Class Notes for Applied Probability and Statistics

Mark Siggers

ver .2015/05/26

These notes are for a upper undergraduate or first year grad course on Probability and Statistics, and are largely based on Hogg, McKean, and Craig's 'Introduction to Mathematical Statistics' (International Seventh edition) which we refer to as [1], or as 'the text'. Section numbering follows the text, and problem numbers often refer to the text. Problems within the notes are usually quite easy, checking that we know definitions, and making simple observations that we will use later. It is important to look also at the problems from the text.

We also add in a couple of applications of probability to problems in graph theory as examples of non-statistical applications of probablility. The reference for this is Alon and Spencer's 'The Probabilistic Method', (or lecture notes of the same name by Matoušek and Vondrák.)

1 Probability and Distributions

1.1 Introduction

Probabililty theory in its pure form has much of the flavour of real analysis. In this course on applied probability theory, we try to avoid this formality not by sacrificing rigour, but by avoiding exceptional cases. We generally assume things to be nice. Our goal is to get an introduction to how probablility can be applied, both in statistics and in mathematics.

Probability theory for statistics is concerned with *random experiments* - experiments that can be repeated several times, under the same conditions, and have different outcomes. They are characterised by the fact that we cannot predict the outcome of an individual experiment but can predict the frequency of the outcome over many repetitions of the experiment.

For example, an experiment might consist of tossing a coin. We cannot predict whether the outcome will be heads or tails, but if we toss the same coin 100 times, we are all going to guess that the outcome will be heads 50 times.

Would we bet on it though? Would you take the following bet? You pay

₩1000 and toss a coin 100 times. If the outcome is heads exactly 50 times, you win ₩2000.

Probably not. Would you take the bet if you win in the case that the outcome is heads between 40 and 60 times? This is the kind of question that probability theory lets us address.

In statistics we will not be really be interested in probability that a coin comes up heads, but perhaps we will be interested in the probability that a given person in a population tests positive for some disease. Our goal will be to look at the data of an experiment, estimate such a parameter, and then give some measure of how good our estimate is.

One the other hand, the application of probability to mathematics is usually a way of counting structures, and through this, determining properties that *most* of the structures have, or showing that a structure with a given property must exist.

For example, by constructing a graph by randomly adding an edge between any two vertices with some given probability, and then calculating the probabilities that the graph has small cycles or large independent sets, we show the existence of graphs of large girth and large chromatic number.

The obvious commonality in these applications is the notion of something happening randomly. This brings us to *random variables*, which are the starting point of our course.

1.2 Set Theory

We will generally consider sets of points in $\mathbb{R}$ or $\mathbb{R}^n$, such as

$$C = \{(x,y) | x \in \mathbb{R}, y = 2x\}.$$

The union and intersection of sets are denoted standardly by such notation as $C_1 \cup C_2$, $\cup_{i=1}^{k} C_i$, $\cup_{i=1}^{\infty} C_i$, $C_1 \cap C_2$, $\cap_{i=1}^{k} C_i$, and $\cap_{i=1}^{\infty} C_i$.

The empty set, or null set is often (but not always) denoted in [1] by ϕ, but I will use the more standard $\emptyset$.

Subsets are denoted $C_1 \subset C_2$ and may be equal. A set C is usually assumed to be a subset of and underlying *universe* $\mathscr{C}$, and the complement of a set C is defined as

$$C^c = \mathscr{C} - C = \{x \in \mathscr{C} \mid x \notin C\}.$$

(The notation $\overline{X}$ will play a different role.)

Recall the following basic set laws.

i) $C \cup C^c = \mathscr{C}$.

ii) $C \cap C^c = \emptyset$.

iii) $C \cup \mathscr{C} = \mathscr{C}$.

iv) $C \cap \mathscr{C} = C$.

v) $(C_1 \cup C_2)^c = C_1^c \cap C_1^c$ (Demorgan).

vi) $(C_1 \cap C_2)^c = C_1^c \cup C_1^c$.

Definition 1.2.1. The powerset $\mathscr{P}(\mathscr{C})$ of a set $\mathscr{C}$ is the set of all sets in $\mathscr{C}$. A σ-algebra of $\mathscr{C}$ is a subset $\mathcal{B} \subset \mathscr{P}(\mathscr{C})$ that contains $\mathscr{C}$, and is closed under complements, countable unions and countable intersections. A *set function* on $\mathscr{C}$ is a function $F : \mathcal{B} \to \mathbb{R}^{\text{ex}} = \mathbb{R} \cup \{\pm\infty\}$ for some σ-algebra of $\mathscr{C}$.

Note

For any set $\mathscr{C}$, the powerset $\mathscr{P}(\mathscr{C})$ is clearly a σ-algebra. A set function F defined on a σ-algebra $\mathcal{B}$ of $\mathscr{C}$ can be inocuously extended to a set function on $\mathscr{P}(\mathscr{C})$ by setting it equal to 0 on any set not in $\mathcal{B}$. So we often omit explicit mention of a σ-algebra, taking it as $\mathscr{P}(\mathscr{C})$.

Example 1.2.2. The cardinality map $|\cdot|$, acting like:

$$|\{1, 3, 5, 6, 7\}| = 5,$$

is a set function on any finite set $\mathscr{C}$. Its range is the natural numbers $\mathbb{N} \subset \mathbb{R}$.

Example 1.2.3. The area function Q is a set function on $\mathbb{R}^2$.

- $Q(\{(x, y) \mid x, y \in [0, 1]\}) = 1$.
- $Q(\{(x, y) \mid |(x, y) - 0| \leq 1\}) = 2\pi$.
- $Q(\{(x, y) \mid y = 2x\}) = 0$.
- $Q(\{(x, y) \mid x \geq 0\}) = \infty$.

The area function is really just an integral. More generally for any function $f : \mathbb{R}^n \to \mathbb{R}$ we have a set function Q_f on $\mathbb{R}^n$ defined by

$$Q_f(C) = \oiint_C f(\vec{x})\, \mathrm{d}\vec{x}.$$

Example 1.2.4. Let $Q = Q_{e^{-x}}$ then for $C = [0, \infty) \in \mathbb{R}$, we have:

$$\begin{aligned} Q(C) &= \int_0^\infty e^{-x}\, \mathrm{d}x = \lim_{N \to \infty} \int_0^N e^{-x}\, \mathrm{d}x \\ &= \lim_{N \to \infty} (-e^{-x}) \mid_{x=0}^N \\ &= \lim_{N \to \infty} (-e^{-N} + e^0) = 0 + 1 = 1 \end{aligned}$$

The *support* of a function $f : \mathscr{C} \to \mathbb{R}$ is the subset $\mathrm{Supp}(f) \subset \mathscr{C}$ defined by

$$\mathrm{Supp}(f) = \{x \in \mathscr{C} \mid f(x) \neq 0\}.$$

Observe that for the set function Q_f, $Q_f(S) = Q_f(S \cap \mathrm{Supp}(f))$ for any $S \in \mathscr{P}(\mathscr{C})$. As such, we often assume that $\mathscr{C} = \mathrm{Supp}(f)$ for some f.

Problem 1.2.5. Show that if $\mathscr{B}$ is a σ-algebra of $\mathscr{C}$, and f is a function on $\mathscr{C}$, then the family

$$\{S \cap \mathrm{Supp}(f) \mid S \in \mathscr{B}\}$$

is a σ-algebra of $\mathrm{Supp}(f)$.

Now, it is clear that for any $f : \mathbb{R}^n \to \mathbb{R}$ with finite or countable support, $Q_f = 0$. In such cases, we will consider a discrete analogue $Q_f(C) = \sum_{x \in C} f(x)$.

Example 1.2.6. Let $f : \mathbb{R} \to R$ be defined by $f(x) = (1/2)^x$ for $x \in \mathbb{Z}^+$ (positive integers) and $f(x) = 0$ otherwise. Then

$$Q_f(\{x \in \mathbb{N} \mid x < 3\}) = 1/2 + 1/4 = 3/4$$

and

$$Q_f(\mathbb{R}) = Q_f(\mathbb{Z}^+) = 1/2 + 1/4 + \cdots = 1.$$

This is our first introduction to what will become a consistant theme in the course: concepts and definitions will frequently have *continuous* and *discrete* versions.

Problems from the Text

Section 1.2: 1,5,6,8,11,14,16

1.3 The Probability Set Function

Definition 1.3.1. A *probablility set function* on a set $\mathscr{C}$ is a set function P on (some σ-algebra $\mathcal{B}$ of) $\mathscr{C}$ such that

i) $P(C) \geq 0$ for all $C \subset \mathcal{B}$.

ii) $P(\mathscr{C}) = 1$.

iii) P is countably additive: for a family $\{C_n\}_{n \in \mathbb{N}}$ of pairwise disjoint sets $C_n \subset \mathcal{B}$,

$$P(\cup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} P(C_n).$$

We are now ready to define the basic setup that we will assume throughout the course.

Definition 1.3.2. A *(random) experiment* or a *probability space* consists of a set $\mathscr{C}$, and a probability set function P defined on a as σ-algebra $\mathcal{B}$ of $\mathscr{C}$. We call $\mathscr{C}$ the *sample space*, and elements of $\mathcal{B}$ the events. Elements $x \in \mathscr{C}$ are called *outcomes*, or sometimes, viewed as singleton sets in $\mathcal{B}$, *elementary events.*

Often a probability function, expecially for finite (discrete) $\mathscr{C}$, is defined additively by its value on elementary events.

Example 1.3.3. Tossing a coin is a random experiment with two outcomes: $\mathscr{C} = \{H, T\}$. The probablility function on $\mathscr{C}$ is defined by $P(\{H\}) = 1/2$ by additivity eg.:

$$P(\{T\}) = P(\mathscr{C} - \{H\}) = P(\mathscr{C}) - P(\{H\}) = 1 - 1/2 = 1/2.$$

For elementary events like $\{H\}$, we will write $P(H)$ for $P(\{H\})$.

Example 1.3.4. Tossing two identical coins is a random experiment with three possible outcomes: $\mathscr{C} = \{HH, HT, TT\}$. The probability function is defined by $P(HH) = P(TT) = 1/4$ and $P(HT) = 1/2$. The event $C = \{HT, HH\}$ has probability $P(C) = 3/4$.

Given a random experiment, we often define events non-formally: the event $C = \{HT, HH\}$ can be described as the event that 'at least one head is tossed'. We would say 'The probability that at least on head is tossed is 3/4.' and write $P($ at least on head is tossed $) = 3/4$.

Problem 1.3.5. Tossing two non-identical coins is an experiment with four possible outcomes: $\mathscr{C} = \{HH, HT, TH, TT\}$. What is the probability that at least one head is tossed?

Problem 1.3.6. Let $\mathscr{C}$ be the set of 36 possible outcomes

$$\{(i, j) \mid i, j \in [6]\}$$

when two different dice are rolled. What is the probablility of the following events (assuming that each outcome is equally likely)?

i) $i + j = 7$

ii) $i + j$ is even

iii) $i > j$.

Example 1.3.7. A p-coin is a coin that when tossed, shows heads with probabilty p and shows tails with probability $1 - p$. Tossing a p-coin is a random experiment wtih $\mathscr{C} = \{H, T\}$ such that $P(H) = p$ and $P(T) = 1 - p$. (A 1/2-coin is called a fair coin.)

Unless otherwise stated, when we talk about events, it is always assumed that they are events of a sample space $\mathscr{C}$ with a probabilitiy set function P.

Theorem 1.3.8. *For events C, C_1 and C_2, the following hold.*

i) $P(C^c) = 1 - P(C)$.

ii) $P(\emptyset) = 0$.

iii) $C_1 \subset C_2 \Rightarrow P(C_1) \leq P(C_2)$.

iv) $0 \leq P(C) \leq 1$.

v) $P(C_1 \cup C_2) = P(C_1) + P(C_2) - P(C_1 \cap C_2)$.

Proof. All of these are pretty easy using the additivity of P. □

Now, the above theorem immediately yields the following which is known as **Bonferroni's Inequality**.

$$P(C_1 \cap C_2) \geq P(C_1) + P(C_2) - 1. \tag{1}$$

A sequence of events $\{C_n\}$ is *non-decreasing* if $C_n \subset C_{n+1}$ for each n. A sequence $\{D_n\}$ is *non-increasing* if $D_n \supset D_{n+1}$. In this case we often write $\lim_{n\to\infty} C_i$ for $\cup_{n=1}^{\infty} C_i$ and $\lim_{n\to\infty} D_i$ for $\cap_{n=1}^{\infty} D_i$.

Given a non-decreasing sequence $\{C_n\}$ of events, if we let $R_{n+1} = C_{n+1} - C_n$ for each n, then the events R_n are pairwise disjoint, and so by the additivity of P we have that

$$\begin{aligned} P(\lim_{n\to\infty} C_n) &= P(\cup_{n=1}^{\infty} C_n) = P(\cup_{n=1}^{\infty} R_n) = \sum_{n=1}^{\infty} P(R_n) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(R_i) = \lim_{n\to\infty} P(C_i). \end{aligned}$$

That is, we can interchange P and the limit. We have essentially shown the 'non-decreasing' part of the following.

Theorem 1.3.9. *Let $\{C_n\}$ be a non-decreasing or a non-increasing sequence of events. Then*

$$\lim_{n\to\infty} P(C_n) = P(\lim_{n\to\infty} C_n).$$

Problem 1.3.10. Prove the above theorem for a non-increasing sequence of events.

Using $C_n' = C_n - \cup_{i=1}^{n-1} C_i$ instead of R_n in given proof of the above theorem, we get the following.

Theorem 1.3.11 (Boole's Inequality)**.** *Let $\{C_n\}$ be a sequence of events. Then*

$$P(\cup_{n=1}^{\infty} C_n) \leq \sum_{n=1}^{\infty} P(C_n).$$

Example 1.3.12. In a experiment, you flip a coin until you get two consecutive heads or two consecutive tails. The sample space is

$$\mathscr{C} = \{HH, TT, HTT, THH, HTHH, THTT, HTHTT, THTHH, \ldots\}.$$

What is the probabilitly that the experiment ends with an H?

Letting C_i be the event that we finish with two heads in at most i flips, we get that $P(C_2) = P(\{HH\}) = 1/4$, $P(C_2) = P(\{HH, THH\}) = 1/4 + 1/8$, and in general that $P(C_n) = 1/4 + 1/8 + \ldots 1/2^n$. By Theorem , we get that the probability $C = \cup_{i=2}^{\infty} C_i$ that the experiment ends in two heads is

$$P(C) = P(\cup_{n=2}^{\infty} C_n) = \lim_{n \to \infty} (\sum_{i=2}^{n} 1/2^i) = \sum_{i=2}^{\infty} 1/2^i = 1/2.$$

There is another easy way to do the above examples. Assuming that the experiment ends, it is easy to see, by symmetry, that it is equally likely to end with H or with T, so with probability $1/2$ it ends with H. This uses conditional probability, which we will see next section, but still it must be shown that the experiment ends. Or more precisely, it must be shown that the probability that the experiment ends is 1.

Problems from the Text

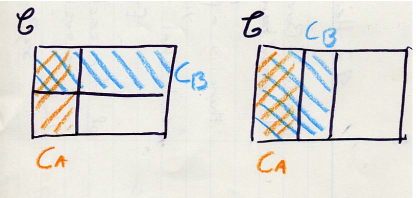
Section 1.3: 1,3,5,8,10,13,15,20

1.4 Conditional Probability

In an experiment with some event C_A of probability $1/3$, and another event C_B of probability $1/2$, does the knowledge that C_A occurs affect the probability that C_B occurs?

It can.

In the following picture let C_A be the event that a randomly placed dot in $\mathscr{C}$ is placed in the orange region and C_B be the event that a randomly placed dot in $\mathscr{C}$ is placed in the blue region.



In the first picture, knowing that event C_A happened doesn't affect the probability of event C_B. In the second picture, the fact that C_A has occured implies that event C_B **definitely** occurs.

In the first picture, the events C_A and C_B are *independent* and in the second picture they are not. Let's give this a mathematical definition.

1.4.1 Conditional probability and independence

Definition 1.4.1. For events C_1 and C_2, the *conditional probability of* C_1 *given* C_2 is

$$P(C_1 \mid C_2) = \frac{P(C_1 \cap C_2)}{P(C_2)}.$$

The events C_1 and C_2 are *independent* if $P(C_1 \mid C_2) = P(C_1)$.

Notice that if two events C_1 and C_2 are independent then we have that $P(C_1) = P(C_1 \mid C_2) = \frac{P(C_1 \cap C_2)}{P(C_2)}$ and so

$$P(C_1) \cdot P(C_2) = P(C_1 \cap C_2). \tag{2}$$

And indeed this is an alternate definition of the independence of events, and because of this, independent is sometimes called multiplicity.

Example 1.4.2. In the experiment $\mathscr{C} = \{(i, j) \mid i, j \in [6]\}$ where we roll two independent dice. We define the events $C_1 : i \leq 3$, $C_2 : j \leq 3$, and $C_3 : i + j = 8$. Intuitively, we feel that the events C_1 and C_2 should be independent, while C_3 should depend on either of them. Indeed, we see, among other things that $P(C_1) = P(C_2) = 1/2$, $P(C_3) = 5/36$, $P(C_1 \cap C_2) = 9/36 = 1/4$, and

$$P(C_3 \cap C_1) = \frac{|\{(2, 6), (3, 5)\}|}{36} = 1/18.$$

This gives the conditional probabilities,

- $P(C_1 \mid C_2) = P(C_1 \cap C_2)/P(C_2) = (9/36)/(3/36) = 1/2 = P(C_1)$,
- $P(C_2 \mid C_1) = P(C_2 \cap C_1)/P(C_1) = (1/6)/(1/3) = 1/2 = P(C_2)$, and
- $P(C_3 \mid C_1) = P(C_3 \cap C_1)/P(C_1) = (2/36)/(3/6) = 1/9 \neq 5/36 = P(C_3)$.

We conclude that C_1 is independent of C_2 and C_2 is independent of C_1, but C_3 is not independent of C_1.

Notice that C_1 and C_2 were independent of each other. This should be expected, as it is clear from (2) that independence is a symmetric relationship. Here are some other obvious facts.

i) $P(C_1 \mid C_1) = 1$.

ii) $P(C_1 \mid C_2) = P(C_1 \cap C_2 \mid C_2)$.

iii) For fixed C_2 the function $P(\cdot \mid C_2) : \mathscr{P}(\mathscr{C}) \to \mathbb{R} : C_1 \mapsto P(C_1 \mid C_2)$ is a probability set function.

iv) $P(C_1 \cap C_2) = P(C_2)P(C_1 \mid C_2) = P(C_2)P(C_2 \mid C_1)$.

This last fact can be extended to more events

$$\begin{aligned} P(C_1 \cap C_2 \cap C_3) &= P(C_1 \cap C_2) \cdot P(C_3 \mid C_1 \cap C_2) \\ &= P(C_1) \cdot P(C_2 \mid C_1) \cdot P(C_3 \mid C_1 \cap C_2) \end{aligned}$$

and used as a way to calculate the probability of an intersection of events.

Example 1.4.3. There is a bucket 10 different coloured jelly-beans. In an experiment you reach your hand into the bucket and pull out 3 jellybeans unseen. The probability of the that we pull blue, green, and red, is $1/\binom{10}{3}$. But we can compute this another way. The probability $P(C_b)$ that one of the chosen jellybeans is blue is $P(C_b) = \binom{9}{2}/\binom{10}{3}$, the probability that one of the other two is green is $P(C_g \mid C_b) = 8/\binom{9}{2}$ and the probability that the final one is red is $P(C_r \mid C_g \cap C_b) = 1/8$.

This checks out, as

$$\frac{\binom{9}{2}}{\binom{10}{3}} \cdot \frac{8}{\binom{9}{2}} \cdot \frac{1}{8} = \frac{1}{\binom{10}{3}}.$$

We have been using the notion of independence implicitly in some of our examples. In the experiment when we tossed two coins, we said the probability of the outcome, say HH, was $P(HH) = 1/4$. We were assuming that the outcome of the second toss was independent of the outcome of the first. In this case we say that the two tosses, or experiments, are *independent*.

We also assumed independence of the two dice rolls in the two dice experiment.

1.4.2 Bayes Theorem

Let the events $C_1, \ldots, C_n$ be a partition of the sample space $\mathscr{C}$; that is, assume that

i) C_i and C_j are independent for $i \neq j \in [n]$, and

ii) $\bigcup C_i = \mathscr{C}$.

The the outcome of an experiment on $\mathscr{C}$ must be in exactly one of the C_1, and so for any event C we have that

$$P(C) = \sum_{i=1}^{n} P(C \cap C_i) = \sum_{i=1}^{n} P(C \mid C_i) \cdot P(C_i). \tag{3}$$

Example 1.4.4. Consider the following experiment:

i) Flip a 1/3-coin A.

ii) If A shows heads, flip a 1/3-coin B_1; if A shows tails, flip a 1/2 coin B_2.

Let C_1 be the event that we flip B_1, and C_2 be the event that we flip B_2. Let C_H be the event that the second coin we flip shows heads.

Now it is easy to compute $P(C_H|C_1) = 1/3$, and say,

$$P(C_H) = P(C_H|C_1) \cdot P(C_1) + P(C_H|C_2) \cdot P(C_2) = (1/3 \cdot 1/3) + (2/3 \cdot 1/2) = 4/9.$$

But what is $P(C_1 \mid C_H)$? Intuitively, we see that in the computation of $P(C_H)$, 1/9 of the 4/9 came from the case when C_1 held. So $P(C_1 \mid C_H) = 1/4$.

This is exactly what Bayes Theorem says.

Theorem 1.4.5 (Bayes Theorem)**.** *Let the events $C_1, \ldots, C_n \in \mathcal{B}$ be a partition of the sample space $\mathscr{C}$, and $C \in \mathcal{B}$. Then for any $j \in [n]$,*

$$P(C_j \mid C) = \frac{P(C \cap C_j)}{\sum_{i=1}^{n} P(C \cap C_i)} = \frac{P(C_j)P(C \mid C_j)}{\sum_{i=1}^{n} P(C_i)P(C \mid C_i)}.$$

Proof. Indeed,

$$P(C_j \mid C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C \mid C_j)P(C_j)}{P(C)}.$$

Putting (3) in the bottom of the right-hand side gives the identity. □

Example 1.4.6. Plants 1, 2 and 3 produce respectively 10%, 50%, and 40% of the light bulb produced by a lightbulb company. Light bulbs made in these plants are defective with probabilities .01, .03, and .04 respectively. What is the probability that a randomly chosen defective lightbulb was produced in plant 1?

By Bayes Theorem the probabilitly is

$$\frac{.10 * .01}{(.10 * .01) + (.50 * .03) + (.40 * .04)} = \frac{1}{32}.$$

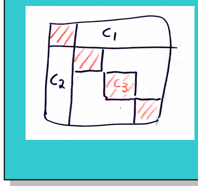
1.4.3 Mutual Independence

Definition 1.4.7. Events $C_1, \dots, C_n$ are *pairwise independent* if for all $i \neq j$, C_i and C_j are independent: $P(C_i \cap C_j) = P(C_i) \cdot P(C_j)$. They are *mutually independent* if for all $S \subset [n]$,

$$P(\bigcap_{i \in S} C_i) = \prod_{i \in S} P(C_i).$$

Problem 1.4.8. Show that a family of independent events need not be mutually independent.

Hint



Problem 1.4.9. Show that if $C_1, \dots, C_n$ are mutually independent, then so are

i) $C_1 \cup C_2$ and C_3, or

ii) $C_1^c \cap C_2$ and C_3.

Problems from the Text

Section 1.4: 6,8,11,18,23,30,34

A The Probabilistic Method

The Probabilistic Method is a technique that uses probability to prove the existence of a structure having certain properties. The main random structure we will consider is the random graph model introduced by Erdős and Rényi in 1959.

A.1 The Erdős Rényi Random Graph

Definition A.1.1 (The Erdős Rényi Random Graph)**.** The *random graph* $G_{n,p}$ on n vertices $[n]$ is constructed as follows. For each pair of vertices u, v the edge (u, v) is in $G_{n,p}$ with probability p, independent of the existence of other edges.

We view the random graph $G_{n,p}$ as a sample space, containing $2^{\binom{n}{2}}$ basic outcomes, each being a (labelled) graph on n vertices. The probability that $G_{n,p}$ is a given graph H on the vertices $[n]$ depends on the number of edges of H. If H has $|E(H)| = m$ edges, then the probability that $G_{n,p} = H$ is $p^m(1-p)^{\binom{n}{2}-m}$.

Problem A.1.2. Show that when $p = 1/2$, $P(G_{n,p} = H)$ is the same for every graph H.

In applications of the *Probabilistic Method* we show that the event S that $G_{n,p}$ has certain properties with has positive probability. In the case that $p = 1/2$, such an argument can usually be reformulated as a counting argument, showing that the number of graphs on n vertices without the property is less than $2^{\binom{n}{2}}$. Our first example is such a case, but we quickly get into examples that are hard to do without probabilistic ideas.

A.2 Ramsey Numbers

Definition A.2.1. Recall that an *independent set* in a graph is a subset of the vertices which induces no edges, and a *clique* in a graph is a subset of the vertices which induces a complete graph. The *ramsey number* $R(i, k)$ is the minimum number of vertices n such that any graph G on n vertices contains either a *independent set* of size i or a clique of size k.

Ramsey showed in 1933 that $R(i, k)$ exists for all i and k, but finding the actual value of $R(i, k)$ is notoriously difficult. The current bounds for the diagonal case $i = k$ are approximately

$$\sqrt{2}^k < R(k, k) < 4^k,$$

and we only know $R(k, k)$ exactly when $k \leq 4$.

We use the Probabilistic Method to get the above lower bound.

Theorem A.2.2. *For $k \geq 3$, $R(k,k) > 2^{k/2-1}$.*

Proof. Let $n \leq 2^{k/2-1}$. The idea is to show that in the random graph $G = G_{n,1/2}$, the probability of the event: 'there is a clique of size k or an independent set of size k' is strictly less than one, so that there exists a graph with no such substructure.

Indeed, the probability that there is a clique of size k on a given set of k vertices in G is $(2)^{-\binom{k}{2}}$ and there are $\binom{k}{2}$ such sets, so using the subadditivity of probability for non-independent events, (part (v) of Theorem 1.3.8) the probability of a clique of size k is at most $\binom{n}{k}2^{-\binom{k}{2}}$. Similarily, the probablility of an independent set of size k is at most $\binom{n}{k}2^{-\binom{k}{2}}$, and so , again by subadditivity, the probability of either a clique or an independent set of size k is less than $2\binom{n}{k}2^{-\binom{k}{2}}$.

But then

$$\begin{aligned} n < 2^{k/2-1} \quad &\Rightarrow \quad n^k < 2^{\frac{k^2-k}{2}} = 2^{\binom{k}{2}} \\ &\Rightarrow \quad 2\binom{n}{k} < 2 \cdot \frac{n^k}{k!} < n^k < 2^{\binom{k}{2}} \\ &\Rightarrow \quad 2\binom{n}{k}2^{-\binom{k}{2}} < 1. \end{aligned}$$

So if $n < 2^{k/2-1}$, there exist graphs on n vertices with no clique or independent set of size k. Thus $R(k,k) > 2^{k/2-1}$. □

A.3 Erdős-Ko-Rado

A family $\mathcal{F} \subset \binom{[n]}{k}$ of k-element subsets of $[n]$, is *intersecting* if every pair $A, B \in \mathcal{F}$ of sets in the family have non-empty intersection: $|A \cap B| > 1$.

The family $\mathcal{F}_0 = \{A \in \binom{[n]}{k} \mid 1 \in A\}$ of all sets containing a given element, is clearly intersecting, and has size $\binom{n-1}{k-1}$. The following theorem shows that no intersecting family can be bigger.

Theorem A.3.1 (Erdős-Ko-Rado)**.** *Let $\mathcal{F} \subset \binom{[n]}{k}$ be intersecting. Then*

$$|\mathcal{F}| \leq \binom{n-1}{k-1}.$$

Further, it is known that the only intersecting families even close to this bound are isomorphic to a subfamily of $\mathcal{F}_0$. The proof of this 'Further' bit is tricky. We give now a beautiful probabilistic proof of the above theorem.

Proof. Let $A_s \in \binom{[n]}{k}$ be the set of k consecutive integers from s to $s+k-1$ (modulo n). It is easy to see that an intersecting family $\mathcal{F}$ can contain at most k of these special sets A_s.

For any permutation σ of $[n]$ we also have that

$$\sigma(A_s) := \{\sigma(x) \mid x \in A_s\}$$

is in $\mathcal{F}$ for at most k values of s, so for a random choice of s the probability that $\sigma(A_s)$ is in $\mathcal{F}$ is at most k/n.

But a random choice of σ and s is a random choice of a set in $\binom{[n]}{k}$ so the probability that $\sigma(A_s)$ is in $\mathcal{F}$ is exactly $|\mathcal{F}|/\binom{n}{k}$.

Thus $k/n \geq |\mathcal{F}|/\binom{n}{k}$ which yields

$$|\mathcal{F}| = \frac{k}{n}\binom{n}{k} = \binom{n-1}{k-1}.$$

□

A.3.1 Colourings

Recall that K_b is the clique on b vertices.

Problem A.3.2. Let $R^*(b,r)$ be the minimum number of vertices in a graph G such that for any 2-colouring of the edges there is a blue copy of K_b or a red copy of K_r in G. Show that $R^*(b,r)$ is the ramsey number $R(b,r)$.

As in our proof for the lower bound for $R(k,k)$ we can view the random graph $G_{n,p}$ as a clique K_n with randomly coloured edges. Using a random colouring of $[n]$ show the following.

Problem A.3.3. Let $\mathcal{F}$ be a (not-necessarily intersecting) subfamily of $\binom{[n]}{k}$, of size $|\mathcal{F}| \leq 2^{k-1}$. Show that there is a 2-colouring of $[n]$ such that no set in $\mathcal{F}$ is monochromatic (i.e. for no set is every element the same colour).

1.5 Random Variables

Definition 1.5.1. A *random variable* or *RV* is a real function

$$X : \mathscr{D} \to \mathbb{R}$$

on the sample space $\mathscr{D}$ of some experiment. The image $\mathscr{C} = X(\mathscr{D})$ is called the *space* of X.

Example 1.5.2. In an experiment, we flip 100 fair coins. So the sample space is $\mathscr{D} = \{H, T\}^{100}$. Let X be the random variable such that counts the number of heads in an outcome of $\mathscr{D}$. Then $\mathscr{C} = X(\mathscr{D}) = \{0, 1, 2, \ldots, 100\}$.

The random variable is used to define a new probablity space on $\mathscr{C} = \{0, 1, 2, \ldots, 100\}$, which is usually a bit easier to work with than the original probability space on $\mathscr{D} = \{H, T\}^{100}$. We now define the probabililty set function of the new space.

Definition 1.5.3. The *cumulative distribution function* or *cdf* of a random variable X is the function $F_x : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = P(X \leq x).$$

Note

The *distribution* of a probability space is a list of the probabilities of each possible outcome. Its definition is slightly different depending on whether the sample space is countable or continuous, but is closely related to the probability set function.
Restricting to the probability space defined by a random variable, it will be defined, in the upcoming sections, as a 'pmf' or 'pdf', which will be (essentially) defined by the cdf.

Because of this, we casually refer to the cdf, or even an RV itself, as a distribution.

Problem 1.5.4. Show that the cdf of a random variable is a probability set function on its space.

Problem 1.5.5. In the above example what is the probability $P(2 \leq X \leq 30)$?

Example 1.5.6. Continuing the above example, we have that $F_X(0) = \left(\frac{1}{2}\right)^{100} = F_X(100)$, that $F_X(1) = F_X(0) + \binom{100}{1}\left(\frac{1}{2}\right)^{100}$ and that in general, for $x \in \mathscr{C}$,

$$F_X(x) = \left(\frac{1}{2}\right)^{100} \sum_{i=0}^{x} \binom{100}{i}.$$

Observe also that we have such values as $F_X(-3) = 0,\ ,\ F_X(499) = 1$, and $F_X(2.3) = F_X(2)$.

Problem 1.5.7. What is the cdf of a random variable X that counts the number of heads in an experiment in which 100 p-coins are tossed?

Problem 1.5.8. In the two dice experiment with the sample space $\mathscr{D} = \{(x, y) \mid x, y \in [6]\}$ of 36 equally likely outcomes, let $X : \mathscr{C} \to \mathbb{R}$ be the random variable defined by $X((x, y)) = x + y$. Find $F_X(4)$.

Example 1.5.9. Let X be the identity on the sample space $\mathscr{C} = [0, 1]$ in which each outcome is equally likely. Then $F_X(x) = P(X \leq x) = x$, for $x \in [0, 1]$. (Also $F_X(x) = 0$ if $x < 0$ and $F_X(x) = 1$ if $x > 1$.)

A random variable X is *discrete* if its sample space $\mathscr{C}$ is finite or countable. Examples 1.5.2 and 1.5.6 have discrete RVs while the RV in Example 1.5.9 is not discrete. The treatment of discrete and non-discrete RVs is a little different, and we will consider them separately in the next two sections. But before we do this, we make a couple more observations about the cdf.

Theorem 1.5.10. *Let F be the cdf of an RV. Then*

i) $a < b \Rightarrow F(a) \leq F(b)$,

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

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

iv) $\lim_{x\to a^+} F(x) = F(a)$.

Problem 1.5.11. Prove the above theorem.

Problem 1.5.12. Show that $\lim_{x\to a^-} F(x) = F(a)$ need not be true.

Problem 1.5.13. For an event $B \subset \mathscr{D}$, the *indicator random variable* $I_B : \mathscr{D} \to [0, 1]$, is defined by

$$I_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise.} \end{cases}$$

Show that $P(I_B = 1) = P(B)$.

1.6 Discrete Random Variables

Recall that a random variable X is discrete if it has a countable sample space $\mathscr{C}$. For such a variable one can talk of the probability of a given outcome $x \in \mathscr{C}$.

Definition 1.6.1. For a discrete random variable X, the *probability mass function* or *pmf* of X is the function $p_X : \mathbb{R} \to [0, 1]$ defined by

$$p_X(x) = P(X = x).$$

Example 1.6.2. Let X be the random variable that counts the number of flips of a fair coin you make until one coin shows up heads. The sample space $\mathscr{C}$ of X is the positive integers. Then we have, for example, $p_X(1) = 1/2$, $p_X(2) = 1/4$, $p_X(n) = 1/2^n$, and $p_X(1/2) = 0$.

Problem 1.6.3. Show that for the pmf p of a discrete RV X, $\sum_{x\in\mathscr{C}} p(x) = 1$.

Example 1.6.4. Let X be the random variable from Problem 1.5.7 that counts the number of heads showing up when we a p-coin 100 times. Then

$$p_X(37) = \binom{100}{37} p^{37} q^{63} = F_X(37) - F_X(36).$$

This exhibits a fundamental relationship between the cdf F_X and the pmf p_X of a discrere RV:

$$F_X(x) = \sum_{i\in\mathscr{C}, i\leq x} p_X(i).$$

Problem 1.6.5. Prove this. Show that it need not hold for non-discrete RVs. (Hint: What is p_X when X is the RV from Example 1.5.9?)

The problem above exhibits the main difference between discrete and non-discrete random variables. The pmf may be trivial for non-discrete RVs. In the next section we will define an analogue of the pmf for certain nice non-discrete RVs. Before we do this though, we talk about transformations of random variables.

Problems from the Text

Section 1.5: 2,3,8,9

1.6.1 Transformations of Discrete RV

Often we will define one random variable as a function of another.

Example 1.6.6. Let X be an RV with space $\mathscr{C} = \{\pm 1, \pm 2, \ldots, \pm 5\}$, and $p_X(x) = 1/10$ for all $x \in \mathscr{C}$

Let $Y = X^2$. Then Y is an RV with space $\mathscr{D} = \{1, 4, 9, 16, 25\}$. For each $y \in \mathscr{D}$ we have that

$$\begin{aligned} p_Y(y) &= P(Y = y) \\ &= P(X^2 = y) \\ &= P(X \in \{\pm\sqrt{y}\}) \\ &= p_X(-\sqrt{y}) + p_X(\sqrt{y}) \end{aligned}$$

So $p_Y(y) = 1/5$ for all $y \in \mathscr{D}$.

It is easy to see that in general, if $Y = g(X)$ then

$$p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x).$$

If g is one-to-one, this simplifies to

$$p_Y(y) = p_X(g^{-1}(x)).$$

Problem 1.6.7. Show that if $Y = g(X)$ for monotone strictly increasing g then $F_y(g(x)) = F_x(x)$. What can we say if g is decreasing?

Problems from the Text

Section 1.6: 1,2,3,5,7,10

1.7 Continuous Random Variables

Recall that if an RV X is not discrete, its pmf $p_X(x) = P(X = x) = \lim_{\varepsilon \to 0}(F_X(x) - F_X(x - \varepsilon))$ may be identically 0. Indeed this is the situation we want.

Definition 1.7.1. A random variable X is *continuous* if its cdf F_X is a continuous function on $\mathbb{R}$.

For such functions p_X is identically 0.

Definition 1.7.2. A random variable X is *absolutely continuous* if

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

for some function f_X. Then function f_X is called the *probability density function* or *pdf* of X.

We will never consider RVs that are continuous but not absolutely continuous (though the exist); so **any time we say an RV X is continuous, we will assume that it has a pdf f_X.** It then follows by the fundamental theorem of calculus that

$$f_X(x) = \frac{d}{dx} F_X(x).$$

The pdf of a continuous RV is the analogue of the pmf of a discrete RV.

Problem 1.7.3. Show that for a continuous RV X,

$$P(a < X < b) = \int_a^b f_X(t)\, dt.$$

Example 1.7.4. Recall the RV X from Example 1.5.9 that had cdf $F_X(x) = x$ for all $x \in [0,1]$. Its pdf is the derivative

$$f_x(x) = F_X'(x) = \frac{d}{dx}(x) = 1.$$

We say that an RV X (or its sample space $\mathscr{C}$) has a *uniform distribution* if the pdf (or pmf) is constant on its support. The RV X from the above example is said to have the *standard* uniform distribution. This is denoted $X \sim \mathrm{Unif}([0,1])$. The RV X from Example 1.6.6 is a uniformly distributed discrete random variable.

> Note
>
> Given a sample space $\mathscr{C}$, we sometimes say that an event is 'chosen at random' to mean we an event is chosen according to a uniform distribution.

Problem 1.7.5. Find the pdf of the uniformly distributed random variable $X \sim \mathrm{Unif}([-1,1])$ on the space $\mathscr{C} = [-1,1]$.

Problem 1.7.6. Show that for a uniformly distributed space $\mathscr{C} \subset \mathbb{R}^n$ the probability of an event C is

$$P(C) = \frac{\oiiint_C 1\,dx}{\oiiint_{\mathscr{C}} 1\,dx}.$$

That is, show that the probability of an event is proportional to its volume.

Example 1.7.7. Let a point (x,y) be chosen randomly from $\mathscr{C} = \{(x,y) \mid x^2 + y^2 < 1\}$ and let X be its distance from $(0,0)$. By definition $\mathscr{C}$ is uniformly distributed, but $\mathscr{C}$ is not the space of X. (The space of X is $[0,1]$.) We observe that X is not uniformly distributed.

Indeed, its cdf is $F_X(t) = P(X \leq t)$ which is the area of the event $\{(x,y) \mid x^2 + y^2 \leq t\}$ over the area of $\mathscr{C}$ (which is π.) So

$$F_X(x) = 1/\pi \int_0^x 2\pi t\,dt = \int_0^x 2t\,dt = x^2$$

for $x \in [0,1]$. And so $f_X(x) = \frac{d}{dx}x^2 = 2x$, which is not constant.

1.7.1 Transformations of Continuous RVs

Now let $Y = X^2$ be a transformation of X from the above example. So $Y = g(X)$ where the function $g(x) = x^2$ is monotone strictly increasing on the space $(0,1]$. It is tempting to follow the discrete case and say that the pdf of f is

$$f_Y(y) = f_X(g^{-1}(y)) = 2\sqrt{y}.$$

But this is not true! Indeed, the cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}),$$

but this is $F_Y(y) = \sqrt{y}^2 = y$ and so the pdf is $f_Y(y) = \frac{d}{dy}y = 1$.

Of course! A transformation is just a change of variables from calculus. In general, differentiating the above equation $F_Y(y) = F_X(g^{-1}(y))$ with respect to y, we get, by the chain rule,

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y) = f_X(g^{-1}(y)) \cdot \frac{dx}{dy}.$$

Indeed, this is what we found:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{dx}{dy} = 2\sqrt{y} \cdot \frac{d}{dy}\sqrt{y} = 2\sqrt{y} \cdot \frac{1}{2\sqrt{y}} = 1.$$

We have proved the following theorem (in the case that g is monotone increasing).

Theorem 1.7.8. *Let X be a continuous RV with pdf $f_X(x)$ and let $Y = g(X)$ where g is one-to-one and differentiable on the support of X. Then the pdf of Y is*

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |J|$$

where $J = \frac{d}{dy}g^{-1}(y)$, for y in the support $\{g(x) \mid x \in \operatorname{Supp} X\}$ of Y.

Problem 1.7.9. Where in the proof are we using that fact that g is monontone increasing?

The value $J = \frac{d}{dy}g^{-1}(y)$ is called the *Jacobian* of the transformation g. In other books it may be called the Jacobian of g^{-1}.

Example 1.7.10. Where $X \sim \text{Unif}((0,1))$ let $Y = -2\log X$. So Y has support $(0,\infty)$. The transformation $h : X \to Y : x \mapsto -2\log x$ is one-to-one with inverse $h^{-1}(y) = e^{-y/2}$, so it has Jacobian

$$J = \frac{d}{dy}e^{-y/2} = -\frac{1}{2}e^{-y/2}.$$

The pdf of Y is thus

$$f_Y(y) = f_X(e^{-y/2}) \cdot |J| = 1 \cdot \frac{1}{2}e^{-y/2} = \frac{1}{2}e^{-y/2}$$

on the support of Y.

Problems from the Text

Section 1.7: 1,2,5,6,8,9,10

1.8 Expectation of a Random Variable

Definition 1.8.1. For a random variable X, the *expected value* or *expectation* of X is

$$E(X) = \sum_{x \in \mathscr{C}} x p_X(x)$$

or

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

depending on whether X is discrete or continuous.

Note

Technically, being an integral or possibly infinite sum, the expectation need not always exist for an RV. And indeed, we should insist that the sum/integral is absolutely convergent so that the expectation is independent of an ordering of the sample space. But this is not an issue for all RVs that we consider. In theorems dealing with expectation, we will implicitly assume sufficiently strong convergence.

Example 1.8.2. The expected value when you roll a die is $(1+2+\cdots+6)/6 = 3.5$.

Example 1.8.3. Let X be the distance from $(0,0)$ of a randomly chosen point in the unit circle $S = \{(x,y) \mid x^2 + y^2 \leq 1\}$. What is $E(X)$?

Using that the pdf of X is $f(x) = 2x$ (from Example 1.7.7), we get that

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot 2x\, dx = \int_0^1 x \cdot 2x\, dx \\ &= 2\int_0^1 x^2\, dx = 2\left(1/3x^3\right)_0^1 = 2/3 \end{aligned}$$

Now, ignoring issues of convergence (dealt with in Theorem 1.8.1 of [1]) we see that for any transformation $Y = g(X)$ of an RV X we have

$$\begin{aligned} E(Y) &= \sum_{\mathscr{C}_y} y p_Y(y) = \sum_{\mathscr{C}_y} y \sum_{g(x)=y} p_X(x) \\ &= \sum_{\mathscr{C}_y} \sum_{g(x)=y} y p_X(x) = \sum_{\mathscr{C}_y} \sum_{g(x)=y} g(x) p_X(x) \\ &= \sum_{\mathscr{C}_x} g(x) p_X(x). \end{aligned}$$

The following tool is immediate from the above using the linearity of sums and integrals.

Theorem 1.8.4 (Linearity of Expectation)**.** *If* $Y = k_1 g_1(X) + k_2 g_2(X)$ *then* $E(Y) = k_1 E(g_1(X)) + k_2 E(g_2(X))$.

Problem 1.8.5. Prove Theorem 1.8.4.

Problem 1.8.6. Where $Y = X^2$ for X from Example 1.8.3, what is $E(Y)$? (Make a guess before you compute it. What should it be?)

Problem 1.8.7. Let I_C be the indicator variable (see Problem 1.5.13) for an event $C \in \mathscr{C}$. Show that $E(I_C) = P(A)$.

Problem 1.8.8. Let X count the number of heads that show up when n independent p-coins are flipped. Find $E(X)$.

Problem 1.8.9. Let v be a randomly chosen vertex in $G_{n,p}$. What is the expected degree $E(\deg(v))$ of v.

Problems from the Text

Section 1.8: 3,4,6,7,8,9,11

1.9 Mean Variance and Moments

Given a random variable X, we will be interested in the expected value of various functions of X. Certain ones get special names and notation. The expected value of X is also called the *mean* $\mu = E(X)$ of X. Then *variance* of X is

$$\sigma^2 = \mathrm{Var}(X) = E\left((X - \mu)^2\right).$$

Expanding the square in expression on the right, and using the linearity of expectation, we get that

$$\begin{aligned} \sigma^2 &= E\left(X^2 - 2\mu X + \mu^2\right) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(x^2) - \mu^2 \end{aligned}$$

Then positive square root σ of the variance is called the *standard deviation* of X.

Example 1.9.1. Let X have pdf $f(x) = \frac{1}{2}(x+1)$ for $x \in [-1, 1]$. Find μ and σ^2.

We get

$$\begin{aligned} \mu = E(X) &= \int_{-1}^{1} x f(X)\, dx = \frac{1}{2}\int_{-1}^{1} x^2 + x\, dx \\ &= \frac{1}{2}\left(\frac{1}{3}(1+1)\right) = \frac{1}{3}, \end{aligned}$$

and

$$\begin{aligned}\sigma^2 = E(X^2) - \mu^2 &= \frac{1}{2}\int_{-1}^{1} x^3 + x^2\,dx - \frac{1}{9}\\ &= \frac{1}{2}\left(\frac{1}{4}(1+1) + \frac{1}{3}(1+1)\right) - \frac{1}{9}\\ &= 17/36.\end{aligned}$$

Problem 1.9.2. In terms of $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$, what is $\mathrm{Var}(X - Y)$?

1.9.1 The moment generating function

The mean $\mu_X = E(X)$ has yet another name. It is also called the *first moment* of X. And the value $E(X^2)$, used in the compuation of the variance, is called the *second moment* of X. In general $E(X^n)$ is the n^{th} *moment* of X. The 0^{th} moment is $E(1) = 1$.

Letting the moments be the coefficients of an exponential generating function:

$$M_X(t) = E(1) + tE(X) + t^2\frac{E(X^2)}{2!} + t^3\frac{E(X^3)}{3!} + \ldots,$$

we get by the linearity of expectation that

$$M_X(t) = E(1 + (tX) + \frac{(tX)^2}{2!} + \ldots) = E(e^{tx}).$$

So we have that the n^{th} moment $E(X^n)$ of X is also denoted $M_X^{[n]}(t)$.

From the cdf of an RV, one can compute the moments, and so find the moment generating function. On the other hand, we know from an analysis class, that the power series of a function expanded on an open interval around a point, is uniqely defined, so the moment generating function uniquely defines the moments of a distribution. The following theorem takes this one step further and asserts that from the moments, we can recover the cdf. The proof of this is beyond our scope, (and beyond the scope of the text).

Theorem 1.9.3. *If $M_X(t) = M_Y(t)$ on some open interval around t, then $F_X(z) = F_Y(z)$ for all z.*

This function $M_X(t) = E(e^{tx})$ is called the *moment generating function* or *mgf* of X. As the taylor expansion of a function about a point does not always converge, not all RVs necessarily have mgfs, and indeed the text gives examples of RVs for which the mgf does not exist. But the mgf does exist for many RVs that we will consider, and it will become a useful tool.

Problems from the Text

Section 1.9: 1,2,3,5,6,18,23

1.10 Important Inequalities and Bounds

We finish the chapter with some basic inequalities.

1.10.1 Markov's Inequality

Theorem 1.10.1 (Markov's Inequality)**.** *For any RV X, any non-negative function u of X and any constant c, the following are both true.*

- $P(X \geq c) \leq E(X)/c$
- $P(u(X) \geq c) \leq E(u(X))/c$

Proof. The first statement is simply a special case of the second, so we prove just the second. We prove it in the case that X is continuous. The proof in the discrete case is essentially the same.

Let $A = \{x \mid u(x) \geq c\}$. (Recall that A^c is its complement.) Then

$$\begin{aligned} E(u(X)) &= \int_{-\infty}^{\infty} u(x) f_X(x)\, dx \\ &= \int_A u(x) f_X(x)\, dx + \int_{A^c} u(x) f_X(x)\, dx \\ &\geq \int_A u(x) f_X(x)\, dx \\ &\geq c \int_A f_X(x)\, dx = cP(x \in A) = cP(u(x) \geq c). \end{aligned}$$

The inequality follows. □

Markov's inequality is crude. Indeed if $X \sim \text{Unif}([0,4])$ then $E(X) = 2$ and taking $c = 1$ the inequality says $P(X \geq 1) \leq 2$. We could certainly give a better bound. However, the inequality is incredibly useful due to its universality.

1.10.2 Chebyshev's Inequality

Corollary 1.10.2 (Chebyshev's Inequality)**.** *Let X be an RV, then for every $\varepsilon, k > 0$, the following, clearly equivalent statements, all hold.*

i) $P(|X - \mu| \geq k\sigma) \leq 1/k^2$

ii) $P(|X - \mu| < k\sigma) > 1 - 1/k^2$

iii) $P(|X - \mu| < \varepsilon) > 1 - \sigma^2/\varepsilon^2$

Proof. Applying Markov wih $u(x) = (X-\mu)^2$ and $c = k^2\sigma^2$ gives

$$P\left((X-\mu)^2 \geq k^2\sigma^2\right) \leq \frac{E\left((X-\mu)^2\right)}{k^2\sigma^2} = \frac{1}{k^2}$$

□

In Problem 1.9.3 of the text, you were asked to find $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$ for an RV X with pdf $f(x) = 6x(1-x)$. Compare this with the quick bound we can now get without even computing μ and σ:

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) > 1 - 1/4 = 3/4$$

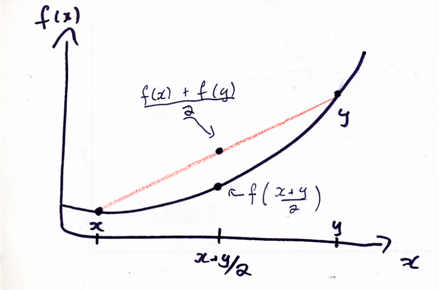
1.10.3 Jensen's Inequality

You have probably seen the next inequality several times, and proved it in a linear algebra class. It won't hurt to see it again. We state it without proof.

Definition 1.10.3. A function f is *convex* on an interval $I = [a,b]$ if for all $x, y \in I$ and all $n > 1$,

$$f(x/n + y\frac{n-1}{n}) < f(x)/n + f(y)\frac{n-1}{n}.$$

The following picture for the case $n = 2$ show that this definition of convexity agrees with the definition for a continuous function that it is convex if the second derivative is positive.



Theorem 1.10.4 (Jensen's Inequality)**.** *If f is convex, then*

$$f\left(\frac{x_1 + x_2 + \ldots x_n}{n}\right) \leq \frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n}.$$

For any RV X, this means that

$$f(E(X)) \leq E(f(X)).$$

Example 1.10.5. The function x^2 is convex so $E(X)^2 < E(X^2)$. Thus $\text{Var}(X) = E(X^2) - E(X)^2$ is non-negative.

Problem 1.10.6. Sometimes the mean of a set of numbers $\{x_1, \ldots, x_n\}$ is called the *arithmetic mean* $AM = \frac{1}{n}\sum x_i$, distinguishing it from the *geometric mean* $GM = (\prod x_i)^{1/n}$ and the *harmonic mean* $HM = (\frac{1}{n}\sum \frac{1}{x_i})^{-1}$.

Using that $-\log x$ is convex, use Jensen's Inequality to show that for any set $\{x_1, \ldots, x_n\}$ of positive numbers, $HM \leq GM \leq AM$.

1.10.4 'Compoud interest' and Stirling's Formula

Recall from your financial math class that investing ₩1000 at .05 interest compounded continuously yield a real yearly return of

$$\lim_{n\to\infty} 1000(1 + .05/n)^n = 1000e^{.05}.$$

This same inequality $1 + x \leq e^x$, is often used with probabilities:

$$(1-p)^n \leq e^{-np}.$$

I call this the *compound interest* bound. We can also get a lower bound

$$e^{-2np} < e^{-np(1-p/2)/(1-p)} < (1-p)^n$$

which we use for probabilities $p < 1/2$.

In applications we will often have to bound $n!$. Often it is enough to write

$$(n/2)^{(n/2)} \leq n! \leq n^n,$$

but we get the following from *stirling's formula*

$$(n/e)^n \leq n! \leq ne(n/e)^n.$$

To bound $\binom{n}{k}$ we can usually use

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k < n^k.$$

Problems from the Text

Section 1.10: 2,3,4,6

2 Multivariate Distributions

In Example 1.5.2 we considered the experiment of tossing 100 p-coins and let Y be the random variable counting the number of heads. The experiment can be viewed as a set of 100 random variables $X_1, \ldots, X_n$ each having the pmf of a p-coin. In this context we can view Y as a function $Y = \sum_{i=1}^{100} X_i$ of the multivariate distribution $\mathbf{X} = (X_1, \ldots, X_{100})$. Lets go into more detail. .

2.1 Distributions of Two Random Variables

Definition 2.1.1. A *random vector* $\mathbf{X} = (X_1, \ldots, X_n)$ is a set of RVs on a sample space $\mathscr{D}$. The *space* of $\mathbf{X}$ is

$$\mathscr{C} = \{(X_1(c), X_2(c), \ldots, X_n(c)) \mid c \in \mathscr{C}\}.$$

Example 2.1.2. Where $\mathscr{D}$ are people in a sample population, $\mathbf{X} = (\text{Height}, \text{Weight}, \text{Age})$ is a random vector.

Example 2.1.3. The random graph $G_{n,p}$ can be viewed as a random vector consisting of $\binom{n}{2}$ RVs X_e each with the distribution of a p-coin, one for each possible edge e on the vertices $[n]$.

We extend many of our definitions for RVs to random vectors. For most definitions the extension from two variables to arbitrarily many is trivial. For those that it isn't, we will revisit them later for more than two variables.

Definition 2.1.4. The *(joint) cdf* of $\mathbf{X} = (X_1, X_2)$ is

$$F_{\mathbf{X}}(\boldsymbol{x}) = F_{X_1,X_2}(x_1, x_2) = P\left((X_1 \leq x_1) \text{ and } (X_2 \leq x_2)\right).$$

The *(joint) pmf* for discrete $\mathbf{X}$ is

$$p_{\mathbf{X}}(\boldsymbol{x}) = p_{X_1,X_2}(x_1, x_2) = P\left((X_1 = x_1) \text{ and } (X_2 = x_2)\right).$$

The *(joint) pdf* for continuous $\mathbf{X}$ is a function $f_{\mathbf{X}}$ such that

$$F_{\mathbf{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{\mathbf{X}}(t_1, t_2)\, dt_1\, dt_2,$$

so almost everywhere we have

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{\partial^2 F_{\mathbf{X}}(x_1, x_2)}{\partial x_1 \partial x_2}.$$

Example 2.1.5. Let $f_{\boldsymbol{x}}(x_1, x_2) = 6x_1^2 x_2$ for $x_i \in (0,1)$ be the joint pdf of $\mathbf{X} = (X_1, X_2)$. Then $P(1/4 < X_1 \leq 3/4, 0 < x_2 < 2)$ is

$$\int_{\frac{1}{4}}^{\frac{3}{4}} \int_0^1 6x_1^2 x_2 \,\mathrm{dx}_2\, \mathrm{dx}_1 \quad = \quad \int_{\frac{1}{4}}^{\frac{3}{4}} 3x_1^2\, \mathrm{dx}_1 = [x_1^3]_{\frac{1}{4}}^{\frac{3}{4}} = 13/32$$

From the joint pmf of a random vector, we can isolate the *(marginal) pmf* of any one component RV as follows,

$$p_{X_1}(x_1) = \sum_{x_2} p_{\mathbf{X}}(x_1, x_2),$$

where $\sum_{x_2}$ is over all x_2 such that $(x_1, x_2) \in \mathscr{C}$.

The marginal pdf of component of a continous random vector is defined analogously.

Problem 2.1.6. Find the marginal pdf of X_1 for the joint distribution $f_{\boldsymbol{x}}(x_1, x_2) = 6x_1^2 x_2$ from the above example.

The following generalisation of Theorem 1.8.4 to random vectors is a key tool in saying anything of substance in statistical inference or with the probabilistic method.

Theorem 2.1.7 (Additivity of Expectation)**.** *Let* $\mathbf{X} = (X_1, \dots, X_n)$ *be a random vector and* $k_1, \dots, k_n$ *be real numbers. Then*

$$E(\sum k_i X_i) = \sum k_i E(X_i).$$

Proof. We do the continuous case for a vector of two RVs:

$$\begin{aligned} E(k_1 X_1 + k_2 X_2) &= \int_{\mathbb{R}} \int_{\mathbb{R}} (k_1 x_1 + k_2 x_2) f_{\mathbf{X}}(x_1, x_2) \, \mathrm{dx}_2 \, \mathrm{dx}_1 \\ &= k_1 \int \int x_1 f_{\mathbf{X}}(x_1, x_2) \, \mathrm{dx}_2 \, \mathrm{dx}_1 + k_2 \int \int x_2 f_{\mathbf{X}}(x_1, x_2) \, \mathrm{dx}_2 \, \mathrm{dx}_1 \\ &= k_1 E(X_1) + k_2 E(X_2) \end{aligned}$$

$\square$

We can talk of the *expected value of a random vector.* It is simply the vector

$$E(\mathbf{X}) = (E(X_1), \dots, E(X_n))$$

of expected values of its components. To find the expected value of a random vector on must find the expected value of the components. To find $E(X_1)$, or more generally $E(g(X_1))$, we can get the marginal distribution of X_1, and then find the expected value as in the previous chapter. Or we can find it directly:

Example 2.1.8. Where $f_{\mathbf{X}}(x_1, x_2) = 6x_1^2 x_2$, the expected value of X_1^2 is

$$\begin{aligned} E(X_1^2) &= \int_0^1 \int_0^1 x_1^2 \cdot 6x_1^2 x_2 \, dx_1 \, dx_2 \\ &= \frac{1}{2} \int_0^1 6x_1^4 \, dx_1 = \frac{6}{10} \end{aligned}$$

Problem 2.1.9. Find $E(X_1^2)$ in the above example by using the marginal distribution of X_1 which you found in Problem 2.1.6.

Definition 2.1.10. The mgf of $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$M_{\mathbf{X}}(\boldsymbol{t}) = E(e^{\boldsymbol{t}\cdot\boldsymbol{x}}) = E(e^{t_1x_1+t_2x_2+\cdots+t_nx_n})$$

Clearly $M_{\mathbf{X}}(t, 0) = M_{X_1}(t)$.

Problems from the Text

Section 2.1: 1,2,3,6,7,9,12

2.2 Transformations of Bivariate RV

Assume that $\mathbf{X}$ is a random vector and Y is some function $Y = g(\mathbf{X})$ of $\mathbf{X}$. Given the joint pdf of $\mathbf{X}$ we can find the pdf of Y by going through the cdf: finding

$$F_Y(y) = \oiint_{S=\{\boldsymbol{x}|g(\boldsymbol{x})\leq y\}} f_{x_1,x_2}(x_1,x_2)\, d\boldsymbol{x}$$

and then differentiatin to get $f_{\mathbf{Y}}(y)$.

In this section we look at the method of transformations for doing the same thing.

2.2.1 Discrete Case

In the one variable case when we had $Y = g(X)$ for one-to-one (and increasing) g we observed that clearly $p_y(y) = p_x(g^{-1}(y))$. Now, when $Y = g(X_1, X_2)$ we are not one-to-one. We overcome this by extending g to a transformation of random vectors.

Example 2.2.1. Let $\mathbf{X} = (X_1, X_2)$ have pmf

$$p_{\boldsymbol{x}}(x_1, x_2) = \frac{\mu_1^{x_1}\mu_2^{x_2}e^{-\mu_1}e^{-\mu_2}}{x_1!x_2!} x_i \in \mathbb{N},$$

and $Y_1 = X_1 + X_2$. This is not one-to-one, but we can make it so by introduction a dummy variable Y_2 and extending it to a one-to-one transformation $\boldsymbol{u}$ of vectors:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} u_1(X_1, X_2) \\ u_2(X_1, X_2) \end{bmatrix} = \begin{bmatrix} X_1 + X_2 \\ X_1 \end{bmatrix}.$$

This is indeed one-to-one as it has the inverse transformation

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} w_1(Y_1, Y_2) \\ w_2(Y_1, Y_2) \end{bmatrix} = \begin{bmatrix} Y_2 \\ Y_1 - Y_2 \end{bmatrix}.$$

So as in the one variable case, we clearly have

$$p_{\boldsymbol{y}}(y_1, y_2) = p_{\boldsymbol{x}}(w_1(y_1, y_2), w_2(y_1, y_2)) = \frac{\mu_1^{y_2} \mu_2^{y_1 - y_2} e^{-\mu_1} e^{-\mu_2}}{y_2!(y_1 - y_2)!}.$$

Now the sample space of $\mathbf{X}$ is $\mathbb{N}^2$, and $\boldsymbol{u}$ maps this to the space

$$\{(Y_1, Y_2) \mid Y_2 \in \mathbb{N}, Y_1 - Y_2 \in \mathbb{N}\}$$

which means $Y_2 \in \mathbb{N}$, $Y_1 \in \mathbb{N}$ and $Y_2 \leq Y_1$.

Then $p_{y_1}(y_1)$ is the marginal pmf

$$\begin{aligned}
p_{y_1}(y_1) &= \sum_{y_2=0}^{y_1} \frac{\mu_1^{y_2} \mu_2^{y_1 - y_2} e^{-\mu_1} e^{-\mu_2}}{y_2!(y_1 - y_2)!} \\
&= \frac{e^{-(\mu_1 + \mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \frac{y_1!}{y_2!(y_1 - y_2)!} \mu_1^{y_1 - y_2} \mu_2^{y_2} \\
&= \frac{e^{-(\mu_1 + \mu_2)}}{y_1!} \sum_{y_2=0}^{y_1} \binom{y_1}{y_2} \mu_1^{y_1 - y_2} \mu_2^{y_2} \\
&= \frac{(\mu_1 + \mu_2)^{y_1} e^{-(\mu_1 + \mu_2)}}{y_1!}
\end{aligned}$$

where the last line uses the binomial expansion of $(\mu_1 + \mu_2)^{y_1}$.

Problem 2.2.2. We might write the transformation $\boldsymbol{u}$ in the above example as

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

calling the square matrix in the middle U. Write the inverse transformation as

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = W \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

for some square matrix W. What do you notice about U and W?

2.2.2 A Continuous Example

Let $\mathbf{X} = (X_1, X_2)$ have the uniform distribution on the unit square $D = \{(x_1, x_2) \mid 0 \leq x_i \leq 1\}$; so the pdf $f_{\mathbf{X}}$ is 1 on D and 0 elsewhere.

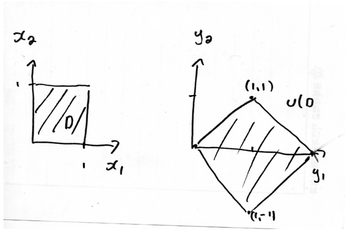
To find the pdf of $Y_1 = X_1 + X_2$, we add the dummy variable Y_2 and extend Y_1 to the transformation

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} u_1(X_1, X_2) \\ u_2(X_1, X_2) \end{bmatrix} = \begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix},$$

with inverse transformation

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} w_1(Y_1, Y_2) \\ w_2(Y_1, Y_2) \end{bmatrix} = \begin{bmatrix} \frac{Y_1+Y_2}{2} \\ \frac{Y_1-Y_2}{2} \end{bmatrix}.$$

The transformation $\boldsymbol{u}$ takes the sample space D of $\mathbf{X}$ to the sample space $\boldsymbol{u}(D)$ shown on the right.



Problem 2.2.3. Describe $\boldsymbol{u}(D)$ mathematically and show that it is as drawn.

Now to get the pmf of $\mathbf{Y}$ we integrate $f_{\mathbf{X}}$ over D and then differentiate with respect to $\mathbf{Y}$. In short, want the function $f_{\mathbf{Y}}$ such that

$$\oiint_{\boldsymbol{u}(D)} f_{\mathbf{Y}}(y_1, y_2)\, dy_1\, dy_2 = \oiint_D f_{\mathbf{X}}(x_1, x_2)\, \mathrm{dx}_1\, \mathrm{dx}_2.$$

Recall from calculus that where J is the Jacobian

$$J = \frac{\partial w_1}{\partial y_1}\frac{\partial w_2}{\partial y_2} - \frac{\partial w_1}{\partial y_2}\frac{\partial w_2}{\partial y_1} = \left(\frac{1}{2}\right)\left(-\frac{1}{2}\right) - \frac{1}{2}\frac{1}{2} = -\frac{1}{2},$$

we have

$$f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{X}}(w(\boldsymbol{y})) \cdot |J| = 1 \cdot \left|-\frac{1}{2}\right| = \frac{1}{2}$$

on $\boldsymbol{u}(D)$ and 0 elsewhere.

To get the marginal pdf f_{Y_1} we then integrate with respect to Y_2. When $y_1 \in [0, 1]$, $f_{\mathbf{Y}}(y_1, y_2)$ is 1 for $y_2 \in [-y_1, y_1]$, so

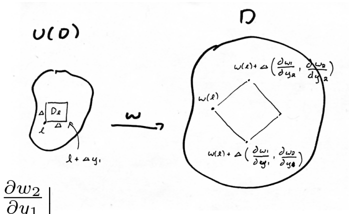
$$f_{Y_1}(y_1) = \int_{-y_1}^{y_1} \frac{1}{2}\, dy_2 = y_1$$

and (as $\boldsymbol{u}(D)$ is symmetric about $y_1 = 1$,) when $y_1 \in [1, 2]$, $f_{\mathbf{Y}}(y_1, y_2) = f_{\mathbf{Y}}(2 - y_1, y_2)$ so $f_{Y_1}(y_1) = 2 - y_1$.

2.2.3 Recalling Jacobian from Calculus

The formula $f_{\mathbf{Y}}(\boldsymbol{y}) = f_{\mathbf{X}}(w(\boldsymbol{y}))\cdot|J|$ can be proved by considering approximation of the integral $\oiint_{\boldsymbol{u}(D)} f_{\mathbf{Y}}(\boldsymbol{y})\, d\boldsymbol{y}$. We would use $\sum_L f_{\mathbf{Y}}(\ell)\Delta^2$: the sum of $f_{\mathbf{Y}}$ at points ℓ of a Δ lattice on $\boldsymbol{u}(D)$, each multiplied by the area of the Δ square D_ℓ in the positive directions from ℓ.

So that $\oiint_{\boldsymbol{u}(D)} f_{\mathbf{Y}}(\boldsymbol{y})\, d\boldsymbol{y} = \oiint_D f_{\mathbf{X}}(\boldsymbol{x})\, d\boldsymbol{x}$, $f_{\mathbf{Y}}(\ell)\Delta^2$ should then be equal to $f_{\mathbf{X}}(\boldsymbol{w}(\ell))A$ where A is the area of the $\boldsymbol{w}(D_\ell)$.


U(D)
D
D_ℓ
Δ
Δ
ℓ
$\ell + \Delta y_1$
w
$w(\ell)$
$w(\ell) + \Delta(\frac{\partial w_1}{\partial y_2}, \frac{\partial w_2}{\partial y_2})$
$w(\ell) + \Delta(\frac{\partial w_1}{\partial y_1}, \frac{\partial w_2}{\partial y_1})$

But $\boldsymbol{w}(D_\ell)$ is approximatley a parallelogram between the vectors $\Delta(\frac{\partial w_1}{\partial y_1}, \frac{\partial w_2}{\partial y_1})$ and $\Delta(\frac{\partial w_1}{\partial y_2}, \frac{\partial w_2}{\partial y_2})$, so has area approximately

$$A = \Delta^2 \begin{vmatrix} \frac{\partial w_1}{\partial y_1} & \frac{\partial w_2}{\partial y_1} \\ \frac{\partial w_1}{\partial y_2} & \frac{\partial w_2}{\partial y_2} \end{vmatrix}.$$

Taking limits in this argument gives the Jacobian formula.

Problems from the Text

Section 2.2: 1,3,5,6

2.3 Conditional Distributions

Definition 2.3.1. Let (X, Y) be a random vector. The *conditional pmf (pdf)* of X, conditioned on Y, is

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

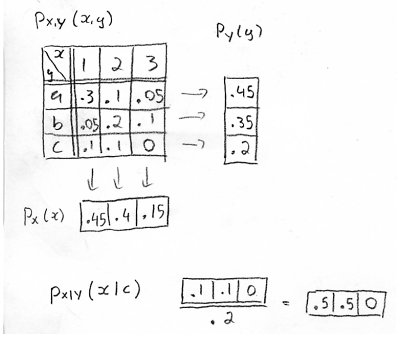
(or the same with f in place of p).

In the discrete case we have that $p_{X|Y}(x|y)$ is the probability that $X = x$ given that $Y = y$. The continuous case is the density analogue.

Note

In the case of the random vector (X_1, X_2) we may use shortcut notation such as $p_{1|2}$ for $p_{X_1|X_2}$.

Example 2.3.2. Let (X, Y) have the joint distribution $p_{X,Y}$ shown.



$p_{x,y}(x,y)$ $p_y(y)$

$y \backslash x$	1	2	3		
a	.3	.1	.05	→	.45
b	.05	.2	.1	→	.35
c	.1	.1	0	→	.2
	↓	↓	↓		
$p_x(x)$	.45	.4	.15		

$$p_{x|y}(x|c) \quad \frac{\boxed{.1 \,|\, .1 \,|\, 0}}{.2} = \boxed{.5 \,|\, .5 \,|\, 0}.$$

Then the marginal pmf p_X takes x to the sum of the x^{th} column, and and the marginal pmf p_Y takes y to the sum of the y^{th} row. So $p_x(3) = .15$ and $p_Y(b) = .35$. The conditional pmf $p_{X|Y}(x|y)$ restricts to the y^{th} row, scales it by $1/p_Y(y)$ (so that it sums to 1) and then returns the x value.

Observe that for fixed y the function

$$p_{X|y} : x \mapsto p_{X|Y}(x|y)$$

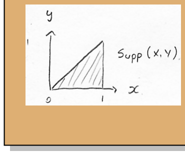
is itself the pmf of a random variable, the *conditional random variable* (or *conditional distribution*) which we denote $X|y$. Being an RV, we can compute its mean and variance.

Note

The notation $p_{X|Y}$ vs $p_{X|y}$ can get confusing. The argument of $p_{X|Y}$ is $x|y$, with two values, the argument of $p_{X|y}$ is simply x. Use this heuristic aid: if the letter in the index is uppercase, the argument has a corresponding lower case value.

Example 2.3.3. Let (X, Y) have the pdf $f_{X,Y}(x, y) = 6y$ on its support $0 \leq y \leq x \leq 1$. We find the mean μ of $X|y$.

Note



First, by definition $\mu = E(X|y) = \int_y^1 x \cdot f_{X|y}(x)\,\mathrm{d}x$, so we need to find $f_{X|y}(x) = f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$. Now the marginal pmf in y is

$$f_Y(y) = \int_y^1 f_{X,Y}(x, y)\,\mathrm{d}x = \int_y^1 6y\,\mathrm{d}x = 6y(1 - y),$$

so

$$f_{X|y}(x) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{6y}{6y - 6y^2} = \frac{1}{1 - y},$$

and so

$$E(X|y) = \frac{1}{1 - y}\int_y^1 x\,\mathrm{d}x = \frac{1}{1 - y}\frac{1}{2}(1 - y^2) = \frac{1 + y}{2}.$$

Problem 2.3.4. Show that the variance of $X|y$ above is $\sigma^2 = \frac{y^2-2y+1}{12}$.

We showed for any y in $[0, 1]$ that $E(X|y) = \frac{1+y}{2}$. This is a transformation $g(Y)$ of Y, where $g(y) = \frac{1+y}{2}$, so itself gives us random variable. We denote this random variable by $E(X|Y)$. Its sample space is $[1/2, 1]$.

Example 2.3.5. By the method of transformations we have that the pdf of $E(X|Y)$ is

$$f_{E(X|Y)}(z) = f_Y(2z - 1)\left|\frac{\mathrm{d}(2\mathrm{z} - 1)}{\mathrm{dz}}\right| = \frac{1 + (2z - 1)}{2} \cdot 2 = 2z.$$

We used that the inverse transformation is $y = g^{-1}(z) = 2z - 1$.

The following simple observation will be a useful tool in finding 'minimum variance estimators.'

Theorem 2.3.6. *For a random vector (X, Y),*

i) $E(E(X|Y)) = E(X)$.

ii) $\mathrm{Var}\,(E(X|Y)) < \mathrm{Var}(X)$.

Proof. For the first statement we have:

$$\begin{aligned} E(E(X|Y)) &= \int_{-\infty}^{\infty} E(X|y) f_Y(y)\,\mathrm{dy} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|y}(x|y) f_y(y)\,\mathrm{dx}\,\mathrm{dy} \\ &= \int\int x \cdot f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} = \mathrm{E(X)}. \end{aligned}$$

For the second, we show

$$E\left(E(X|Y)^2\right) \le E(X^2);$$

taking μ^2 from both sides of this equation then gives the result we are looking for.

$$\begin{aligned} E(E(X|Y)^2) &= \int_{\mathbb{R}} E(X|y)^2 \cdot f_Y(y)\,\mathrm{dy} \\ &\le \int_{\mathbb{R}} E(X^2|y) \cdot f_Y(y)\,\mathrm{dy} \qquad \left(\text{applying } E(X)^2 \le E(X^2) \text{ to the RV } X|y\right) \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} x^2 f_{X|Y}(x|y) \cdot f_Y(y)\,\mathrm{dx}\,\mathrm{dy} \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} x^2 f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} \\ &= E(X^2) \end{aligned}$$

□

Problems from the Text

Section 2.3: 1,2,3,5,7

2.5 Independent Random Variables

Definition 2.5.1. Random variables X and Y, with joint pdf f_{XY} and marginal pdfs f_X and f_Y, are *independent* if

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$$

(holds with probability 1). They are *dependent* otherwise.

There are several equivalent definitions of independence.

Theorem 2.5.2. *The following are equivalent for RVs X and Y.*

i) X and Y are independent

ii) $f_{XY}(x,y) = g(x)h(y)$ (almost everywhere) for some non-negative functions g and h

iii) The cdfs satisty $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all x and y.

iv) For all intervals S_X and $S_Y \subset \mathbb{R}$,

$$P(X \in S_X, Y \in S_Y) = P(X \in S_X)P(Y \in S_Y).$$

Proof. That i) impies ii) is immediate from the definition. We first show ii) implies i). Assuming ii), the marginal pdfs are

$$f_X(x) = \int_{\mathbb{R}} g(x)h(y)\,\mathrm{d}y = g(x)\int_{\mathbb{R}} h(y)\,\mathrm{d}y = c_1 g(x)$$

for some constant c_1 and $f_Y(y) = c_2 h(y)$ for some constant c_2. As

$$\begin{aligned} 1 &= \int_{\mathbb{R}}\int_{\mathbb{R}} f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{\mathbb{R}}\int_{\mathbb{R}} g(x)h(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{\mathbb{R}} g(x)\,\mathrm{d}x \int_{\mathbb{R}} h(y)\,\mathrm{d}y = c_1 c_2 \end{aligned}$$

we get that $c_1c_2 = 1$. So

$$f_{XY}(x,y) = g(x)h(y) = \frac{f_X(x)f_Y(y)}{c_1c_2} = f_X(x)f_Y(y).$$

For i) implies iii):

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(s,t)\,\mathrm{d}t\,\mathrm{d}s = \int\int f_X(s)f_Y(t)\,\mathrm{d}t\,\mathrm{d}s \\ &= \int f_X(s)\,\mathrm{d}s \int f_Y(t)\,\mathrm{d}t = F_X(x)F_Y(y) \end{aligned}$$

For iii) implies i):

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

The proof of the equivalence of iii) and iv) is just as straight forward, so we skip it. □

Theorem 2.5.3. *If X and Y are independent, then $E(XY) = E(X)E(Y)$.*

Proof.

$$\begin{aligned} E(XY) &= \int\int xy f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int xy f_X(x) f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int x f_X(x)\,\mathrm{d}x \int y f_Y(y)\,\mathrm{d}y = E(X)E(Y) \end{aligned}$$

□

Problem 2.5.4. Show that if X and Y are independent, then $E(X|y) = E(X)$.

Problems from the Text

Section 2.5: 1,3,4,5,8,9,12

2.4 The Correlation Coefficients

The following is done for continuous variables, which are assumed to be nice, (ie., all necessary expectations are assumed to exist). It holds though for discrete variables as well.

For random variables X and Y with means μ_X and μ_Y respectively, the *covariance* is $\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$.

Problem 2.4.1. Show that $\mathrm{Cov}(X,Y) = E(XY) - \mu_X\mu_Y$.

Observe that if $X = Y$ then $\mathrm{Cov}(X,Y) = E(X^2) - E(X)^2 = \mathrm{Var}(X)$; so the covariance can be seen as a generalisation of the variance.

Problem 2.4.2. Show that if X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$. Show if $E(X|y)$ is an increasing (decreasing) function of y then $\mathrm{Cov}(X,Y) > 0$ ($\mathrm{Cov}(X,Y) < 0$).

The magnitude of $\mathrm{Cov}(X,Y)$ is hard to interpret, but the normalised version, the *correlation coefficient*

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

has the property that $-1 \leq \rho \leq 1$. If $X = Y$ then $\rho = 1$ and if $X = -Y$ then $\rho = -1$. So $|\rho|$ is a measure of how closely X and Y are related.

Problems from the Text

Section 2.4: 1,3,4,10

2.6 Extension to more Random Variables

Let $\mathbf{X} = (X_1, \ldots X_n)$ be an n dimensional random vector. Its joint cdf is

$$F_{\mathbf{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots X_n \leq x_n)$$

and its joint pdf is a function $f_{\mathbf{X}}$ such that

$$F_{\mathbf{X}}(\boldsymbol{y}) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f_{\mathbf{X}}(\boldsymbol{x})\,\mathrm{d}x_n \ldots \mathrm{d}x_1.$$

The conditional pdfs are

$$f_{\mathbf{X}|X_i}(\boldsymbol{x}|x_i) = \frac{f_{\mathbf{X}}(\boldsymbol{x})}{f_{X_i}(x_i)}.$$

The variables $X_1, \ldots X_n$ are *mutully independent* if

$$f_{\mathbf{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

(with probability 1.)

In this case

$$E(\prod u_i(X_i)) = \prod E(u_i(X_i))$$

for any transformations u_i of X_i. In particular:

Theorem 2.6.1. *Let $T = \sum_{i=1}^{n} k_i X_i$ where $X_1, \ldots, X_n$ are mutually independent RVs having respective mgfs $M_1(t), \ldots, M_n(t)$. The RV T has mgf*

$$M_T(t) = \prod M_i(k_i t).$$

Proof.

$$\begin{aligned} M_T(t) &= E(e^{tT}) = E(e^{t\sum k_i X_i}) = E(\prod e^{tk_i X_i}) \\ &= \prod E(e^{tk_i X_i}) = \prod M_i(k_i t) \end{aligned}$$

□

A vector of RVs is *independent identically distributed* or *iid* if the components are mutually independent and all have the same pdfs. An n-dimensional iid random vector of variables all having the same pdf as a RV X is an *random sample of distribution* X ; it has n *tests*, or n *samples*, or simply has *size* n. Often we implicitly assume n is the size of the sample.

Corollary 2.6.2. *If* $\mathbf{X}$ *is a random sample of distribution* X *then* $M_{\sum X_i}(t) = (M_{X_1}(t))^n$.

Problems from the Text

Section 2.6: 1,2

2.7 Transformations for more Variables

This is mostly the same as Section 2.2 so we skip it, except for noting what the jacobian looks like for more variables.

For n-dimensional random vectors $\mathbf{X}$ and $\mathbf{Y}$ a transformation $\boldsymbol{u} : \mathbf{X} \to \mathbf{Y}$ is described by $Y_i = u_i(X_1, \ldots X_n)$ for $i = 1, \ldots, n$ and its inverse is described by $X_i = w_i(Y_1, \ldots, Y_n)$.

The jacobian is the determinant

$$\begin{vmatrix} \frac{\partial w_1}{\partial y_1} & \cdots & \frac{\partial w_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial w_n}{\partial y_1} & \cdots & \frac{\partial w_n}{\partial y_n} \end{vmatrix}$$

2.8 Linear Combinations of Random Variables

Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors. By the linearity of expectation

$$E(\sum a_i X_i) = \sum a_i E(X_i).$$

Further

$$\begin{aligned} \operatorname{Cov}(\sum a_i X_i, \sum b_i Y_i) &= E\left((\sum a_i X_i - \sum a_i E(X_i))(\sum b_i Y_i - \sum b_i E(Y_i))\right) \\ &= E(\sum\sum a_i b_j X_i Y_j - \sum\sum a_i b_j E(X_i) Y_j + \ldots \\ &= \sum\sum a_i b_j E[X_i Y_j - X_i E(Y_j) - E(X_i) Y_j + E(X_i) E(Y_j)] \\ &= \sum\sum a_i b_j E[(X_i - E(X_i))(Y_j - E(Y_j))] \\ &= \sum\sum a_i b_j \operatorname{Cov}(X_i, Y_j) \end{aligned}$$

In the case that $\mathbf{X} = \mathbf{Y}$ this gives that

$$\operatorname{Var}(\sum a_i X_i) = \sum\sum a_i a_j \operatorname{Cov}(X_i, X_j) = \sum a_i^2 \operatorname{Var}(X_i)$$

where the last inequality uses that the non-diagonal terms are 0 by the independence of the variables.

If $\mathbf{X}$ is a random sample of a distribution X having mean μ and variance σ^2, then the *sample mean* is

$$\overline{X} = \frac{\sum X_i}{n}.$$

It has expected value

$$E(\overline{X}) = E\left(\frac{1}{n}\sum X_i\right) = \frac{nE(X)}{n} = E(X)$$

and variance

$$\operatorname{Var}(\overline{X}) = \frac{1}{n^2}\sum \operatorname{Var}(X_i) = \frac{\sigma^2}{n}.$$

The *sample variance* is

$$S^2 = \frac{\sum(X_i - \overline{X})^2}{n-1} = \frac{\sum X_i^2 - n\overline{X}^2}{n-1}$$

We get the second equality above as follows.

$$\begin{aligned}
\sum(X_i - \overline{X})^2 &= \sum(X_i^2 - 2X_i\overline{X} + \overline{X}^2) \\
&= \sum X_i^2 - 2\overline{X}\sum X_i + n\overline{X}^2 \\
&= \sum X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum X_i^2 - n\overline{X}^2
\end{aligned}$$

The sample variance is a random variable. It has expected value

$$\begin{aligned}
E(S^2) &= \frac{1}{n-1}\left(\sum E(X_i^2) - nE(\overline{X}^2)\right) \\
&= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n(\mu^2 + \frac{\sigma^2}{n})\right) = \sigma^2
\end{aligned}$$

Problems from the Text

Section 2.8: 2,3,10

B Cycles in $G_{n,p}$

We give an analysis of the number of cycles in $G_{n,p}$, exhibiting how we use the concepts (to be) developed in Chapters 1, 2 and 3.

B.1 The probability that $G_{n,p}$ is a cycle

In Chapter 1, we viewed $G_{n,p}$ as a sample space containing every possible graph on n vertices. Where $N = \binom{n}{2}$, any graph H with m edges occured with probability $p^m q^{N-m}$.

To calculate the probability that $G_{n,p}$ is isomorphic to some graph H, we must count the number of isomorphic copies of H in $G_{n,p}$, and then use the additivity of probability on these (disjoint) basic event.

Example B.1.1. The event C_n that $G = G_{n,p}$ is an n-cycle contains $n!/2n$ different outcomes, as this is the number of different n-cycles on n labelled vertices, and each occurs with probability $p^n(1-p)^{\binom{n}{2}-n}$. So the probability that G is an n-cycle is

$$P(C_n) = \frac{n!}{2n} p^n (1-p)^{\binom{n}{2}-n} = \frac{n!}{2n}(p^n - p^{\binom{n}{2}}).$$

Problem B.1.2. What is the probability that $G = G_{n,p}$ is a k-cycle? That it is a cycle of any girth?

B.2 Expected number of cycles in $G_{n,p}$

Finding the probability that $G_{n,p}$ is a cycle is easy, it is a much harder task to find the probability that $G_{n,p}$ **contains** a cycle.

For a cycle c on the vertices $[n]$, let Y_c be the event that $G_{n,p}$ contains c. The following is not so hard.

Problem B.2.1. Find $P(Y_c)$.

However, for two different n-cycles c and c', Y_c and $Y_{c'}$ are not independent as they were when the event were that $G_{n,p}$ is c, so we cannot get a precise value for the probability that $G_{N,p}$ contains any n-cycle by using the subadditivity of probability.

But we can get a bound using the expected number of cycles. If the expected number of cycles is much less than 1, then the probability of a cycle is low. If the expected number of cycles is much more than 1, then the probability of a cycle is high.

We observed in Example 2.1.3 that $G_{n,p}$ can be viewed as a random vector of $N = \binom{n}{2}$ independent RVs X_i. To count the expected number of edges we find the expected value of the RV $M = \sum_{i=1}^{N} X_i$ which essentially counts edges of $G_{n,p}$.

We use the same setup to count the number of copies of any substructure: define an indicator variable for each possible occurence, and a 'counting' variable as the sum of the indicators. Then the expected value of this counting variable is the expected number of copies of the substructre in $G_{n,p}$.

Example B.2.2. Every permutation of the set $[n]$ determines an n-cycle on $[n]$, but each cycle is determined by $2n$ such permutations, so there are $n!/2n$ different n-cycles in $G_{n,p}$. Let $Y_c = \prod_{e \in c} X_e$ indicate the event that the cycle c is in G. Then $E(Y_c) = E(X_e)^n = p^n$. Let $C_n = \sum Y_c$ count the number of n-cycles in G. Then by the additivity of expectation we have that the expected number of n-cycles in G is

$$E(C_n) = \sum_c E(Y_c) = n!/2n \cdot p^n \approx (np)^n.$$

Problem B.2.3. Show that the expected number of k-cycles is

$$E(C_k) = \frac{n!}{(n-k)!2k} \cdot p^k < (np)^k/k$$

and that the expected number of cycles of any girth is

$$E(C) = \sum_{k=3}^{n} (np)^k/k.$$

Now when $p = 1/n$ we get that

$$E(C) = \sum_{k=3}^{n} (np)^k/k < \sum_{k=3}^{n} 1/k < \int_3^{n-1} 1/x \, dx < \ln(n/2),$$

which is a pretty small number; bigger than, but close to 1. By taking p a tiny bit bigger, we will have a very good probabilty that $G_{n,p}$ contains a cycle. On the other hand, by taking p a bit smaller, we will see that $G_{n,p}$ almost never contains a cycle.

We have to introduce some language for these ideas.

B.3 Asymptotics and thresholds

Definition B.3.1. Recall that for functions $f, g : \mathbb{N} \to \mathbb{R}$ we write $f = o(g)$ if

$$\lim_{n \to \infty} \frac{f(n)}{g(n)} = 0.$$

Problem B.3.2. Show that

i) $f = o(1)$ if and only if $\lim_{n\to\infty} f(n) = 0$.

ii) $p = o(1/n)$ if and only if $\lim_{n\to\infty} pn = 0$.

Example B.3.3. We show that if $p = o(1/n)$ then $E(C) = o(1)$, where C is the random variable counting the number of cycles in $G_{n,p}$.

Indeed, if $p = o(1/n)$ then for every $\varepsilon > 0$ there is some N_ε such that $n > N_\varepsilon$ implies that $np < \min(\varepsilon, 1/2)$. So

$$\begin{aligned} E(C) = \sum_{k=3}^{n} (np)^k/k \quad &< \quad \varepsilon \sum_{k=3}^{n} 1/(2^{k-1}k) \\ &< \quad \varepsilon \sum_{k=2}^{n} 1/2^k < \varepsilon \end{aligned}$$

This gives us that $E(C) \to 0$, as needed.

Definition B.3.4. Any event $C \subset \mathscr{C}$ in the sample space of the random graph $G_{n,p} \sim (X_1, \ldots, X_m)$ is a *property.* The *probability that $G_{n,p}$ has a property C* is $P(C)$. If $P(C) \to 1$ as $n \to \infty$ then C occurs *asymptotically almost surely* (aas) or *with high probability* (whp) aas.

Problem B.3.5. Using Markov's inequality, show that if $p = o(1/n)$, then where C is again the random variable counting the cycles in $G_{n,p}$, show that the property $C = 0$ occurs aas.

It might be suprising therefore that if $p = (1+\varepsilon)/n$ then ass $C \geq 1$. This is called a threshold.

Definition B.3.6. For a property C of the random graph $G_{n,p}$, a *threshold for C* is a probability p_0 such that if $p/p_0 = o(1)$ then aas C doesn't occur, and if $p_0/p = o(1)$ then aas C does occur.

Theorem B.3.7. *Containing a cycle is a property of $G_{n,p}$. The value $p_0 = 1/n$ is a threshhold for this property.*

Proof. We have shown that if $p = o(1/n)$, so if $p/p_0 = o(1)$, then ass $G_{n,p}$ contains no cycles. We have to show that if $p_0/p = o(1)$, so if $p/n \to \infty$ then $G_{n,p}$ almost surely has a cycle. This uses the simple observation that if $G_{n,p}$ has at least n edges, then it must contain a cycle. So we show that if $p/n \to \infty$, then ass $G_{n,p}$ has at least n edges.

As before, let $M = \sum_{i=1}^{N} X_i$ be the random variable counting the edges of $G_{n,p}$. We saw that $\mu = E(M) = pN = \frac{p \cdot n \cdot n-1}{2}$. Lets compute the variance σ^2

of M. As

$$\begin{aligned} E(M^2) &= E(M \cdot \sum X_i) = \sum E(M \cdot X_i) = NE(M \cdot X_i) \\ &= NE(\sum X_j \cdot X_i) = N(N-1)E(X_i \cdot X_j) + NE(X_i^2) \\ &= N(N-1)p^2 + Np = Np(p(N-1)+1), \end{aligned}$$

we get that

$$\sigma^2 = E(M^2) - \mu^2 = Np(p(N-1) + 1 - Np) = Np(1-p) < \mu.$$

Taking $p \geq 5/n$ we get that $\mu > 2n$ so by Chebyshev's inequality we there for get that

$$P(M > n) > P(|M - 2n| < n) = P(|M - \mu| < n) > 1 - \frac{\sigma^2}{n^2} > 1 - \frac{2}{n}.$$

As $p/n \to \infty$, we have for large enough n that $p > 5/n$, and so $P(M > n) > (1 - \frac{2}{n}) \to 1$. □

What we have done in this proof is shown that the distribution M is 'concentrated' around its mean μ, and that this concentration increases as n does. An important fact that we will soon expand on.

3 Some Special Distributions

3.1 Binomial and Related Distributions

We have given a name to only one distribution so far: the uniform distribution whose pdf or pmf is constant on its support. There are several other distributions that occur repeatedly in mathematics and statistics. One of the most basic is the Binomial Distribution, which we build from the following distribution, which we will recognise as the distribution of outcomes when tossing a p-coin.

3.1.1 The Bernoulli Distribution

Bernoulli $X \sim b(1, p)$

$p_X(x)$	$\begin{cases} p & \text{if } x = 1 \\ 1-p & \text{otherwise.} \end{cases}$
μ	p
σ^2	pq
$M_X(t)$	$pe^t + q$

Definition 3.1.1. An RV X has a *Bernoulli distribution*, or is a *Bernoulli RV*, if its support is $\{0, 1\}$. Its pmf is

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{otherwise,} \end{cases}$$

for some probability $p \in [0, 1]$.

If X is a Bernoulli distribution with probability p, then its mean is $\mu = p$ and its variance is $\sigma^2 = p(1-p) = pq$.

Indeed, $\mu = p \cdot 1 + (p-1) \cdot 0 = p$, and $E(X^2) = p \cdot 1^2 + (p-1) \cdot 0^2 = p$, so $\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1-p)$.

The Bernoulli distribution of probabilitly p is thus more often called the *Bernoulli distribution of mean p*.

3.1.2 The Binomial Distribution

Binomial $X \sim b(n, p)$

$p_X(x)$	$\binom{n}{x}p^x(1-p)^{n-x}$
μ	np
σ^2	npq
$M_X(t)$	$(pe^t + q)^n$

Often a probabilitly space consists of n independent Bernoulli spaces. The random graph $G_{n,p}$ is an example of this, the experiment of tossing 100 p-coins is a simpler example. When tossing 100 p-coins, the only random variable of any interest is that which counts the number of times we the outcome is 'heads'. This is the Binomial distribution.

Definition 3.1.2. A random variable Y has the *Binomial distribution* $b(n,p)$ if

$$Y = \sum_{i=1}^{n} X_i$$

for a family $\{X_i\}_{i\in[n]}$ of iid Bernoulli RVs with mean p.

Clearly the pmf of $Y \sim b(n,p)$ is

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

on its support $y = 0, 1, \ldots, n$. By the linearity of expectation, its mean is

$$\mu = E(Y) = \sum_{i=1}^{n} E(X_i) = \sum p = np.$$

The random variable M counting the edges of $G_{n,p}$ is the binomial distribution $b(n,p)$. We did the following in the proof of Theorem B.3.7.

Problem 3.1.3. Show that the variance of $Y \sim b(n,p)$ is $\sigma^2 = np(1-p)$.

The 'concentration' part of the proof of Theorem B.3.7 can be viewed in the following way.

Example 3.1.4. If $Y \sim b(n,p)$, then Y/n can be viewed as the 'rate of success' of the trials $X_1, \ldots, X_n$ making up Y. Clearly $E(Y/n) = E(Y)/n = np/n = p$, and one can show that $\mathrm{Var}(Y/n) = \frac{p}{1-p}n$. So by Chebyshev,

$$P(|Y/n - p| \geq \varepsilon) \leq \frac{\mathrm{Var}(Y/n)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \to 0.$$

This means that the rate of success of the trials X_i is more and more concentrated around p as n gets bigger. This is an example of the Law of Large Numbers.

The mgf of $X \sim b(n,p)$ is

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} e^{xt} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x} \\ &= (pe^t + q)^n. \end{aligned}$$

Problem 3.1.5. Use the mgf to find μ and σ^2 of $X \sim b(n,p)$.

Problem 3.1.6. Show that if $X_i \sim b(n_i, p)$ for $i = 1, \ldots, d$ then $Y = \sum_{i=1}^{d} X_i$ has distribution $b(\sum_{i=1}^{d} n_i, p)$.

We do not do much with the rest of the distributions in this section. We simply define them so that we have seen them.

3.1.3 The geometric and negative binomial distributions

For the binomial distribution $b(n,p)$ we conducted n independent Bernoulli trials with mean p and counted the number of successes. For the Geometic distribution Y, we conduct Bernoulli trials with mean p until there is a success. We let Y count the number of failures.

Formally, the *geometric RV with parameter* p is the RV with pmf

$$p(y) = (1-p)^y \cdot p.$$

More generally the *negative binomial RV with parameters* p *and* r is the RV that counts the number of failures that occur, when conducting Bernoulli trials $b(1,p)$, until the r^{th} success. It has pmf

$$p(y) = \binom{y+r-1}{r-1} p^r (1-p)^y.$$

3.1.4 The Hypergeometric Distribution

Hypergeometric

$p_X(x)$	$\frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}$
μ	$n\frac{D}{N}$
σ^2	$n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1}$

In a lot of N items, D are defective. We choose n items. The RV X that counts the number of chosen items that are defective is a *hypergeometric* distribution. Its pmf is

$$p(x) = \frac{\binom{N-D}{n-x}\binom{D}{x}}{\binom{N}{n}}.$$

The expected value of X is

$$E(X) = \sum_{x=0}^{n} xp(x) = \sum \binom{N-D}{n-x}\binom{D}{x}\binom{N}{n}^{-1}.$$

Using that $b\binom{a}{b} = a\binom{a-1}{b-1}$ and so

$$\binom{a}{b} = \frac{a}{b}\binom{a-1}{b-1}$$