increasing sequence $\{a_n\}$, we have $\lim_{x\to\infty} F_X(x) = 1$. We immediately obtain $\lim_{x\to-\infty} F_X(x) = 0$ as well by taking complements in the above argument.

(iv) Take sequences again. Note that $\bigcap_{n=1}^{\infty}\{X \le a_n\} = \{X \le y\}$ if $\{a_n\}$ is decreasing and $a_n \to y$.

(v) We informally write

$$\begin{aligned} F_X(x^-) = \lim_{y\to x^-} F_X(y) = \lim_{y\to x^-} \mathbb{P}(X \le y) &= \lim_{y\to x^-} \mathbb{P}(X \in (-\infty, y]) \\ &= \mathbb{P}(X \in (-\infty, x)) = \mathbb{P}(X < x). \end{aligned}$$

To justify this, one can take sequences and argue as before.

(vi) This is because $F_X(x) - F_X(x^-) = \mathbb{P}(X \le x) - \mathbb{P}(X < x) = \mathbb{P}(X = x)$.

(vii) Write $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \le b) - \mathbb{P}(X < a) = F_X(b) - F_X(a^-)$. □

Remark. Observe that we also have

$$\mathbb{P}(a < X \le b) = \mathbb{P}(X \in (a,b])) = \mathbb{P}(X \le b) - \mathbb{P}(X \le a) = F_X(b) - F_X(a).$$

We get analogous results for the probability of X being in any interval.

5.2 Continuous Random Variables

Definition 5.2. A random variable X is *continuous* if its distribution function F_X is continuous.

Definition 5.3. A random variable X is *absolutely continuous* if F_X is absolutely continuous, i.e.

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

for some function $f_X : \mathbb{R} \to \mathbb{R}$ such that $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

In this case, f_X is called a *(probability) density function (pdf)* of X.

Remark. Note that density functions of X may differ on a set of measure zero, e.g. at a point.

Remark. If X is discrete, then F_X is *not* continuous, and thus certainly not absolutely continuous.

Remark. If X is absolutely continuous with pdf f_X, then

$$\mathbb{P}(X \in (a, b]) = \int_a^b f_X(x)\,dx.$$

In particular, this says that for all $x \in \mathbb{R}$, we must have

$$\int_{\{x\}} f_X(t)\,dt = 0.$$

More generally, if X is continuous, then $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$.

5.3 Common Continuous Random Variables

Example 5.3.1. The following are common continuous random variables:

- *Uniform random variables*: We say that $X \sim \text{Unif}([0, 1])$ if

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

  In general, for $-\infty < a < b < \infty$, we say that $X \sim \text{Unif}([a, b])$ if $(X - a)/(b - a) \sim \text{Unif}([0, 1])$. In this case, X has pdf

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

  Note that f_X is continuous and also differentiable everywhere except at $x = a$ and $x = b$.

- *Exponential random variables*: For $\lambda > 0$, we say that $X \sim \text{Exp}(\lambda)$ if X has density function

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The cdf of X is

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt = \begin{cases} 0 & \text{if } x \le 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

Note that F_X is continuous and also differentiable everywhere except $x = 0$.

Remark. The *fundamental theorem of calculus* says that for a fixed $c \in \mathbb{R}$,

$$\frac{d}{dx}\int_c^x f(t)\,dt = f(x).$$

When applied to distribution functions, this says that

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\int_{-\infty}^{x} f_X(t)\,dt = f_X(x),$$

for almost all $x \in \mathbb{R}$ (whenever F_X is differentiable at x).

Example 5.3.2. A random variable X is said to be *normal* (or *Gaussian*) with parameters $\mu \in \mathbb{R}$ and σ^2 such that $\sigma > 0$, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is given by

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

for $-\infty < x < \infty$. When $\mu = 0$, $\sigma^2 = 1$, then X is said to be the *standard normal* distribution, i.e.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

To verify that this is a density, first note that clearly $f_X(x) \ge 0$. Now set $I = \displaystyle\int_{-\infty}^{\infty} e^{-x^2/2}\,dx$. Then

$$I^2 = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx \int_{-\infty}^{\infty} e^{-y^2/2}\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dx\,dy.$$

This step is justified by the Fubini-Tonelli theorem since $e^{-(x^2+y^2)/2} \ge 0$. Then change to polar coordinates (with $(x, y) = (r\cos\theta, r\sin\theta)$ and $dxdy = r\,drd\theta$) to get

$$I^2 = \int_0^{\infty}\int_0^{2\pi} re^{-r^2/2}\,dr\,d\theta = 2\pi\int_0^{\infty} re^{-r^2/2}\,dr. = 2\pi\left[e^{-r^2/2}\right]_{r=0}^{r=\infty} = 2\pi.$$

Thus $I = \sqrt{2\pi}$ (since $I \ge 0$), so $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = I/\sqrt{2\pi} = 1$. Thus f_X is a density. The fact that the general pdf f_X is a density follows from a change of coordinates $y = (x-\mu)/\sigma$ with $dx = \sigma\,dy$, so that

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}\,dx = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-y^2/2}\sigma\,dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\,dy = 1,$$

where the latter integral is what we computed previously.

Example 5.3.3. A random variable X has the *standard Cauchy* distribution if its pdf is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

Clearly we have $f_X(x) \geq 0$, and also we can compute that

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)}\,dx = \frac{1}{\pi}\lim_{N,M\to\infty}\int_{-M}^{N}\frac{1}{1+x^2}\,dx = \frac{1}{\pi}\lim_{N,M\to\infty}[\arctan x]_{-M}^{N} = \frac{1}{\pi}\left(\frac{\pi}{2}+\frac{\pi}{2}\right) = 1.$$

5.4 Expectation of Continuous Random Variables

Theorem 5.1 (Law of the subconscious statistician)**.** *Let X be a random variable with pdf f_X, and let $h : \mathbb{R} \to \mathbb{R}$ be a sufficiently nice function.*[1] *Let $Y = h(X)$. Then*

$$\mathbb{E}[Y] = \mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f_X(x)\,dx,$$

provided that $\displaystyle\int_{-\infty}^{\infty} h(x)f_X(x)\,dx$ converges absolutely, i.e. $\displaystyle\int_{-\infty}^{\infty} |h(x)|f_X(x)\,dx < \infty$.[2]

Proof. Use the Riemann sum approximation. First restrict to a compact interval, e.g. $[0,1]$. Then

$$\int_0^1 h(x)f_X(x)\,dx \approx \sum_{i=1}^{n} h(x_i)f_X(x_i)\Delta x$$

for large n. Now appeal to the discrete version of this result. □

Example 5.3.4. Let $X \sim \mathcal{N}(0,1)$. Then we would like to find

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-x^2/2}\,dx.$$

First we must check that this integral converges absolutely. We write

$$\mathbb{E}[|X|] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} |x|e^{-x^2/2}\,dx = \lim_{M_1,M_2\to\infty}\frac{1}{\sqrt{2\pi}}\int_{-M_1}^{M_2}|x|e^{-x^2/2}\,dx.$$

In the finite setting, we can split the integral to get

$$\begin{aligned}
\mathbb{E}[|X|] &= \lim_{M_1,M_2\to\infty}\left[\frac{1}{\sqrt{2\pi}}\int_0^{M_2} xe^{-x^2/2}\,dx - \frac{1}{\sqrt{2\pi}}\int_{-M_1}^{0} xe^{-x^2/2}\,dx\right] \\
&= \lim_{M_2\to\infty}\left[\frac{1}{\sqrt{2\pi}}(-e^{-x^2/2})\right]_{x=0}^{x=M_2} + \lim_{M_1\to\infty}\left[\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right]_{x=-M_1}^{x=0} \\
&= \lim_{M_2\to\infty}\frac{1}{\sqrt{2\pi}}(1-e^{-M_2^2/2}) + \lim_{M_1\to\infty}\frac{1}{\sqrt{2\pi}}(1-e^{-M_1^2/2}) = \sqrt{2/\pi}.
\end{aligned}$$

[1] Here h must be *measurable*.

[2] The integral of $h(x)F_X(x)$ must be absolutely convergent for its *Lebesgue integral* to exist.

This establishes that $\mathbb{E}[X]$ converges absolutely. Thus we can argue that

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2}\,dx = 0$$

since $xe^{-x^2/2}$ is an odd function on $\mathbb{R}$ (we can restrict to $[-M, M]$ and then take $M \to \infty$).

Example 5.3.5. Now let $X \sim \mathcal{N}(\mu, \sigma^2)$, and we find $\mathbb{E}[X]$. We have

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} xe^{-(x-\mu)^2/(2\sigma^2)}\,dx.$$

Make the change of variables $y = (x - \mu)/\sigma$ with $dx = \sigma\,dy$ to get

$$\begin{aligned}\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sigma y + \mu)e^{-y^2/2}\sigma\,dy &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ye^{-y^2/2}\,dy + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy \\ &= 0 + \mu = \mu\end{aligned}$$

Example 5.3.6. Let $X \sim \mathcal{N}(0, 1)$, and we find $\mathbb{E}[X^2]$. We have

$$\mathbb{E}[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx.$$

This integrand is nonnegative, so no need to justify absolute convergence. Now integrate by parts with $u(x) = x$ and $v'(x) = xe^{-x^2/2}$ (with $u'(x) = 1$ and $v(x) = -e^{-x^2/2}$) to get

$$\mathbb{E}[X^2] = \frac{1}{\sqrt{2\pi}} \left[-xe^{-x^2/2}\right]_{x=-\infty}^{x=\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = 0 + 1 = 1.$$

In particular, this gives $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1 - 0 = 1$.

Example 5.3.7. In general, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we can compute

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - \mu)^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-(x-\mu)^2/2\sigma^2}\,dx.$$

Now change variables with $y = (x - \mu)/\sigma$ and $dx = \sigma\,dy$ to get

$$\mathrm{Var}[X] = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \sigma^2 y^2 e^{-y^2/2}\sigma\,dy = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2}\,dy = \sigma^2.$$

Remark. The previous computations show that for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.

Example 5.3.8. Let X be a standard Cauchy random variable. We compute

$$\begin{aligned}\mathbb{E}[|X|] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{|x|}{1 + x^2}\,dx &\geq \frac{1}{\pi} \int_{1}^{\infty} \frac{x}{1 + x^2}\,dx = \frac{1}{\pi} \lim_{N\to\infty} \int_{1}^{N} \frac{x}{1 + x^2}\,dx \\ &= \frac{1}{\pi} \lim_{N\to\infty} \left[\frac{\ln(1 + x^2)}{2}\right]_{x=1}^{x=N} = \frac{1}{\pi} \lim_{N\to\infty} \left[\frac{\ln(1 + N^2)}{2} - \frac{\ln 2}{2}\right] = \infty.\end{aligned}$$

In particular, this means that $\mathbb{E}[X]$ does not exist. This is an example of a *heavy-tailed* random variable.

Remark. For any random variable X, if $\mathbb{E}[|X|] = \infty$ then $\mathbb{E}[X^2] = \infty$ also. This is because

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx \geq \int_{|x|\geq 1} x^2 f_X(x)\,dx \geq \int_{|x|\geq 1} |x| f_X(x)\,dx = \infty,$$

since $x^2 \geq |x|$ for $|x| \geq 1$. Note that $|x| f_X(x)$ is bounded on $|x| \leq 1$, so its integral on $[-1, 1]$ is finite.

5.5 The Gamma and Beta Random Variables

Definition 5.4. Define the *gamma function* Γ for $\alpha > 0$ by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx.$$

Proposition 5.2. *The gamma function* Γ *has the following properties:*

(i) $\Gamma(1) = 1$,

(ii) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, *and in particular* $\Gamma(n) = (n-1)!$ *for positive integers* n,

(iii) and $\Gamma(1/2) = \sqrt{\pi}$.

Proof. (i) We have

$$\Gamma(1) = \int_0^\infty e^{-x}\, dx = 1.$$

(ii) We can integrate by parts with $u(x) = x^\alpha$ and $v'(x) = e^{-x}$ to get

$$\Gamma(\alpha+1) = \int_0^\infty x^\alpha e^{-x}\, dx = \left[x^\alpha(-e^{-x})\right]_0^\infty + \alpha \int_0^\infty x^{\alpha-1} e^{-x}\, dx = \alpha\Gamma(\alpha).$$

The second part follows by induction on n.

(iii) Change variables with $u = \sqrt{x}$ and $dx/\sqrt{x} = du$ to get

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-x}}{\sqrt{x}}\, du = \int_0^\infty 2e^{-u^2}\, du = \sqrt{\pi}.$$

Note that one can extend this by induction to compute $\Gamma(\alpha)$ at positive half integers. □

Example 5.4.1. A random variable X has the *standard gamma* distribution if its pdf is given by

$$f_X(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, \quad x \in (0, \infty)$$

for a constant $\alpha > 0$.[3] In general, X is a *gamma random variable* with parameters $\alpha > 0$ and $\lambda > 0$ if

$$f_X(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0.$$

Clearly $f_X(x) \geq 0$. For the integral, make the change of variables $y = \lambda x$ with $dx = dy/\lambda$ to get

$$\int_0^\infty \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}\, dx = \frac{1}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha \left(\frac{y}{\lambda}\right)^{\alpha-1} e^{-y}\, \frac{dy}{\lambda} = \frac{\lambda^\alpha}{\Gamma(\alpha)\lambda^\alpha} \int_0^\infty y^{\alpha-1} e^{-y}\, dy = 1.$$

Remark. There are relationships between the gamma, Poisson, and geometric random variables, which are useful in industrial engineering.

[3] If $\alpha = 1$, then this reduces to the exponential distribution.

Example 5.4.2. A random variable X is said to be a *beta random variable* with parameters $s, t > 0$ if

$$f_X(x; s, t) = \frac{1}{\beta(s,t)} x^{s-1}(1-x)^{t-1}, \quad 0 < x < 1,$$

where β is the *beta function* given by

$$\beta(s,t) = \int_0^1 x^{s-1}(1-x)^{t-1}\, dx = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}.$$

Definition 5.5. Taking a gamma random variable with parameters $\alpha = n/2$ and $\lambda = 1/2$, one gets

$$f_X\left(x; \frac{n}{2}, \frac{1}{2}\right) = \frac{1}{2\Gamma(n/2)} \left(\frac{x}{2}\right)^{n/2-1} e^{-x/2}, \quad x > 0,$$

which is called a χ^2 *(chi-squared) random variable* with n degrees of freedom.

Remark. The reason we care about χ^2 random variables is that if $Z \sim \mathcal{N}(0,1)$, then $Z^2 = \chi^2(1)$. In general, if $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$ are independent, then $Z_1^2 + \cdots + Z_n^2 = \chi^2(n)$.

5.6 Functions of Continuous Random Variables

Remark. Suppose X has pdf f_X and $h : \mathbb{R} \to \mathbb{R}$, so $Y = h(X)$ is a random variable. If h is "nice," then Y has a density f_Y. What is the relationship between f_X and f_Y?

Remark. Note that f_Y need not have a density in general. Let $X \sim \text{Unif}([0,1])$ and

$$h(x) = \begin{cases} \pi & \text{if } 0 \le x < 1/2 \\ e & \text{if } 1/2 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then $Y = h(X)$ only takes two values with positive probability, so it cannot have a density (it is discrete).

Example 5.5.1. Suppose first that $h(x) = ax + b$ with $a > 0$. Then $Y = h(X) = aX + b$, and

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(aX + b \le y) = \mathbb{P}\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

Then we can differentiate to get

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

If $a < 0$, then the inequality flips and we get

$$F_Y(y) = \mathbb{P}\left(X \ge \frac{y-b}{a}\right) = 1 - \mathbb{P}\left(X < \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right),$$

so differentiating yields

$$f_Y(y) = \frac{d}{dy} F_Y(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

Note that $a < 0$, so $-a > 0$ and thus $f_Y(y) \ge 0$.

Theorem 5.2. *Let X be an absolutely continuous random variable with density function f_X, and let $h : \mathbb{R} \to \mathbb{R}$ be strictly increasing and differentiable. Then $Y = h(X)$ has a density function given by*

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y) \quad \textit{for all } y \in \mathbb{R}.$$

Proof. Since h is strictly increasing, applying its inverse preserves the inequality, so that

$$F_Y(y) = \mathbb{P}(h(X) \leq y) = \mathbb{P}(X \leq h^{-1}(y)).$$

Differentiating this via the chain rule gives the desired formula for f_Y. □

Remark. Note that

$$\frac{d}{dy}h^{-1}(y) = \frac{1}{h'(h^{-1}(y))}.$$

This is because $h^{-1}(h(x)) = x$, so differentiating via the chain rule gives

$$1 = \frac{d}{dx}x = \frac{d}{dx}h^{-1}(h(x)) = (h^{-1})'(h(x))h'(x).$$

Dividing by $h'(x)$ and setting $x = h^{-1}(y)$ gives the desired formula.

Remark. Using the previous theorem, we can create random number generators following a specified distribution, given a random number generator following a uniform distribution on $[0, 1]$. Suppose that X has cdf F_X which is strictly increasing (if X has a density f_X, then this holds on the support of X), so that F_X is invertible. Suppose that we are given $U \sim \text{Unif}([0, 1])$. Then

$$\mathbb{P}(F_X^{-1}(U) \leq x) = \mathbb{P}(F_X(F_X^{-1}(U)) \leq F_X(x)) = \mathbb{P}(U \leq F_X(x)) = F_X(x)$$

since F_X is increasing (so the direction of the inequality is preserved), and the cdf of U is simply the identity function on $[0, 1]$. This says that $F_X^{-1}(U) \sim X$, so being able to sampling from U is enough. In general, when F_X is not strictly increasing, one can take the *generalized inverse*:

$$F_X^-(y) = \inf\{x \in \mathbb{R} : F_X(x) \geq y\}.$$

Then one can show that again we have $F_X^-(U) \sim X$.

5.7 Geometric Probability

Example 5.5.2 (Bertrand's paradox)**.** Consider a unit disk U, and choose a chord C of the disk at random. Let T be the equilateral triangle with one side C (there are two, choose the one that points towards the center of the disk). Then what is $\mathbb{P}(T \subseteq U)$? The issue comes from how we choose C:

1. Suppose we choose angles $X, Y \in [0, 2\pi]$ uniformly at random, and then choose the chord C to go through X, Y. This amounts to choose the difference uniformly at random in $[0, 2\pi]$, since the problem is rotationally invariant. We need $|X - Y| \leq 2\pi/3$, so

$$\mathbb{P}(T \subseteq U) = \mathbb{P}(-2\pi/3 \leq X - Y \leq 2\pi/3) = \frac{2}{3}$$

2. Alternatively we could choose a point X at random and then choose a random angle $\theta \in [0, \pi/2]$ (from the tangent line to U at X). Clearly X does not matter, and we need $\theta \leq \pi/3$. This gives
$$\mathbb{P}(T \subseteq U) = \mathbb{P}(\theta \leq \pi/3) = 1/3 = \frac{\pi/3}{\pi/2} = \frac{2}{3}.$$

3. We could also choose a direction θ, which determines a line through the circle. Then choose a point $M \in [-1, 1]$ uniformly at random on this line, which determines a chord if we let M be the midpoint and say that the chord is perpendicular to the line. Then we need $-1/2 \leq M \leq 1/2$, os
$$\mathbb{P}(T \subseteq U) = \mathbb{P}(-1/2 \leq M \leq 1/2) = \frac{1}{2}.$$

4. Finally, we could simply choose the midpoint $M \in U$ at random. We need M to be within the circle of radius $1/2$ with the same center as M. Thus in this case
$$\mathbb{P}(T \subseteq U) = \frac{\pi - \pi/4}{\pi} = \frac{3}{4}.$$

We see that, depending on the method in which we choose the chord, the answer may be different.

Example 5.5.3 (Buffon's needle)**.** Consider horizontal lines in the plane, each spaced distance 1 apart. Drop a needle of length 1 at random. What is the probability that the needle intersects a line?

- One way to proceed is to choose (X, Y) first as the position of the center of the needle. Then choose $\theta \in [0, \pi]$ to be the angle of the needle. Note the actual position of the needle does not matter, so we can simply choose $Y \in [0, 1]$. Now there are two cases:

  1. If the needle intersects with the upper line, then we need
  $$Y - \frac{1}{2}\sin\theta \leq 1 \leq Y + \frac{1}{2}\sin\theta \iff 1 - \frac{1}{2}\sin\theta \leq Y \leq 1 + \frac{1}{2}\sin\theta.$$
  Note that the second inequality is always true, so we only need $Y \geq 1 - \frac{1}{2}\sin\theta$.

  2. If the needle intersects with the lower line, then we need
  $$Y - \frac{1}{2}\sin\theta \leq 0 \leq Y + \frac{1}{2}\sin\theta \iff -\frac{1}{2}\sin\theta \leq Y \leq \frac{1}{2}\sin\theta.$$
  This time the first inequality is always true, so we just need $Y \leq \frac{1}{2}\sin\theta$.

  So for the needle to intersect the lines, we need $Y \leq \frac{1}{2}\sin\theta$ or $Y \geq 1 - \frac{1}{2}\sin\theta$. Consider the subset of $[0, \pi] \times [0, 1]$ which satsifies the above condition. There are two symmetric regions of area
  $$\int_0^\pi \frac{1}{2}\sin\theta \, d\theta = \frac{1}{2}[-\cos\theta]_{\theta=0}^{\theta=\pi} = \frac{1}{2}[1 - (-1)] = 1.$$
  Thus the area is 2, and the probability is $2/\pi$.

This gives one way to estimate hte value of π. Let I_n be the number of times the needle intersects with a line in n trials. Then $I_n/n \to 2/\pi$ as $n \to \infty$, so
$$\widehat{\pi}_n = \frac{2n}{I_n} \to \pi.$$

This is called the *Monte Carlo method*, which can also be used to estimate integrals.[4]

[4]Note that there are much more efficient methods than Monte Carlo for estimating π.

Example 5.5.4 (Stick breaking)**.** Choose $X, Y \in [0,1]$ uniformly at random. This determines two cuts of $[0,1]$, which breaks the interval into three segments. What is the probability that these segments form the sides of a triangle?

- The pieces are of length $X, Y - X, 1 - Y$ (assume that $X \leq Y$, which we can do since we are choosing uniformly at random). By the triangle inequality, we need

$$\begin{cases} X + (Y - X) \geq 1 - Y \\ (Y - X) + (1 - Y) > X \\ X + (1 - Y) > Y - X \end{cases} \implies \begin{cases} Y > 1/2 \\ X < 1/2 \\ Y < 1/2 + X. \end{cases}$$

  Draw the region in $[0,1] \times [0,1]$ that satisfies the above, and we find that the probability is $1/4$.

Example 5.5.5. Choose n points on a circle uniformly at random. What is the probability that all n points lie in some semicircle?

- Let A be the event $\{$all n points lie in some semicircle$\}$. The goal is to work with easier events. Number the points by $P_1, P_2, \ldots, P_n$. Then let A_i be the event that $P_1, P_2, \ldots, P_n$ lie in the semicircle which begins at P_i. Then we easily find that $\mathbb{P}(A_i) = 1/2^{n-1}$ for each $1 \leq i \leq n$. Also

$$A = A_1 \cup A_2 \cup \ldots A_n,$$

  where the union is disjoint, since we cannot have two different points both be the first point in a semicircle (note that $\mathbb{P}(P_i = P_j) = 0$ for $i \neq j$). Thus

$$\mathbb{P}(A) = \mathbb{P}(A_1 \cup \cdots \cup A_n) = \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) = n\mathbb{P}(A_1) = \frac{n}{2^{n-1}}$$

  since the union is disjoint.

If the problem is more difficult (the union is not disjoint), one can try to apply inclusion-exclusion.

5.8 Homework Problems

Problems #1, 2, 3, 4, 7, 9, 10, 14 from Grimmett and Welsh.

Chapter 6

Multivariate Continuous Distributions

6.1 Multivariate Absolutely Continuous Distributions

Definition 6.1. Let (X, Y) be a two-dimensional random vector, its *distribution function* is given by

$$F_{(X,Y)}(x, y) = \mathbb{P}(X \leq x, Y \leq y) \quad \text{for } (x, y) \in \mathbb{R}^2.$$

Proposition 6.1. *We have the following properties for the distribution function of* (X, Y)*:*

(i) $\lim_{x\to-\infty} F_{(X,Y)}(x, y) = 0$ *and* $\lim_{y\to-\infty} F_{(X,Y)}(x, y) = 0$.

(ii) $\lim_{x,y\to\infty} F_{(X,Y)}(x, y) = 1$.

(iii) $F_{(X,Y)}$ *is weakly increasing in each variable: If* $x_1 \leq x_2$ *then* $F_{(X,Y)}(x_1, y) \leq F_{(X,Y)}(x_2, y)$ *and if* $y_1 \leq y_2$ *then* $F_{(X,Y)}(x, y_1) \leq F_{(X,Y)}(x, y_2)$.

(iv) If $x_1 \leq y_1$ *and* $x_2 \leq y_2$*, then*[1]

$$\int_{x_1}^{x_2} \int_{y_1}^{y_2} dF(u, v) \geq 0.$$

Proof. (i) We have

$$\begin{aligned}\lim_{x\to-\infty} F_{(X,Y)}(x, y) &= \lim_{x\to-\infty} \mathbb{P}(X^{-1}((-\infty, x]) \cap Y^{-1}((-\infty, y])) \\ &= \mathbb{P}(Y^{-1}((-\infty, y]) \cap \bigcap_{x\in\mathbb{R}} X^{-1}((-\infty, x]) \\ &= \mathbb{P}(Y^{-1}((-\infty, y]) \cap \varnothing) = 0.\end{aligned}$$

(ii) We have

$$\lim_{x\to\infty} F_{(X,Y)}(x, y) = \mathbb{P}(\{X \leq \infty\} \cap \{Y \leq y\}) = \mathbb{P}(Y \leq y) = F_Y(y).$$

Then taking $y \to \infty$ gives $\lim_{y\to\infty} F_Y(y) = 1$.

(iii) If $x_1 \leq x_2$ and y is fixed, then $\mathbb{P}(X \leq x_1, Y \leq y) \leq \mathbb{P}(X \leq x_2, Y \leq y)$ since

$$X^{-1}((-\infty, x_1]) \cap Y^{-1}((-\infty, y]) \subseteq X^{-1}((-\infty, x_2]) \cap Y^{-1}((-\infty, y]).$$

(iv) The integral corresponds to $\mathbb{P}(X \in (x_1, x_2), Y \in (y_1, y_2))$, which must be nonnegative. □

Remark. The above properties characterize distribution functions of two-dimensional random vectors.

[1] One can interpret this integral as a *Riemann-Stieltjes integral*.

6.2 Independence of Continuous Random Variables

Definition 6.2. Two random variables X, Y are said to be *independent* if for all $x, y \in \mathbb{R}$, the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent.

Proposition 6.2. *Two continuous random variables X, Y are independent if and only if*

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$$

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

Proof. ($\Rightarrow$) If X and Y are independent, then $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for all $x, y \in \mathbb{R}$, so

$$\mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{P}(\{X \leq x\})\mathbb{P}(\{Y \leq y\}),$$

i.e. we have $F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$.

($\Leftarrow$) If $F_{(X,Y)}(x, y) = F_X(x)F_Y(y)$, then for all $x, y \in \mathbb{R}$, then

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

i.e. the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent. □

Definition 6.3. The pair of random variables (X, Y) is said to be *jointly (absolutely continuous)* if its joint cdf can be written as

$$F_{(X,Y)}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{(X,Y)}(u, v)\, dudv$$

for all $x, y \in \mathbb{R}$ and $f_{(X,Y)} : \mathbb{R}^2 \to [0, \infty)$. If this is the case, then $f_{(X,Y)}$ is called the *joint pdf* of (X, Y).

Remark. We have

$$\frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x, y) = f_{(X,Y)}(x, y)$$

if this derivative exists at (x, y). If it does not exist then we set $f_{X,Y}(x, y) = 0$ (this is fine since an absolutely continuous function is differentiable *almost everywhere*, i.e. except on a set of measure zero).

Remark. In the context of densities, we have for $A \subseteq \mathbb{R}^2$ (which is "nice enough") that

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{(X,Y)}(u, v)\, dudv.$$

Here the "nice enough" sets are the ones which can be built up from rectangles.

Definition 6.4. Let (X, Y) be a continuous random vector with joint pdf $f_{(X,Y)}$. The *marginal pdf* of X, denoted by f_X is given by

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

The marginal pdf f_Y of Y is defined similarly, integrating out the x variable instead.

Remark. Clearly f_X (as defined above) is nonnegative and also

$$\int_{-\infty}^{\infty} f_X(x)\, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(X,Y)}(x, y)\, dy\, dx = 1.$$

Thus f_X is indeed a valid density function.

Proposition 6.3. *Let X and Y be jointly absolutely continuous random variables with joint pdf $f_{(X,Y)}$. Then X and Y are independent if and only if $f_{(X,Y)}$ can be expressed as*

$$f_{(X,Y)}(x,y) = g(x)h(y)$$

for every $x, y \in \mathbb{R}$, where $g, h : \mathbb{R} \to [0,\infty)$.

Proof. ($\Rightarrow$) By independence, we have $F_{(X,Y)}(x,y) = F_X(x)F_Y(y)$, so

$$\begin{aligned} f_{(X,Y)}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{(X,Y)}(x,y) &= \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} F_X(x)F_Y(y)\right) = \frac{\partial}{\partial x}\left(F_X(x)\frac{\partial}{\partial y}F_Y(y)\right) \\ &= \frac{\partial}{\partial x}\left(F_X(x)f_Y(y)\right) = f_X(x)f_Y(y). \end{aligned}$$

This gives the desired conclusion in this direction.

($\Leftarrow$) Check this as an exercise. □

Proposition 6.4. *Random variables X, Y are independent if and only if*

$$\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]$$

for all $h_1, h_2 : \mathbb{R} \to \mathbb{R}$.

Proof. ($\Rightarrow$) This is clear by the independence of X and Y.

($\Leftarrow$) As in the discrete case, use indicator functions. Then argue using limits of indicators. □

Example 6.4.1. Let X, Y be two random variables with joint pdf given by

$$f_{(X,Y)}(x,y) = 2e^{-x-y}, \quad 0 < x < y < \infty.$$

Are X and Y independent? The answer is no. It might be tempting to say that $f_{(X,Y)}$ factors, but it does not factor *over all of* $\mathbb{R}^2$: The region where the above formula holds adds dependence between x and y. To verify this, we can compute the marginal densities. We have

$$f_X(x) = \int_x^\infty f_{(X,Y)}(x,y)\,dy = \int_x^\infty 2e^{-x}e^{-y}\,dy = 2e^{-x}\left[-e^{-y}\right]_{y=x}^{y=\infty} = 2e^{-2x}$$

for $x > 0$. This says that $X \sim \text{Exp}(2)$. Similarly, we can find that

$$f_Y(y) = \int_0^y 2e^{-x}e^{-y}\,dx = 2e^{-y}\left[-e^{-x}\right]_{x=0}^{x=y} = 2e^{-y}(-e^{-y}+1) = 2e^{-y} - 2e^{-2y}$$

for $y > 0$. Note that $f_Y(y) \geq 0$ since $1 - e^{-y} \geq 0$ for $y > 0$, and

$$\int_0^\infty f_Y(y)\,dy = \int_0^\infty 2e^{-y}(-e^{-y}+1)\,dy = 2\int_0^\infty e^{-y}\,dy - \int_0^\infty 2e^{-2y}\,dy = 2 - 1 = 1,$$

where we recognize the first integral as the integral of a standard exponential density and the second integral as the integral of an exponential density with parameter $\lambda = 2$. Now we can see that

$$f_{(X,Y)}(x,y) = 2e^{-x-y} \neq 2e^{-2x}(2e^{-y} - 2e^{-2y}) = f_X(x)f_Y(y),$$

which shows that X and Y are not independent.

6.3 Transformations of Random Variables

Theorem 6.1. *Let X and Y be two independent random variables with respective density functions f_X and f_Y. Then $Z = X + Y$ has density function given by*

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$$

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

Proof. For fixed $z \in \mathbb{R}$, we have

$$\mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) = \iint_{A_z} f_{(X,Y)}(x, y)\, dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dydx,$$

where $A_z = \{(x, y) \in \mathbb{R}^2 : x + y \leq z\}$. By substituting $u = x$ and $v = x + y$ (and exchanging the order of integration, which is permissible by Tonelli's theorem since $f_{(X,Y)}(x, y) \geq 0$), we get

$$\mathbb{P}(Z \leq z) = \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_{(X,Y)}(u, v - u)\, dudv = \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_X(u) f_Y(v - u)\, dudv$$

by the independence of X, Y. Now differentiating both sides in z gives the desired result. □

Theorem 6.2. *Let X and Y be jointly absolutely continuous with joint density $f_{(X,Y)}$. Let*

$$D = \{(x, y) \in \mathbb{R}^2 : f_{(X,Y)}(x, y) > 0\} \quad \text{and} \quad T : (x, y) \to T(x, y) = (u(x, y), v(x, y)).$$

Assume that T is a bijection from D to $S \subseteq \mathbb{R}^2$. Then the pair $(U, V) = (u(X, Y), v(X, Y))$ is jointly absolutely continuous with density function

$$f_{(U,V)}(u, v) = \begin{cases} f_{(X,Y)}(x(u, v), y(u, v))|J(u, v)| & \text{if } (u, v) \in S, \\ 0 & \text{otherwise,} \end{cases}$$

where $J(u, v)$ is the Jacobian determinant of T^{-1} at (u, v).

Proof. Let $A \subseteq D$ and $T(A) = B$. Since T is a bijection, we have

$$\mathbb{P}((U, V) \in B) = \mathbb{P}(T(X, Y) \in B) = \mathbb{P}((X, Y) \in A).$$

Now by the change of variables formula from multivariable calculus, we have

$$\begin{aligned} \mathbb{P}((U, V) \in B) = \mathbb{P}((X, Y) \in A) &= \iint_A f_{(X,Y)}(x, y)\, dxdy \\ &= \iint_B f_{(X,Y)}(x(u, v), y(u, v))|J(u, v)|\, dudv. \end{aligned}$$

From this we obtain

$$\iint_B f_{(U,V)}(u, v)\, dudv = \mathbb{P}((U, V) \in B) = \iint_B f_{(X,Y)}(x(u, v), y(u, v))|J(u, v)|\, dudv.$$

Since this holds for every $B \subseteq S$, we deduce that the integrands must be equal (if they differed on a set E of positive measure, then their integrals will differ on a subset of E). □

6.4 Conditional Distributions

Definition 6.5. The *conditional density* of X given $\{Y = y\}$ is denoted by

$$F_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{f_{(X,Y)}(x,y)}{\displaystyle\int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\,dx}$$

for all $y \in \mathbb{R}$ with $f_Y(y) > 0$.

Example 6.5.1. Let

$$f_{(X,Y)}(x,y) = \begin{cases} 2e^{-x-y} & \text{if } 0 < x < y < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{2e^{-x-y}}{2e^{-2x}} = e^{x-y} \quad 0 < x < y < \infty$$

and

$$f_{X|Y}(x|y) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)} = \frac{2e^{-x-y}}{2e^{-y}(1-e^{-y})} = \frac{e^{-x}}{1-e^{-y}} \quad 0 < x < y < \infty,$$

using the marginal pdfs we computed from a previous example.

Definition 6.6. The *conditional expectation* of X given $\{Y = y\}$ is defined by

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,dx = \int_{-\infty}^{\infty} x \frac{f_{(X,Y)}(x,y)}{f_Y(y)}\,dx$$

for all $y \in \mathbb{R}$ with $f_Y(y) > 0$.

Proposition 6.5. *For continuous random variables X, Y with density functions f_X, f_Y, we have*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] f_Y(y)\,dy \quad \textit{and} \quad \mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X=x] f_X(x)\,dx.$$

Proof. We have

$$\begin{aligned} \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] f_Y(y)\,dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x \frac{f_{(X,Y)}(x,y)}{f_Y(y)} f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{(X,Y)}(x,y)\,dy = \mathbb{E}[X]. \end{aligned}$$

The computation for $\mathbb{E}[Y]$ is identical. □

6.5 Bivariate Normal Distribution

Definition 6.7. The random vector (X, Y) is said to be *normally distributed* if

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

for all $x, y \in \mathbb{R}$ and $\rho \in (-1, 1)$.

Remark. Note that if $\rho = 0$, then the joint density becomes

$$f_{(X,Y)}(x,y) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x^2+y^2)\right) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\cdot\frac{1}{\sqrt{2\pi}}e^{-y^2/2}$$

for every $x, y \in \mathbb{R}$. So X, Y are independent and $X, Y \sim \mathcal{N}(0,1)$.

Example 6.7.1. We can compute the marginal distribution of X by

$$\begin{aligned} f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)\,dxdy &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty} e^{-(x^2-2\rho xy+y^2)/2(1-\rho^2)}\,dy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{\infty} e^{-((y-\rho x)^2+x^2(1-\rho^2))/2(1-\rho^2)}\,dy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}}e^{-x^2(1-\rho^2)/2(1-\rho^2)}\int_{-\infty}^{\infty} e^{-(y-\rho x)^2/2(1-\rho^2)}\,dy \\ &= \frac{e^{-x^2/2}}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{2\pi(1-\rho^2)}}\int_{-\infty}^{\infty} e^{-(y-\rho x)^2/2(1-\rho^2)}\,dy = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \end{aligned}$$

since the last integral is the integral of the pdf of a $\mathcal{N}(\rho x, 1-\rho^2)$ random variable. From this we can recognize that $X \sim \mathcal{N}(0,1)$, and by symmetry we also see that $Y \sim \mathcal{N}(0,1)$.

Example 6.7.2. Now we compute the conditional pdf of Y given $X = x$. This is

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} = \frac{e^{-((y-\rho^2)+x^2(1-\rho^2))/2(1-\rho^2)}}{\sqrt{2\pi}\sqrt{2\pi(1-\rho^2)}}\cdot\frac{\sqrt{2\pi}}{e^{-x^2/2}} = \frac{e^{-(y-\rho x)^2/2(1-\rho^2)}}{\sqrt{2\pi(1-\rho^2)}}.$$

This shows that $Y|X = x$ has a normal density with mean ρx and variance $1-\rho^2$. We also get the formula for $f_{X|Y}(x|y)$ by symmetry by exchanging the roles of x and y. Note that this also gives

$$\mathbb{E}[Y|X=x] = \rho x \quad \text{and} \quad \mathbb{E}[X|Y=y] = \rho y.$$

Note that the conditional expectation of X given $Y = y$ is linear in y (and linear in x for $Y|X = x$).

Example 6.7.3. We compute $\mathbb{E}[XY]$. We can write

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty}\mathbb{E}[XY|X=x]f_X(x)\,dx.$$

Since we condition on $X = x$, we have $\mathbb{E}[XY|X=x] = \mathbb{E}[xY|X=x] = x\mathbb{E}[Y|X=x]$, and so

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} x\mathbb{E}[Y|X=x]f_X(x)\,dx = \int_{\infty}^{\infty} x(\rho x)f_X(x)\,dx = \rho\int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \rho$$

since we recognize the last integral as the integral for $\mathrm{Var}[X] = 1$.

Remark. The above computations show that

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \rho - 0\cdot 0 = \rho.$$

So the parameter ρ in the bivariate normal pdf is the covariance of X and Y.

Remark. If (X, Y) is a normal random vector, then X and Y are independent *if and only if* X and Y are uncorrelated, i.e. $\mathrm{Cov}(X, Y) = 0$. Recall that being uncorrelated does not imply independence in general. This special case works because the conditional expectations are linear.

Definition 6.8. The *correlation coefficient* of two random variables X and Y is given by

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}},$$

provided that $\mathrm{Var}[X], \mathrm{Var}[Y] > 0$, i.e. $X \neq \mathbb{E}[X]$ and $Y \neq \mathbb{E}[Y]$.

Definition 6.9. Let $\mu \in \mathbb{R}^d$ and Σ be a positive definite symmetric $d \times d$ matrix. We say that X has the multivariate normal distribution with mean μ and covariance matrix Σ, written $X \sim \mathcal{N}(\mu, \Sigma)$, if

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \quad \text{for all } x \in \mathbb{R}^d.$$

Remark. When $d = 2$, we get

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

as the joint pdf. Here we have

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

If $X = (X_1, X_2)$, then $\mathbb{E}[X_1] = \mu_1$, $\mathbb{E}[X_2] = \mu_2$, $\mathrm{Var}[X_1] = \sigma_1^2$, $\mathrm{Var}[X_2] = \sigma_2^2$, and $\rho(X_1, X_2) = \rho$.

6.6 Homework Problems

Problems #1, 3, 4, 9, 11, 15, 20, 21, 24, 26 from Grimmett and Welsh.

Chapter 7

Moment Generating Functions

7.1 Other Types of Random Variables

Remark. Random variables may be neither discrete nor absolutely continuous. Let $X \sim \text{Exp}(1)$ with

$$f_X(x) \begin{cases} e^{-x} & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-x} & \text{if } x > 0. \end{cases}$$

Let Y be a discrete random variable, say given by

$$\mathbb{P}(Y = k) = 2^{k-1}, \quad k = 1, 2, \ldots.$$

Define

$$F(x) = \frac{1}{2}F_Y(x) + \frac{1}{2}F_X(x) = \frac{1}{2}F_Y(x) + \frac{1}{2}(1 - e^{-x}).$$

One can show that F is increasing, $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$, and F is right-continuoous, so F is a distribution function. But F is not a step function or continuous, so the random variable with distribution function F is neither discrete nor continuous.

Remark. In general, there is another type of random variable called a *singular random variable*, which has zero density almost everywhere but is not constant. One can show that every random variable X can be uniquely decomposed as

$$X = X_{\text{disc}} + X_{\text{cont}} + X_{\text{sing}},$$

where X_{disc} is discrete, X_{cont}, is absolutely continuous, and X_{sing} is singular continuous.

7.2 Covariance

Definition 7.1. For two random variables X, Y with $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$, their *covariance* is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Remark. Why do we need $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$? This is because of the *Cauchy-Schwarz inequality*.

Proposition 7.1 (Cauchy-Schwarz inequality)**.** *Let X, Y be two random variables with $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$. Then we have*

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}.$$

Proof. Let $x \in \mathbb{R}$ and consider $A(x) = xX + Y$. Then

$$0 \leq P(x) = \mathbb{E}[(xX+Y)^2] = \mathbb{E}[x^2X^2 + 2xXY + Y^2] = x^2\mathbb{E}[X^2] + 2x\mathbb{E}[XY] + \mathbb{E}[Y^2].$$

Since $P(x) \geq 0$ for all $x \in \mathbb{R}$, its discriminant must be non-positive, i.e.

$$(2\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0,$$

which implies that $\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$, or $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}$. □

Remark. Why does $\mathbb{E}[X^2] < \infty$ imply that $\mathbb{E}[XY] < \infty$? This is because

$$\mathbb{E}[|X|] = \mathbb{E}[|X| \cdot 1] \leq \sqrt{\mathbb{E}[|X|^2]}\sqrt{\mathbb{E}[1]} = \sqrt{\mathbb{E}[X^2]} < \infty$$

by the Cauchy-Schwarz inequality, when $\mathbb{E}[X^2] < \infty$.

Remark. Recall the correlation coefficient $\rho(X,Y)$. When $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$, we have

$$\begin{aligned}|\mathrm{Cov}(X,Y)| = |\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]| &\leq \sqrt{\mathbb{E}[(X-\mathbb{E}[X])^2]}\sqrt{\mathbb{E}[(Y-\mathbb{E}[Y])^2]} \\ &= \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}\end{aligned}$$

by the Cauchy-Schwarz inequality. This implies that $-1 \leq \rho(X,Y) \leq 1$. Note that $\rho(X,Y)$ is analogous to the cosine of the angle θ between two vectors, where for $x, y \in \mathbb{R}^n$ we defined

$$\cos\theta = \frac{\langle x, y\rangle}{\|x\|\|y\|} = \frac{\langle x, y\rangle}{\sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle}}.$$

Recall that $\cos\theta$ if and only if $x \perp y$. Similarly, $\rho(X,Y) = 0$ if and only if $\mathrm{Cov}(X,Y) = 0$, which is a sense of *orthogonality* for X and Y. This is justification for the notation $X \perp\!\!\!\perp Y$ for independence. Also,

$$\begin{aligned}\cos\theta = 1 \quad &\text{if and only if} \quad x = ay,\ a > 0, \\ \cos\theta = -1 \quad &\text{if and only if} \quad x = ay,\ a < 0.\end{aligned}$$

Similarly, for random variables $X = aY + b$ with $a \neq 0$ and $b \in \mathbb{R}$,

$$\begin{aligned}\rho(X,Y) &= \frac{\mathrm{Cov}(aY+b, Y)}{\sqrt{\mathrm{Var}[aY+b]}\sqrt{\mathrm{Var}[Y]}} = \frac{\mathbb{E}[(aY+b)Y] - \mathbb{E}[aY+b]\mathbb{E}[Y]}{\sqrt{\mathrm{Var}[aY]}\sqrt{\mathrm{Var}[Y]}} \\ &= \frac{\mathbb{E}[aY^2] + b\mathbb{E}[Y] - a\mathbb{E}[Y]\mathbb{E}[Y] - b\mathbb{E}[Y]}{\sqrt{a^2\,\mathrm{Var}[Y]}\sqrt{\mathrm{Var}[Y]}} = \frac{a\,\mathrm{Var}[Y]}{|a|\,\mathrm{Var}[Y]} = \frac{a}{|a|}.\end{aligned}$$

This is similar to the result for $\cos\theta$. Thus $\rho(X,Y)$ is a measure of the linear dependence of X and Y. Note that we also have the converse: If $\rho(X,Y) = \pm 1$, then $X = aY + b$ for some $a \neq 0$ and $b \in \mathbb{R}$. This is due to the sharpness of the Cauchy-Schwarz inequality (double root if and only if discriminant is 0).

Proposition 7.2. *Suppose $X_1, \ldots, X_n$ are independent random variables, and let $L = \sum_{k=1}^n a_k X_k$ for some $a_k \in \mathbb{R}$. Then*

$$\mathrm{Var}[L] = \sum_{k=1}^n a_k^2\,\mathrm{Var}[X_k].$$

Proof. By independence (recall that $\operatorname{Var}[X+Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$ when $X \perp\!\!\!\perp Y$), we have

$$\operatorname{Var}[L] = \operatorname{Var}\left[\sum_{k=1}^{n} a_k X_k\right] = \sum_{k=1}^{n} \operatorname{Var}[a_k X_k] = \sum_{k=1}^{n} a_k^2 \operatorname{Var}[X_k].$$

This is the desired formula. □

Proposition 7.3. *Suppose $X_1, \ldots, X_n$ are random variables (not necessarily independent), and let $L = \sum_{k=1}^{n} a_k X_k$ for some $a_k \in \mathbb{R}$. Then*

$$\operatorname{Var}[L] = \sum_{k=1}^{n} a_k^2 \operatorname{Var}[X_k] + 2 \sum_{1 \le k < \ell \le n} a_k a_\ell \operatorname{Cov}(X_k, X_\ell).$$

Proof. First observe that

$$\mathbb{E}[L] = \mathbb{E}\left[\sum_{k=1}^{n} a_k X_k\right] = a_k \mathbb{E}[X_k],$$

so we get that

$$\operatorname{Var}[L] = \mathbb{E}[L^2] - \mathbb{E}[L]^2 = \mathbb{E}\left[\left(\sum_{k=1}^{n} a_k (X_k - \mathbb{E}[X_k])\right)^2\right].$$

Expanding the square, we obtain

$$\begin{aligned}
\operatorname{Var}[L] &= \mathbb{E}\left[\sum_{k=1}^{n}\sum_{\ell=1}^{n} a_k a_\ell (X_k - \mathbb{E}[X_k])(X_\ell - \mathbb{E}[X_\ell])\right] = \sum_{k=1}^{n}\sum_{\ell=1}^{n} a_k a_\ell \mathbb{E}[(X_k - \mathbb{E}[X_k])(X_\ell - \mathbb{E}[X_\ell])] \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} a_k a_\ell \operatorname{Cov}(X_k, X_\ell) = \sum_{k=1}^{n} a_k^2 \operatorname{Var}[X_k] + \sum_{1 \le k \ne \ell \le n} a_k a_\ell \operatorname{Cov}(X_k, X_\ell) \\
&= \sum_{k=1}^{n} a_k^2 \operatorname{Var}[X_k] + 2 \sum_{1 \le k < \ell \le n} a_k a_\ell \operatorname{Cov}(X_k, X_\ell),
\end{aligned}$$

where we have used $\operatorname{Var}[X] = \operatorname{Cov}(X, X)$. This is the desired formula. □

Corollary 7.0.1. *If $X_1, \ldots, X_n$ are pairwise uncorrelated, i.e. $\operatorname{Cov}(X_k, X_\ell) = 0$ for all $k \ne \ell$, then*

$$\operatorname{Var}\left[\sum_{k=1}^{n} a_k X_k\right] = \sum_{k=1}^{n} a_k^2 \operatorname{Var}[X_k].$$

Proposition 7.4. *Let $L_1 = \sum_{k=1}^{n} a_k X_k$ and $L_2 = \sum_{k=1}^{n} b_k X_k$ for some $a_k, b_k \in \mathbb{R}$. Then*

$$\operatorname{Cov}(L_1, L_2) = \sum_{k=1}^{n}\sum_{\ell=1}^{n} a_k b_\ell \operatorname{Cov}(X_k, X_\ell),$$

i.e. the covariance is bilinear.

Proof. Since $\mathbb{E}[L_1] = \sum_{k=1}^n a_k \mathbb{E}[X_k]$ and $\mathbb{E}[L_2] = \sum_{k=1}^n b_k \mathbb{E}[X_k]$, we have

$$\begin{aligned}\operatorname{Cov}(L_1, L_2) &= \mathbb{E}\left[\left(\sum_{k=1}^n a_k(X_k - \mathbb{E}[X_k])\right)\left(\sum_{\ell=1}^n b_\ell(Y_\ell - \mathbb{E}[Y_\ell])\right)\right] \\ &= \mathbb{E}\left[\sum_{k=1}^n \sum_{\ell=1}^n a_k b_\ell (X_k - \mathbb{E}[X_k])(X_\ell - \mathbb{E}[X_\ell])\right] \\ &= \sum_{k=1}^n \sum_{\ell=1}^n a_k b_\ell \mathbb{E}[(X_k - \mathbb{E}[X_k])(X_\ell - \mathbb{E}[X_\ell])] = \sum_{k=1}^n \sum_{\ell=1}^n a_k b_\ell \operatorname{Cov}(X_k, X_\ell),\end{aligned}$$

which is the desired result. □

7.3 Moment Generating Functions

Definition 7.2. Let X be a random variable. The *moment generating function* of X, if it exists, is

$$M_X(t) = \mathbb{E}[e^{tX}] \quad \text{for } t \in \mathbb{R}.$$

Remark. Since $e^{tX} \geq 0$, $M_X(t) \geq 0$. So the only way that $M_x(t)$ might not exist is if it is infinite.

Example 7.2.1. If X is absolutely continuous with density f_X, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx.$$

This is essentially the *Laplace transform* of f_X. Note that for $t = 0$, we have

$$M_X(0) = \mathbb{E}[e^{0X}] = \mathbb{E}[1] = 1.$$

In particular, $M_X(0)$ is always finite. However, $M_X(t)$ could be infinite for every $t \neq 0$.

Example 7.2.2. In the discrete case where X has pmf p_X, we have

$$M_X(t) = \sum_{x \in \mathcal{R}(X)} e^{tx} p_X(x).$$

Remark. Why is M_X called the moment generating function? This is because

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[\frac{t^k X^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}[X^k]}{k!}.$$

The exchange of the sum and expectation is justified by uniform convergence when $M_X(t)$ exists in an open neighborhood of 0, By the uniqueness of Taylor expansions, we obtain that

$$M_X^{(k)}(0) = \mathbb{E}[X^k] \quad \text{for } k \in 0, 1, 2, \ldots,$$

i.e. M_X generates the moments of X.

Theorem 7.1. *If a moment generating function $M_X(t)$ exists for all $t \in (-\delta, \delta)$ for some $\delta > 0$, then M_X uniquely determines the distribution of X. Moreover, $\mathbb{E}[|X|^k] < \infty$ for all $k = 1, 2, \ldots$ and*

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k \mathbb{E}[X^k]}{k!} \quad \text{for } t \in (-\delta, \delta).$$

Proof. Look this up. The idea is that if $M_X(t)$ exists for $t \in (-\delta, \delta)$, then M_X is differentiable there. □

Example 7.2.3. The contrapositive of the above theorem says that if $\mathbb{E}[|X|^k] = \infty$ for any $k \geq 1$, then then M_X cannot exist on any open neighborhood of 0. For instance, take a Cauchy random variable.

Example 7.2.4. On the other hand, $M_X(t)$ can exist for all $t \in \mathbb{R}$. Let $X \sim \text{Ber}(p)$, then

$$M_X(t) = \mathbb{E}[e^{tX}] = e^{0t}(1-p) + e^t p = 1 - p + pe^t,$$

which exists on all of $\mathbb{R}$.

Proposition 7.5. *We have the following properties of moment generating functions:*

(i) Let M_X be the mgf of X. Then $M_{aX+b}(t) = e^{tb} M_X(at)$.

(ii) Let X and Y be independent random variables with mgfs M_X existing for $t \in (-\delta_1, \delta_2)$ and M_Y existing for $t \in (-\delta_2, \delta_2)$. Then for $t \in (-\min\{\delta_1, \delta_2\}, \min\{\delta_1, \delta_2\})$, $M_{X+Y}(t)$ exists and

$$M_{X+Y}(t) = M_X(t) M_Y(t).$$

Proof. (i) We have

$$M_{aX+b}(t) = \mathbb{E}[e^{t(aX+b)}] = e^{tb}\mathbb{E}[e^{t(aX)}] = e^{tb} M_X(at).$$

(ii) We have

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t)M_Y(t)$$

by the independence of X and Y. □

7.4 Characteristic Functions

Definition 7.3. Let X be a random variable. The *characteristic function* of X, denoted by φ_X, is

$$\varphi_X(t) = \mathbb{E}[e^{itX}] \quad \text{for } t \in \mathbb{R}.$$

Remark. If X has pdf f_X, then we define the expectation of the complex-valued function e^{itX} by

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx = \int_{-\infty}^{\infty} \cos(tx) f_X(x)\, dx + i \int_{-\infty}^{\infty} \sin(tx) f_X(x)\, dx.$$

In particular, unlike the moment generating function, φ_X exists for all $t \in \mathbb{R}$. This is because

$$|\varphi_X(t)| = \left| \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx \right| \leq \int_{-\infty}^{\infty} |e^{itx} f_X(x)|\, dx = \int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

We similarly have the power series expansion

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(it)^k X^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{(it)^k \mathbb{E}[X^k]}{k!},$$

which says that $\varphi_X^{(k)}(0) = i^k \mathbb{E}[X^k]$ by the uniqueness of Taylor expansions.

Remark. Similar uniqueness results and properties also hold for characteristic functions, as for mgfs. We also have some additional results for characteristic functions.

Theorem 7.2. *Let X be a random variable. If $\mathbb{E}[|X|^N] < \infty$ for some integer $N \geq 1$, then*

$$\varphi_X(t) = \sum_{k=0}^{N} \frac{(it)^k \mathbb{E}[X^k]}{k!} + o(t^N) \quad \text{as } t \to 0.$$

Proof. Look this up. □

Theorem 7.3. *We have the following properties of characteristic functions:*

(i) Let φ_X be the characteristic function of X. Then $\varphi_{aX+b}(t) = e^{itb}\varphi_X(at)$.

(ii) Let X and Y be independent random variables. Then for all $t \in \mathbb{R}$, $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.

Proof. The proofs carry over almost identically from the mgf case. One detail to notice is that we have only shown $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B]$ for *real-valued* random variables. But this can be remedied by writing

$$\begin{aligned}\mathbb{E}[e^{itX}e^{itY}] &= \mathbb{E}[(\cos(tX) + i\sin(tX))(\cos(tY) + i\sin(tY))] \\ &= \mathbb{E}[\cos(tX)\cos(tY) - \sin(tX)\sin(tY)] + i\mathbb{E}[\cos(tX)\sin(tY) + \mathbb{E}[\sin(tX)\cos(tY)].\end{aligned}$$

These are now real-valued random variables and we can proceed as before. □

7.5 Important Inequalities

Theorem 7.4 (Bienaymé-Chebyshev-Markov inequality)**.** *Let X be a nonnegative random variable. Then for all $t > 0$, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. We can split X into two parts and get

$$\mathbb{E}[X] = \mathbb{E}[X\mathbb{1}_{X \geq t}] + \mathbb{E}[X\mathbb{1}_{X<t}] \geq \mathbb{E}[X\mathbb{1}_{X\geq t}] \geq t\mathbb{E}[\mathbb{1}_{X\geq t}] = t\mathbb{P}(X \geq t).$$

Note that $\mathbb{E}[X\mathbb{1}_{X<t}] \geq 0$ as $X \geq 0$ and we trivially have $X \geq t$ on the set $\{X \geq t\}$. □

Remark. The previous inequality says that if $\mathbb{E}[X] < \infty$, then the tail behaves like $1/t$. If instead we have $\mathbb{E}[X^2] < \infty$ and $X \geq 0$, then we can write

$$\mathbb{P}(X \geq t) = \mathbb{P}(X^2 \geq t^2) = \mathbb{P}(Y \geq t^2) \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\mathbb{E}[X^2]}{t^2}.$$

This formulation of Markov's inequality is known as *Chebyshev's inequality*. Another version is that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2} \quad \text{for all } t > 0.$$

Theorem 7.5 (Jensen's inequality)**.** *Let $h : (a, b) \to \mathbb{R}$ be a convex function, i.e. such that*

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) \quad \text{for all } 0 < \lambda < 1.$$

Then we have $h(\mathbb{E}[X]) \leq \mathbb{E}[h(X)]$.

Proof. Refer to Grimmett and Welsh. □

Example 7.3.1. The particular case $h(x) = x^2$ (note that $h''(x) \geq 0$ is sufficient to imply convexity) of Jensen's inequality yields $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, i.e. Cauchy-Schwarz.

7.6 Homework Problems

Problems #2, 3, 8, 10, 14, 17, 18, 20, 24 from Grimmett and Welsh.

Chapter 8

Nov. 19 — Limit Theorems

8.1 Convergence of Random Variables

Example 8.0.1. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of real numbers such that $a_n \to 0$ as $n \to \infty$. Then

$$\frac{a_1 + \cdots + a_n}{n} \xrightarrow[n\to\infty]{} 0$$

as well. This is known as *Cesàro convergence*. We sometimes write $a_1 + \cdots + a_n = \sum_{k=1}^{n} a_k = s_n$.

Remark. What does it mean for a sequence of random variables to converge? Let $X_1, \ldots, X_n, \ldots$ be a sequence of random variables. Each X_k is a function $X_k : \Omega \to \mathbb{R}$. One possible notion of convergence is the pointwise convergence of the X_k. But this is not very useful as it is not very likely. If the X_k are coin flips, then we get a random sequence of 0s and 1s. This probably does not converge pointwise.

Definition 8.1. A sequence $\{X_n\}_{n=1}^{\infty}$ of random variables *converges in mean-square* to X if

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0.$$

Convergence in mean-square is often denoted by $X_n \xrightarrow[n\to\infty]{\text{m.s.}} X$.

Definition 8.2. A sequence $\{X_n\}_{n=1}^{\infty}$ of random variables *converges in probability* to X if for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \epsilon) \xrightarrow[n\to\infty]{} 0.$$

Convergence in probability is often denoted by $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Example 8.2.1. Define X_n by $\mathbb{P}(X_n = 0) = 1 - 1/n$ and $\mathbb{P}(X_n = n) = 1/n$. In this case, we have $X_n \to 0$ in probability. To see this, fix $\epsilon > 0$, and we can compute that

$$\mathbb{P}(|X_n| \geq \epsilon) = \mathbb{P}(X_n = n) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0.$$

Remark. The idea behind mean-square convergence is that we have $\mathbb{E}[Y^2] = 0$ if and only if $Y = 0$ almost surely. One difference is that the X_n must have finite second moments in order to talk about mean-square convergence, whereas this is not necessary for convergence in probability.

Example 8.2.2. Define X_n the same way as in the previous example. We can compute that

$$\mathbb{E}[X_n]^2 = \frac{n^2}{n} = n,$$

which does not converge to 0 as $n \to \infty$. Thus X_n does not converge to 0 in mean-square. In particular, this means that convergence in probability does not imply convergence in mean-square.

Example 8.2.3. Now define X_n by $\mathbb{P}(X_n = 1) = 1/n$ and $\mathbb{P}(X_n = 2) = 1 - 1/n$. We can compute

$$\mathbb{E}[(X_n - 2)^2] = (1-2)^2\frac{1}{n} + (2-2)^2\left(1 - \frac{1}{n}\right) = \frac{1}{n} \xrightarrow[n\to\infty]{} 0,$$

so we can conclude that $X_n \xrightarrow[n\to\infty]{\text{m.s.}} 2$.

Proposition 8.1. *Convergence in mean-square implies convergence in probability.*

Proof. Recall that if $X \geq 0$ and $x > 0$, then by Markov's inequality we have

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[X]}{x}.$$

Applying this inequality to $|X_n - X|$ and $\epsilon > 0$ gives us

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \mathbb{P}(|X_n - X|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[(X_n - X)^2]}{\epsilon^2} \xrightarrow[n\to\infty]{} 0$$

when $X_n \xrightarrow[n\to\infty]{\text{m.s.}} X$, since $\epsilon > 0$ is fixed. This shows convergence in probability. □

8.2 Weak Law of Large Numbers

Theorem 8.1 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent random variables, each with mean μ and variance σ^2. Then*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow[n\to\infty]{\text{m.s.}} \mu.$$

Proof. Let $S_n = X_1 + \cdots + X_n$. Since $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots = \mathbb{E}[X_n] = \mu$, we have

$$\mathbb{E}\left[\frac{S_n}{n}\right] = \frac{\mu_1 + \cdots + \mu_n}{n} = \frac{n\mu}{n} = \mu.$$

Now we can compute that

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{S_n}{n} - \mu\right)^2\right] &= \mathbb{E}\left[\left(\frac{S_n}{n} - \frac{\mathbb{E}[S_n]}{n}\right)\right] = \frac{1}{n^2}\mathbb{E}[(S_n - \mathbb{E}[S_n])^2] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{k=1}^n X_k - \sum_{k=1}^n \mathbb{E}[X_k]\right)^2\right] = \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{k=1}^n (X_k - \mathbb{E}[X_k])\right)^2\right] \\
&= \frac{1}{n^2}\operatorname{Var}[S_n] = \frac{1}{n^2}\sum_{k=1}^n \operatorname{Var}[X_k] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

where $\operatorname{Var}[S_n] = \sum_{k=1}^n \operatorname{Var}[X_k]$ by independence. This implies that $S_n/n \xrightarrow[n\to\infty]{\text{m.s.}} \mu$. □

Remark. In the above proof, it is enough to assume that the X_i are uncorrelated in place of independence. It is also enough to assume that the variances $\operatorname{Var}[X_i]$ are uniformly bounded, i.e. $\operatorname{Var}[X_i] \leq M$ for every i. In this case, we can replace σ^2 with M and still get $M/n \to 0$ as $n \to \infty$.

Remark. In ergodic theory, this type of result is known as an *ergodic theorem*.

Example 8.2.4. Suppose that $X_i \sim \text{Ber}(p)$ for $i \geq 1$ are independent. Note that $\mathbb{E}[X_i] = p$. Then the weak law of large numbers says that we have

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow[n\to\infty]{\text{m.s.}} \mathbb{E}[X_1] = p = \mathbb{P}(X_1 = 1).$$

Mean-square convergence then implies convergence in probability, so this says that averaging the flips of a coin will give the probability of getting heads as the number of flips $n \to \infty$.

Remark. The (strong) law of large numbers says that we also get convergence *almost surely*.

8.3 Central Limit Theorem

Remark. The weak law of large numbers says that $S_n/n \approx \mu$, i.e. $S \approx n\mu$. But this is a first order approximation: $n\mu \in \mathbb{R}$ is a constant, but S_n is random. Observe that $S_n - n\mu$ is random and

$$\mathbb{E}[S_n - n\mu] = 0$$

if all of the X_k have the same mean. This means that $S_n - n\mu$ has mean 0 and variance

$$\text{Var}[S_n - n\mu] = \text{Var}[S_n] = n\sigma^2$$

if all of the X_k have the same variance. In particular, these calculations imply that

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

has mean 0 and variance 1. The central limit theorem says that in fact $Z_n \approx \mathcal{N}(0, 1)$.

8.4 Homework Problems

Problems #1, 2, 5, 7, 8, 11, 14, 15 from Grimmett and Welsh.