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 $\{\text{heads on the } n\text{th throw}\}$ 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

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

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] To see that convolution is associative, one can note that the sum 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.