INTRODUCTION TO PROBABILITY LECTURE NOTES

JOHN PIKE

These lecture notes were written for MATH 4710 at Cornell University in the Fall semester of 2014. They are intended for personal educational use only. Almost all of the material and structure (as well as some of the language) comes directly from the course text, *A First Course in Probability* by Sheldon Ross. There are likely typos and mistakes in these notes. All such errors are the author's fault and corrections are greatly appreciated.

Day 1: Section 1
Day 2: Section 2
Day 3: Up to Example 3.6
Day 4: Through set algebra
Day 5: Up to inclusion-exclusion
(Explained Vitali sets and it seemed to go over really well!)
Day 6: Up to poker hands
Day 7: Finished Section 5
Day 8: Up to Borel-Cantelli I
Day 9: Finished Section 7
Day 10: Finished Section 8
Day 11: Up to Problem of Points
Day 12: Through Example 9.6
Day 13: Finished Section 9
Day 14: Finished Section 10
Day 15: Discussed cardinality, Through Theorem 11.1
Day 16: Prelim 1
Day 17: Up to Example 11.8
Day 18: Finished Section 11
Day 19: Up to negative binomial
Day 20: Up to Poisson
Day 21: Finished Section 12
Day 22: Through Example 13.3
Day 23: Up to Example 13.6
Day 24: Through "normal integrates to 1"
Day 25: Up to memorylessness of exponential
Day 26: Up to joint c.d.f.
Day 27: Finished Section 15
Day 28: Through Example 16.1
Day 29: Up to Example 17.1
Day 30: Up to multivariate change of variables
Day 31: Through Example 18.3
Day 32: Finished Section 19
Day 33: Prelim 2
Day 34: Through Example 20.3
Day 35: Finished Section 21
Day 36: Finished Section 22
Day 37: Up to $\to_{a.s.}$ implies $\to_p$
Day 38: Up to Example 24.1
Day 39: Finished Section 24

1. Advertisement

Example 1.1. (Monty Hall)

You are a contestant on a game show and asked to choose one of three doors. A prize is behind a random door, and the other two are empty. The host knows where the prize is. You pick a door. At least one of the doors you did not pick does not have a prize and the host reveals that there is nothing behind that door. (If you initially picked an empty door, the host reveals the other empty door and if you initially picked the prize door, the host flips a coin to decide which of the others to open.) The host then asks if you want to switch to the door he did not open. Should you switch?

Many people find this unintuitive, but indeed you are better off switching. If you don't switch doors, you win if and only if you initially picked the door with the prize. By assumption, the chance that this happens is $\frac{1}{3}$. If you do switch, you win if and only if you initially picked a door without the prize, which happens with probability $\frac{2}{3}$.

Example 1.2. (Sibling Paradox)

- You visit a town in which every household has exactly two children, each of which is equally likely to be a boy or a girl. You meet a parent at random who says no when you ask if they have two sons. What is the probability that they have two daughters?

  Writing the sibling combinations as (oldest sex, youngest sex), the possibilities are (G,G), (G,B), (B,G), and each is equally likely. Since only one of these three corresponds to two daughters, the probability is $\frac{1}{3}$.

  (To convince yourself, note that this is equivalent to tossing two coins and asking what is the chance that you get two heads if you know that you got at least one. Choosing the convention of how to order the siblings is like distinguishing between the two coins. You are as likely to get two different values as two of the same and just as likely to get two heads as two tails...)

- What if they say yes when you ask if their oldest child is a girl?

  The possibilities now are (G,G) and (G,B) and they are equally likely, so $\frac{1}{2}$.

- What if they say yes if you ask whether they have a daughter who was born on a Tuesday?

  Writing (sex,birthday of oldest; sex,birthday of youngest), the possibilities are

  $$\begin{array}{llll} (G,Tu;G,Mo) & (G,Tu;G,Tu) & \cdots & (G,Tu;G,Su) \\ (G,Tu;B,Mo) & (G,Tu;B,Tu) & \cdots & (G,Tu;B,Su) \\ (G,Mo;G,Tu) & (G,We;G,Tu) & \cdots & (G,Su;G,Tu) \\ (B,Mo;G,Tu) & (B,Tu;G,Tu) & \cdots & (B,Su;G,Tu) \end{array}$$

  so the probability is $\frac{13}{27}$.

- The manner in which we obtained this information was important. For example, if we ascertained that the parent had at least one girl because a girl walked up and introduced herself as the daughter, this is presumably twice as likely to happen in a two girl household, so the probability of two girls is $\frac{1}{2}$ when we learn that there is at least one girl in this fashion.

Example 1.3. (Birthday paradox)

Assuming that everyone is equally likely to be born on any of the 365 days of the year (so we are excluding leap year and ignoring things like people being more likely to be born on weekdays when more doctors are working), the probability that any two people share a birthday is 1/365.

With this in mind, how many people do you need to assemble in a room so that there is at least a 50% chance that at least one pair of people in the room share a birthday?

The answer here turns out to be 23. One explanation is that the probability that no two people in a group of n share a birthday is

$$\frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-(n-1)}{365}.$$

To see this, label the people $1, 2, \ldots, n$. The probability that person 2 has a different birthday than person 1 is $\frac{364}{365}$, the probability that person 3 has a different birthday than persons 1 and 2 is $\frac{363}{365}$,..., and the probability that person n has a different birthday than persons $1, \ldots, n-1$ is $\frac{365-(n-1)}{365}$.

You can check that this is greater than 1/2 when $n \leq 22$ and less than 1/2 when $n \geq 23$.

Roughly, what's going on is that with 23 people in a room, there are $(23 \cdot 22)/2 = 253$ different pairs of people: 23 choices for the first, 22 for the second, and we divide by 2 since (i, j) is the same pair as (j, i).

Ignoring the fact that there are some small dependence relations (if we know that i and j have the same birthday and that j and k do as well, then we automatically know that i and k have the same birthday), we can approximate the probability of no common birthdays in a group of n people as

$$(364/365)^{\binom{n}{2}}$$

where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the number of pairs in a group of n and 364/365 is the chance that a particular pair does not share a birthday. We can compute that $\binom{22}{2} = 231$ and $(364/365)^{231} \approx 0.5306$ while $(364/365)^{253} \approx 0.4995$.

How many people do you need to assemble in a room so that it is more likely than not that at least one of them has your birthday?

The chance that any particular person does not have your birthday is 364/365, so the chance that n people all have different birthdays than you is $(364/365)^n$. We need to choose n big enough that

$$(364/365)^n \leq 1/2$$

or

$$n \log(364/365) \leq \log(1/2)$$

or

$$n \geq \frac{\log(1/2)}{\log(364/365)} = 252.6519...,$$

hence you need 253 additional people in the room to make it more likely than not that someone has your birthday.

How many people do you need to guarantee that at least one pair shares a birthday?

How many to guarantee at least one shares your birthday?

(Answer: 366 and no answer, respectively)

Example 1.4. (Infinite Urn)

At $\frac{1}{2}$ hr til midnight you put balls labeled $1, 2, ..., 10$ into an urn and remove one of them.
At $\frac{1}{4}$ hr til midnight you add balls labeled $11, 12, ..., 20$ and remove one of the balls from the urn.
At $\frac{1}{8}$ hr til midnight you add balls labeled $21, 22, ..., 30$ and remove one of the balls from the urn.
And so forth... At midnight, you inspect the urn.

- If at each stage you remove the ball with the highest label, how many balls are in the urn at midnight?

  There are infinitely many - each ball with label in $\mathbb{N} \setminus 10\mathbb{N}$ remains.

- If at each stage you remove the ball with the lowest label, how many remain?

  Zero! For each $n \in \mathbb{N}$, the ball labeled n was removed $\frac{1}{2^n}$ hours til midnight and so cannot be in the urn.

- Suppose that you add balls $1, ..., 9$ at stage 1, balls $11, ..., 19$ at stage 2, and so on. You never remove any balls, but at each stage you paint a zero after the label of the lowest ball. Then at all times before midnight, the labels of balls in the urn are exactly as in the previous example, so the urn must be empty at midnight, despite the fact that you never removed any balls!
- We will show that if at each stage a random ball is removed, then the urn is empty at midnight with probability one.

- Suppose there is a lamp in the room. At the first stage you turn it on. At the second you turn it off. In general, at each stage you flip the lamp switch. Is the light on or off at midnight?

The way to interpret this stuff is in terms of formal limits of sets and to acknowledge that some liberties are being taken with the language:
Let $A_n \subseteq \mathbb{N}$ be the set of labels in the urn $\frac{1}{2^n}$ hours before midnight. We say that $A_n \to A$ if for every $x \in A$, there is an $N \in \mathbb{N}$ such that $x \in A_n$ for all $n \geq N$. Using this definition, we are thinking of the contents of the urn at midnight as A and the reasoning is perfectly sound.
In the third case, the balls got pushed off to infinity. The last case can be represented as $A_i = \begin{cases} \{1\}, & i \text{ odd} \\ \{0\}, & i \text{ even} \end{cases}$ so $A = \emptyset$.

Induction. The *Principle of Induction* states that if P is a statement about integers such that

(1) $P(i)$ is true for some $i \in \mathbb{Z}$ (base case)
(2) For every $k \in \mathbb{Z}$ with $k \geq i$, if $P(k)$ is true, then $P(k+1)$ is true (inductive step)

Then $P(n)$ is true for all $n \in \mathbb{Z}$ with $n \geq i$.

(The inductive hypothesis can be strengthened to "$P(j)$ is true for all $i \leq j \leq k$...")

Example 2.1. $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$ for all $n \in \mathbb{N}$.

Proof.

Base Case: $\sum_{k=1}^{1} k = 1 = \frac{1(1+1)}{2}$.

Inductive step: Suppose that $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$. Then

$$\sum_{k=1}^{n+1} k = \sum_{k=1}^{n} k + (n+1) = \frac{n(n+1)}{2} + (n+1) = \frac{n(n+1)+2(n+1)}{2} = \frac{(n+1)[(n+1)+1]}{2}. \qquad \square$$

Example 2.2. $3^{2n} - 1$ is divisible by 8 for all $n \in \mathbb{N}_0$.

Proof.

Base Case: $3^{2 \cdot 0} - 1 = 0$ is divisible by 8.

Inductive Step: Suppose that $8 | 3^{2n} - 1$. Then

$$3^{2(n+1)} - 1 = 3^{2n} \cdot 9 - 1 = 3^{2n}(8+1) - 1 = 3^{2n} \cdot 8 + (3^{2n} - 1)$$

is divisible by 8 since it is a sum of multiples of 8. $\square$

Basic Principle of Counting. The *basic principle of counting* states that if one experiment has m possible outcomes and another has n possible outcomes, then there are mn possible outcomes when both are performed. To see that this is so, observe that we can enumerate the simultaneous outcomes of the two experiments as

$$\begin{array}{cccc} (1,1) & (1,2) & \cdots & (1,n) \\ (2,1) & (2,2) & \cdots & (2,n) \\ \vdots & \vdots & \ddots & \vdots \\ (m,1) & (m,2) & \cdots & (m,n) \end{array}$$

where outcome (i,j) means that the first experiment resulted in outcome i and the second in outcome j.

In general, if r experiments are performed with experiment i having n_i possible outcomes, $i = 1, ..., n$, then there are $\prod_{i=1}^{r} n_i = n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

This can be proved by induction. The base case ($r = 2$) has already been established, and for the inductive step, lump the first n of the $n+1$ experiments together as a single experiment and apply the previous result.

Example 2.3. Suppose a computer username consists of 2 letters followed by 3 numbers. How many possible usernames are there?

There are 26 choices for the first character, 26 choices for the second, 10 for the third,..., so $26^2 \cdot 10^3$.

Example 2.4. Suppose a coin is flipped n times. How many possible sequences of heads and tails are there?

There are 2 possibilities for each of the n terms in the sequence, so 2^n.

Example 2.5. Let S be a set with n elements. How many subsets does S have?

Subsets are uniquely determined by checking whether or not each element of S is a member. That is, in response to n yes/no questions, so 2^n.

Example 2.6. In the username example, what if repeated characters are forbidden?

There are 26 choices for the first character, 25 for the second, 10 for the third,..., so $26 \cdot 25 \cdot 10 \cdot 9 \cdot 8$.

Example 2.7. How many arrangements are there of a deck of 52 cards?

There are 52 choices for the top card, 51 remaining choices for the next card,...,$52-k$ for the card k^{th} from the top,..., so $52 \cdot 51 \cdot 50 \cdots 3 \cdot 2 \cdot 1 \approx 8.066 \times 10^{67}$.

* The sheer magnitude of this number not only makes brute force computations impossible, but it also means that there is a good chance that when you shuffle a deck of cards that have been thoroughly mixed, the result is an ordering that has never before been produced!

In general, there are $n! = \prod_{k=1}^{n} k$ different ways to (linearly) arrange n objects. By labeling the objects $1, ..., n$, we see that each such arrangement corresponds to a bijection $\pi : [n] \to [n]$ (where $[n] = \{1, 2, ..., n\}$). Here, $\pi(i)$ is the label of the object in position i. Such bijections are called permutations, and the set of all permutations of $[n]$ is denoted S_n. Reasoning as before (or using induction), we see that $|S_n| = n!$.

Example 2.8. We know that there are $n!$ ways to arrange n people in a line. How many ways can we arrange them in a circle if two arrangements are considered equivalent whenever one can be obtained from the other by rotation?

Each permutation gives a circular arrangement by placing the person at the head of the line at the top of the circle and then lining the rest up in clockwise order. But this overcounts by a factor of n, so the answer is $\frac{n!}{n} = (n-1)!$.
Alternatively, we could adopt the convention that person 1 is always at the top. Then the circular arrangements are determined by the $(n-1)!$ ways to arrange the remaining people.

Example 2.9. Suppose we want to line up n people, but person i and person j are in a fight and cannot be next to one another.

The number of allowed configurations is given by subtracting the number of arrangements with i and j adjacent from $n!$. If we treat the pair as a single person, we see that there are $(n-1)!$ such arrangements with i before j and $(n-1)!$ with j before i. The answer is thus $n! - 2(n-1)! = (n-2)(n-1)!$.
Alternatively, there are $(n-1)!$ ways to line up everyone but person j and $[(n-1)+1]-2 = n-2$ available positions to then place person j.

Example 2.10. If I have 4 math books, 3 chemistry books, 3 philosophy books, and 2 history books, how many ways can they be lined up so that books concerning the same topic are adjacent?

There are 4! ways to permute the math books amongst themselves, 3! for the chemistry books, and so forth, and there are 4! ways to arrange the subjects. The answer is thus 4!4!3!3!2!.

Example 2.11. How many anagrams are there of the word PEPPER?

There are 6! arrangements of $P_1E_1P_2P_3E_2R$. But we don't want to distinguish between the $P's$ or the $E's$ so we have to divide out the overcount to obtain $\dfrac{6!}{3!2!}$.

In general, if there are $n = \sum_{i=1}^{k} n_i$ objects of which there are n_i of type i, $i = 1, ..., k$, there are $\dfrac{n!}{n_1!n_2!\cdots n_k!}$ ways to arrange them when objects of the same type are regarded as indistinguishable.

Example 2.12. Consider the library a couple of examples back. There are $\dfrac{12!}{4!3!3!2!}$ ways to arrange the books (without constraints) if we only care about the relative ordering of the subjects.

3. Combinatorics II

Combinations.

Suppose that we want to choose r objects from a set of n objects where $n \geq r \geq 0$.

If we care about the order in which the objects are selected, then there are

$$n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

possibilities - n choices for the first object sampled, $n-1$ for the next, all the way down to $n-(r-1)$ choices for the last.

Note that each group of objects $\{x_{i_1}, ..., x_{i_r}\}$ appears in $r!$ of the ordered r-tuples in this enumeration, once for each permutation of $\{i_1, ..., i_r\}$.

Thus if we do not care about the order of selection, the total number of samples of size r from a set of size n is $\dfrac{n!}{r!(n-r)!}$.

This expression shows up so often in counting formulas that we give it a special symbol, $\binom{n}{r}$, which is read as "n choose r."

A simple but useful observation is that $\binom{n}{r} = \binom{n}{n-r}$.

By convention $0! = 1$, so we have $\binom{n}{0} = \binom{n}{n} = 1$, which makes sense as there is only one way to choose all or no objects from a set of size $n \in \mathbb{N}_0$.

Though we have only defined $\binom{n}{r}$ for $n \geq r \geq 0$, we will find it convenient to extend the definition to all $r, n \in \mathbb{Z}$ by adopting the convention that $\binom{n}{r} = 0$ whenever $r < 0$ or $r > n$.

Example 3.1. A group of 8 men and 6 women are to be divided into a committee consisting of 3 men and 4 women. The number of possible committees is $\binom{8}{3}\binom{6}{4}$.

Example 3.2. A group of 12 people are to be divided into 3 committees of sizes 3, 4, and 5. There are $\binom{12}{3}\binom{9}{4}\binom{5}{5} = \binom{12}{3}\binom{9}{4}$ ways for this to happen.

Example 3.3. Consider a communication network consisting of n antennae arranged in a row. Suppose that the antennae are spaced so that the network is able to transmit signals as long as no two consecutive antennae are broken. If m of the antennae are broken, how many possible orderings leave the network functional (where we only distinguish between functioning and broken antennae)?

Imagine that we line up the $n-m$ functioning antennae amongst themselves. The network will remain operative as long as the spaces between the functioning antennae have at most 1 broken one. That is, we must choose m of the $n-m+1$ possible slots to place the broken antennae. The answer is thus $\binom{n-m+1}{m}$.

Example 3.4. Suppose that a committee of size r is to be formed from a group containing n women and 1 man. We know that there are $\binom{n+1}{r}$ possibilities. The number of possible committees that are composed entirely of women is $\binom{n}{r}$ and the number of possible committees which include the man is $\binom{n}{r-1}$. Since these are the only two possibilities and they are distinct, we have that

$$\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}.$$

Observe that with the convention $\binom{m}{k} = 0$ for $k < 0$ or $k > m$, this identity holds for all $n, r \in \mathbb{Z}$.

The binomial theorem.

The symbols $\binom{n}{r}$ are commonly referred to as *binomial coefficients* because of their prominence in the binomial theorem.

Theorem 3.1. *For any $x, y \in \mathbb{R}$, $n \in \mathbb{N}$, we have*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Proof. We will proceed by induction. When $n = 1$, we have

$$(x+y)^1 = y + x = \binom{1}{0} x^0 y^1 + \binom{1}{1} x^1 y^0.$$

Now suppose that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Then

$$\begin{aligned}
(x+y)^{n+1} &= (x+y)(x+y)^n = (x+y)\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \\
&= \sum_{k=0}^{n} \binom{n}{k} x^{k+1} y^{n-k} + \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k+1} \\
&= \sum_{j=1}^{n+1} \binom{n}{j-1} x^j y^{n-(j-1)} + \sum_{k=0}^{n} \binom{n}{k} x^k y^{n+1-k} \\
&= \sum_{j=1}^{n+1} \binom{n+1}{j} x^j y^{n-(j-1)} - \sum_{j=1}^{n+1} \binom{n}{j} x^j y^{n-(j-1)} + \sum_{k=0}^{n} \binom{n}{k} x^k y^{n+1-k} \\
&= \sum_{j=1}^{n+1} \binom{n+1}{j} x^j y^{(n+1)-j} - \sum_{k=1}^{n+1} \binom{n}{k} x^k y^{n+1-k} + \sum_{k=0}^{n} \binom{n}{k} x^k y^{n+1-k} \\
&= \sum_{j=1}^{n+1} \binom{n+1}{j} x^j y^{(n+1)-j} - \binom{n}{n+1} x^{n+1} y^0 + \binom{n}{0} x^0 y^{n+1} \\
&= \sum_{j=1}^{n+1} \binom{n+1}{j} x^j y^{(n+1)-j} + \binom{n+1}{0} x^0 y^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} x^k y^{(n+1)-k}. \quad \square
\end{aligned}$$

The binomial theorem also follows from a simple combinatorial argument:

Expanding $(x+y)^n = (x+y)(x+y)\cdots(x+y)$, we see that every term is a product of k x's and $n-k$ y's for some $k = 0, 1, ..., n$ since each of the n binomial terms contributes either an x or a y. The coefficient of $x^k y^{n-k}$ is the number of ways of selecting k of the terms to contribute an x.

Example 3.5. The binomial theorem provides another proof that a set of size n has 2^n distinct subsets. Namely, for each $k = 0, 1, ..., n$, there are $\binom{n}{k}$ subsets of size k. The total number of subsets is thus

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^k \cdot 1^{n-k} = (1+1)^n.$$

Sampling and Integer Solutions to Equations.

Often in probability and statistics, one is interested in drawing a sample of size r from a population of size n. This can be done with replacement (so that the same object may be counted more than once in the sample) or without. (Note that if we sample without replacement, then we must assume that $n \geq r$.)
Also, we may or may not care about the order in which the objects were sampled.

If we sample with replacement and order matters, there are n^r ways to draw a sample of size r from a population of size n. (There are n possibilities for the first, second, third,...)

If we sample without replacement and order matters, there are $\dfrac{n!}{(n-r)!}$ ways to draw a sample of size r from a population of size n. (There are n possibilities for the first, $n-1$ for the second,..., and $n-r+1$ for the rth.)

If we sample without replacement and order doesn't matter, there are $\binom{n}{r}$ such samples. (This is essentially the definition of $\binom{n}{r}$ and is obtained by dividing out the number of ways to permute the r sampled objects from the previous case.)

To count the number of samples with replacement where order doesn't matter is a bit more subtle.

Perhaps the easiest way to think of the situation is to picture r identical balls distributed amongst bins labeled $1, ..., n$. The number of balls in bin i corresponds to the number of times object i was sampled.

The n bins can be represented by $n-1$ interior walls where wall i separates bins i and $i+1$ for $i = 1, ..., n-1$. An arrangement of r balls in n bins is then represented as a sequence of r balls and $n-1$ walls.

In terms of sampling, the number of balls to the left of all walls is the number of times object 1 was sampled, the number of balls between walls i and $i+1$ is the number of times object $i+1$ was sampled for $i = 1, ..., n-2$, and the number of balls to the right of all walls is the number of times object n was sampled.

Since an arrangement of r balls and $n-1$ walls is determined by choosing which of the $r+n-1$ positions the balls occupy, the desired number of samples is $\binom{r+n-1}{r}$.

Example 3.6. This kind of "balls and walls" argument can also be used to study integer compositions. Given $n, k \in \mathbb{N}$, a composition of n into k parts is a solution of $x_1 + ... + x_k = n$ with $x_1, ..., x_k \in \mathbb{N}$. (In contrast to integer partitions, the order of the summands is considered relevant.)

Writing $n = 1 + ... + 1$, we see that a composition of n into k parts is determined by selecting $k-1$ of the the $n-1$ positions between the ones (balls) to place right parentheses (walls).

For example, if $k = 3$ and $n = 7$, the composition $x_1 = 2$, $x_2 = 4$, $x_3 = 1$ is represented as $1 + 1) + 1 + 1 + 1 + 1) + 1$. It follows that the number of compositions of n into k parts is $\binom{n-1}{k-1}$.

An immediate corollary is that the total number of compositions of n is

$$\sum_{k=1}^{n} \binom{n-1}{k-1} = \sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1}.$$

Example 3.7. Suppose that instead we are interested in the number of nonnegative integer solutions to $x_1 + ... + x_k = n$.
One approach would be to note that this is equivalent to the number of ways of distributing n balls amongst k bins, which we now know to be $\binom{n+k-1}{n}$.

Alternatively, there is a natural bijection between nonnegative integer solutions to $x_1 + ... + x_k = n$ and positive integer solutions to $y_1 + ... + y_k = n + k$: Just take $y_i = x_i + 1$. Thus, by the previous example, the number of nonnegative integer solutions to $x_1 + ... + x_k = n$ is $\binom{n+k-1}{k-1} = \binom{n+k-1}{n}$.

This gives another way to enumerate the number of samples without replacement where order is ignored. It's not quite as obvious as the original argument, but is perhaps easier to follow since the problem is broken up into more stages.

Multinomial Coefficients. Suppose that a group of n objects is to be divided into r distinct groups of sizes $n_1, ..., n_r$, respectively, $n_1 + \ldots n_r = n$.
There are $\binom{n}{n_1}$ choices for the members of the first group, $\binom{n-n_2}{n_2}$ choices for the second group, and so forth, giving a total of

$$\begin{aligned}\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-...-n_{r-1}}{n_r} \\ = \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!}\cdots\frac{(n-n_1-...-n_{r-1})!}{n_r!0!} \\ = \frac{n!}{n_1!n_2!\cdots n_r!}\end{aligned}$$

possibilities.

Alternatively, for each of the $n!$ permutations of the objects, we can assign the first n_1 to group 1, the next n_2 to group 2, etc... Since permuting objects within a group does not change the grouping as a whole, we have to divide out by $n_1!\cdots n_r!$ to get the total number of groupings.

These expressions are a natural generalization of binomial coefficients (which correspond to $r = 2$), so they get an analogous symbol:
If $n_1, ..., n_r \in \mathbb{N}_0$ with $n_1 + ... + n_r = n$, we define

$$\binom{n}{n_1, ..., n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.$$

The above is known as a multinomial coefficient because of its role in the multinomial theorem

Theorem 3.2. *For any $x_1, ..., x_r \in \mathbb{R}$, $n \in \mathbb{N}$,*

$$(x_1 + ... + x_r)^n = \sum_{\substack{(n_1,...,n_r): \\ n_1+...+n_r=n}} \binom{n}{n_1, ..., n_r} x_1^{n_1} \cdots x_r^{n_r}.$$

Theorem 3.2 can be proved combinatorially or by induction on r using the binomial theorem to treat the last term of $(x_1 + ... + x_r + x_{r+1})^n = (x_1 + ... + (x_r + x_{r+1}))^n$.

An easy corollary to the multinomial theorem is the identity

$$r^n = \sum_{\substack{(n_1,...,n_r): \\ n_1+...+n_r=n}} \binom{n}{n_1, ..., n_r}.$$

To connect with previous stories, we can think of $\binom{n}{n_1, ..., n_r}$ as the number of ways to distribute n distinctly labeled balls in r bins, with n_i the number of balls in bin i.

4. The Basic Setup

The typical paradigm in probability is that one is interested in the outcome of a random *experiment*, which we define to be a process whose outcome is unknown in advance.

The set of all possible outcomes is called the *sample space*, which we write as Ω. A generic *outcome* is typically denoted ω.

- If the experiment is rolling a six-sided die, then the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- If the experiment is flipping two quarters, the sample space is $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$.
- If the experiment is shuffling a deck of cards labeled $1, 2, ..., n$, then we can write the sample space as $\Omega = S_n$ where the outcome π means that the label of the card at the top of the deck is $\pi(1)$, that of the second card is $\pi(2)$, and so forth.
- If the experiment is measuring the lifetime (in hours, say) of a light bulb, the sample space is $\Omega = \{x \in \mathbb{R} : x \geq 0\}$.

Prior to performing the experiment, there are a number of questions we can ask, all of which can be phrased as "Will the experiment result in this particular set of outcomes?"

The collection of all subsets of Ω corresponding to potentially answerable questions is the *σ-field* $\mathcal{F}$. An element $E \in \mathcal{F}$ is called an *event*.

- In the roll of a die experiment, $\{1, 3, 5\}$ is the event that an odd number was rolled.
- In the two quarters experiment, $\{(H, H), (T, T)\}$ is the event that both quarters landed on the same side. The event $\{(H, H)\}$ consists of the single outcome that both quarters showed heads.
- In the light bulb experiment, $\{x \in \mathbb{R} : x > 2\}$ is the event that the bulb lasted more than 2 hours.

Events can be combined combined using various set theoretic operations.

- If E and F are events, then their union is the new event $E \cup F = \{\omega \in \Omega : \omega \in E \text{ or } \omega \in F\}$, the set of all outcomes which belong to either E or F.
- Similarly, the intersection of E and F is the event $E \cap F = \{\omega \in \Omega : \omega \in E \text{ and } \omega \in F\}$, the set of all outcomes which belong to both E and F.
- Finally, we the complement of E is the event $E^C = \{\omega \in \Omega : \omega \notin E\}$, the set of all outcomes which do not belong to E.

As a running example, we consider the roll of a die experiment. Let $E = \{1, 2, 3\}$ be the event that a value less than 4 is rolled, $F = \{1, 3, 5\}$ that an odd value is rolled, and $G = \{2\}$ that a 2 is rolled.

Note that the event G is contained in the event E, which we write as $G \subseteq E$.

Also, G and F have no outcomes in common, and we say that F and G are *disjoint*.

- The unions are given by $E \cup F = \{1, 2, 3, 5\}$, $E \cup G = \{1, 2, 3\}$, and $F \cup G = \{1, 2, 3, 5\}$.
- The intersections are $E \cap F = \{1, 3\}$, $E \cap G = \{2\}$, and $F \cap G = \emptyset$.

  * Observe that $E \cup G = E$ and $E \cap G = G$. This is true in general for $G \subseteq E$.
  The expression $F \cap G = \emptyset$ is equivalent to the statement that F and G are disjoint.
  ($\emptyset$ is called the empty set because it is the set containing no elements.)
- The complements are $E^C = \{4, 5, 6\}$ (a value larger than 4), $F^C = \{2, 4, 6\}$ (an even value), and $G^C = \{1, 3, 4, 5, 6\}$ (a value other than 2).

One can take unions and intersections of more than one event:

If $E_1, ..., E_n$ are events, then their union is defined as

$$\bigcup_{i=1}^{n} E_i = \{\omega \in \Omega : \omega \in E_i \text{ for some } i = 1, ..., n\},$$

and their intersection is defined by

$$\bigcap_{i=1}^{n} E_i = \{\omega \in \Omega : \omega \in E_i \text{ for all } i = 1, ..., n\}.$$

Likewise, we can take countable unions and intersections:

Given events $E_1, E_2, ...$, we have

$$\bigcup_{i=1}^{\infty} E_i = \{\omega \in \Omega : \omega \in E_i \text{ for some } i \in \mathbb{N}\},$$
$$\bigcap_{i=1}^{\infty} E_i = \{\omega \in \Omega : \omega \in E_i \text{ for all } i \in \mathbb{N}\}.$$

Intersections and unions over other index sets are defined similarly.

These set operations can also be combined to create more complicated events.

Useful examples are the set difference

$$E \setminus F = E \cap F^C = \{\omega \in \Omega : \omega \in E \text{ and } \omega \notin F\}$$

and the symmetric difference

$$E \triangle F = (E \setminus F) \cup (F \setminus E) = (E \cup F) \setminus (E \cap F),$$

the latter consisting of all outcomes that are in exactly one of E or F.

The operations of union and intersection have nice algebraic properties:

Commutativity: $E \cup F = F \cup E$ and $E \cap F = F \cap E$
Associativity: $(E \cup F) \cup G = E \cup (F \cup G)$ and $(E \cap F) \cap G = E \cap (F \cap G)$
Distributivity: $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$ and $E \cup (F \cap G) = (E \cup F) \cap (E \cap G)$

Complementation is a self-inverse operation in the sense that $\left(E^C\right)^C = E$.

Complements also distribute over unions and intersections, but they interchange the two.

This is the content of *DeMorgan's laws*: If $\{E_i\}_{i \in I}$ is a collection of events, then

$$\left(\bigcup_{i \in I} E_i\right)^C = \bigcap_{i \in I} E_i^C,$$
$$\left(\bigcap_{i \in I} E_i\right)^C = \bigcup_{i \in I} E_i^C.$$

A facility with manipulating sets is extremely useful in probability computations.

Having laid the necessary groundwork, we are finally able to introduce randomness into the picture. There are many interpretations of what probability means in the real world – for example, a measure of propensity borne out in long-term relative frequencies, or a measure of belief (or lack of information) constantly being updated in the face of new evidence. The jury is still out on these metaphysical and epistemological issues, but this is math class and we can get around this obstacle by treating the subject axiomatically: We will set forth a couple of reasonable assumptions and then derive consequences which will largely agree with the computations advocated by the various schools of thought on the subject and are remarkably effective at modeling real world phenomena.

The general setting in which we will work is a probability space $(\Omega, \mathcal{F}, P)$. Ω is the sample space consisting of all possible outcomes of the experiment, $\mathcal{F}$ is the collection of all events about which we can make meaningful statements, and P is a function which assigns probabilities to events in $\mathcal{F}$ in a consistent fashion.

In this class, we will always take $\mathcal{F}$ to be the set of all subsets of Ω, and will thus deemphasize its role. This is a pretty innocuous assumption when Ω is countable, but it turns out that there are subsets of uncountable sample spaces to which one cannot reasonably assign certain natural probabilities.

The probability measure P is an $\mathbb{R}$-valued function defined on $\mathcal{F}$ which satisfies

(1) $0 \leq P(E) \leq 1$ for all $E \in \mathcal{F}$
(2) $P(\Omega) = 1$
(3) For any sequence of pairwise disjoint subsets $E_1, E_2, ...$, $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$.

Observe that if we set $E_1 = \Omega$ and $E_i = \emptyset$ for $i > 1$, then we have

$$1 = P(\Omega) = P\left(\bigcup_{i=1}^{\infty} E_i\right) = P(\Omega) + \sum_{i=2}^{\infty} P(E_i) = 1 + \sum_{i=2}^{\infty} P(\emptyset),$$

so $P(\emptyset) = 0$.

Also, note that finite additivity is a special case of Axiom 3: If $E_1, ..., E_n$ are disjoint events, then we can set $E_i = \emptyset$ for $i > n$ to get

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \sum_{i=1}^{n} P(E_i) + \sum_{i=n+1}^{\infty} P(\emptyset) = \sum_{i=1}^{n} P(E_i).$$

Example 4.1. In the roll of a die example, a natural choice of P is $P(E) = \frac{|E|}{6}$ for all $E \subseteq \{1, 2, ..., 6\}$. If we were working with a loaded die, then some faces may be more likely than others. One way to represent this is as follows: Let $p_1, p_2, ..., p_6 \geq 0$ satisfy $p_1 + p_2 + ... + p_6 = 1$. Then $P(E) = \sum_{i \in E} p_i$. The interpretation is that p_i represents the probability that the the value i is observed.

(Actually, all probabilities on countable sample spaces arise in this fashion since Axiom 3 implies that P is determined by its value on singletons in such cases.)

Example 4.2. In the light bulb lifetime experiment, a typical choice for P is something like $P(E) = \int_E e^{-x}dx$ for $E \subseteq [0,\infty)$. Since $\int_0^\infty e^{-x}dx = 1$, we see that Axiom 2 is satisfied. Also, since $e^{-x} \geq 0$ for all $x \geq 0$, it follows from basic properties of the integral that $0 \leq \int_E e^{-x}dx \leq \int_0^\infty e^{-x}dx = 1$ for any $E \subseteq [0,\infty)$ for which the integral is defined. Using the correct definition of the integral, Axiom 3 is also satisfied.
Though this makes good sense for the kind of events one usually cares about in practice, such as $E = (a,b)$, the integral is not defined over all possible subsets of $[0,\infty)$, which indicates that a rigorous treatment of probability needs to be careful about the domain of P. (Maybe explain Vitali sets...)

From these 3 simple axioms, one can derive a whole host of other useful consequences.

Proposition 4.1. *For any event A, $P(A^C) = 1 - P(A)$.*

Proof. Axioms 2 and 3 imply that

$$1 = P(\Omega) = P(A \cup A^C) = P(A) + P(A^C). \qquad \square$$

This is a useful result because it is sometimes much easier to compute $P(A^C)$ than $P(A)$.

Proposition 4.2. *If $E \subseteq F$, then $P(E) \leq P(F)$.*

Proof. Since F is the disjoint union of $F \setminus E$ and $E \cap F = E$, Axioms 3 and 1 imply

$$P(F) = P(E \cap F) + P(F \setminus E) = P(E) + P(F \setminus E) \geq P(E). \qquad \square$$

Proposition 4.3. *For any events E and F, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.*

Proof. Writing "$\sqcup$" for "disjoint union," we have $E \cup F = (E \setminus F) \sqcup (E \cap F) \sqcup (F \setminus E)$, $E = (E \setminus F) \sqcup (E \cap F)$, and $F = (F \setminus E) \sqcup (E \cap F)$. Thus the third axiom implies

$$\begin{aligned} P(E \cup F) &= P(E \setminus F) + P(E \cap F) + P(F \setminus E) \\ &= P(E \setminus F) + P(E \cap F) + P(F \setminus E) + P(E \cap F) - P(E \cap F) \\ &= P(E) + P(F) - P(E \cap F). \end{aligned} \qquad \square$$

Example 4.3. A certain electronic device consists of two components. The first component has probability 0.4 of failing within the first year, the second component has probability 0.6 of failing within the first year, and the probability that both components fail within a year is 0.25. What is the probability that the device is operational after a year?

If we let A be the event that the first component fails and let B be the event that the second component fails, then $A \cup B$ is the event that at least one component has failed by the year's end. By Proposition 4.3,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.6 - 0.25 = 0.75,$$

so Proposition 4.1 implies that the probability that the device is still working after a year is

$$P\left((A \cup B)^C\right) = 1 - P(A \cup B) = 0.25.$$

Proposition 4.3 has a natural extension to unions of more than two events known as the *inclusion-exclusion* identity.

Proposition 4.4. *For any events $E_1, E_2, ..., E_n$,*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} E_{i_j}\right)$$
$$= \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - ... + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right).$$

Proof. The argument in the text is enlightening and it works well for countable probability spaces, but it needs more work to apply in full generality.

(Roughly, it says fix any $\omega \in \bigcup_{i=1}^{n} E_i$. Then ω is in exactly m of the E_i's for some $m = 1, ..., n$. Thus its probability, $P(\{\omega\})$, contributes to exactly $\binom{m}{k}$ of the summands in the k-fold intersections on the right. It's total contribution to the right-hand side is thus $P(\{\omega\})\left(\sum_{k=1}^{m}(-1)^{k+1}\binom{m}{k}\right)$, which is equal to its contribution, $P(\{\omega\}) = \binom{m}{0}P(\{\omega\})$, to the left-hand side since $\sum_{k=0}^{m}\binom{m}{k}(-1)^{k+1} = -(1-1)^m = 0$.)

Instead we give an inductive proof. Proposition 4.3 is the base case.
Now assume that $P\left(\bigcup_{i=1}^{n} F_i\right) = \sum_{k=1}^{n}(-1)^{k+1}\sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} F_{i_j}\right)$ for any events $F_1, ..., F_n$ and let $E_1, ..., E_{n+1}$ be given. Then

$$P\left(\bigcup_{i=1}^{n+1} E_i\right) = P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cup E_{n+1}\right) = P\left(\bigcup_{i=1}^{n} E_i\right) + P(E_{n+1}) - P\left(\left(\bigcup_{i=1}^{n} E_i\right) \cap E_{n+1}\right)$$
$$= P\left(\bigcup_{i=1}^{n} E_i\right) + P(E_{n+1}) - P\left(\bigcup_{i=1}^{n} (E_i \cap E_{n+1})\right)$$
$$= \sum_{k=1}^{n}(-1)^{k+1} \sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} E_{i_j}\right) + P(E_{n+1}) - \sum_{k=1}^{n}(-1)^{k+1} \sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} (E_{i_j} \cap E_{n+1})\right)$$
$$= \sum_{k=1}^{n+1}(-1)^{k+1} \sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} E_{i_j} \cap E_{n+1}\right).$$

□

It can be shown that the probability of the union is upper-bounded by truncating the alternating sum in Proposition 4.4 after a term has been added, and the probability is lower-bounded by truncating the sum after a negative term. This is the content of the *Bonferroni inequalities*.

5. Equiprobable Outcomes

Historically, probability was first considered in terms of finite sample spaces where all outcomes were judged equally likely - that is, $P(E) = \frac{|E|}{|\Omega|}$ for all $E \subseteq \Omega$. We will hone our counting skills by working through several such examples.

Example 5.1. If two dice are rolled, what is the probability that the sum of their values is 7? What is the probability that the second die shows a higher value than the first?

There are 36 possible and equally likely outcomes of the experiment, (i, j) with $i, j = 1, ..., 6$. For each value of i, there is exactly one value of j for which $i + j = 7$. Thus

$$P(i + j = 7) = \frac{6}{36} = \frac{1}{6}.$$

Similarly, for each value of i, there are $6 - i$ values of j which are greater, so

$$P(j > i) = \frac{1}{36}\sum_{i=1}^{6}(6 - i) = \frac{36}{36} - \frac{1}{36} \cdot \frac{6 \cdot 7}{2} = \frac{5}{12}.$$

Example 5.2. In a certain batch of n widgets, m are known to be defective. Suppose that a random subset of k widgets is selected. What is the probability that at most 2 of them are defective? (Let's assume that $n > m$ and $k \geq 2$.)

There are $\binom{n}{k}$ possible samples and all are assumed to be equally likely. The number of such samples that contain exactly j defective widgets for $j \leq k$ is $\binom{m}{j}\binom{n-m}{k-j}$. Thus $P(j \text{ defective}) = \dfrac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}$.

In particular,

$$\begin{aligned} P(\text{at most 2 defective}) &= P(0 \text{ defective}) + P(1 \text{ defective}) + P(2 \text{ defective}) \\ &= \frac{\binom{n-m}{k}}{\binom{n}{k}} + \frac{m\binom{n-m}{k-1}}{\binom{n}{k}} + \frac{\binom{m}{2}\binom{n-m}{k-2}}{\binom{n}{k}}. \end{aligned}$$

Example 5.3. In the game of poker, players are dealt 5 cards each from a 52 card deck. We assume that the deck is thoroughly shuffled so that all $\binom{52}{5}$ possible hands are equally likely. Cards have thirteen values $2, ..., 10, J, ..., A$ (which we take to be ordered as indicated) and each value comes in four suits.

- A full house consists of three cards of one value and two cards of another.
- A straight consists of five consecutive cards not all having the same suit.
- A three of a kind consists of three cards of one value and two cards having two other values.
- A big bobtail consists of four cards of the same suit in consecutive order and another card which does not extend the straight.

How likely is each hand? Should a big bobtail be worth more or less than a full house?

- For a full house, there are 13 choices for the value appearing three times and 12 remaining choices for the value appearing twice. There are 4 possible choices for the suits of the former (by specifying the suit not chosen) and $\binom{4}{2}$ choices for the suits of the latter. Thus

$$P(\text{Full House}) = \frac{13 \cdot 12 \cdot 4 \cdot \binom{4}{2}}{\binom{52}{5}}.$$

- For a straight, there are 9 choices for the value of the lowest card, the other values then being determined. There are 4 choices for the suit of each card, but we must discount the 4 cases for each choice of values where all cards have the same suit. Thus

$$P(\text{Straight}) = \frac{9 \cdot \left(4^5 - 4\right)}{\binom{52}{5}}.$$

- For a three of a kind, we have 13 choices for the value appearing three times and $\binom{12}{2}$ choices for the other values. There are 4 choices for the suits of the first value and 4 choices each for the suits of the other two. Thus

$$P(\text{3-of-a-kind}) = \frac{13 \cdot \binom{12}{2} \cdot 4^3}{\binom{52}{5}}.$$

- For a big bobtail, there are 8 choices for the smallest value in the "mini-straight" which leave 7 choices for the other value $(3, ..., 10)$, and 2 choices for the smallest value $(2, J)$ which leave 8 for the other value. In all cases, there are 4 for the suit of the mini-straight and 4 for the suit of the other card. Thus

$$P(\text{Big Bobtail}) = \frac{(8 \cdot 7 + 2 \cdot 8) \cdot 4^2}{\binom{52}{5}}.$$

Since

$$(8 \cdot 7 + 2 \cdot 8) \cdot 4^2 = 1,152 < 3,744 = 13 \cdot 12 \cdot 4 \cdot \binom{4}{2},$$

a big bobtail should be worth more than a full house.

Example 5.4. A deck of cards is completely shuffled and the cards are turned up one at a time until the first ace appears. Is the next card more likely to be the ace of spades or the two of clubs?

To find the probability that the ace of spades immediately follows the first ace, note that there are 51! ways to order all cards other than the ace of spades, and for each ordering, there is exactly one way to insert the ace of spades so that it is immediately after the first ace. Thus

$$P\left(A\spadesuit \text{ after first } A\right) = \frac{51!}{52!} = \frac{1}{52}.$$

Of course, the exact same argument applies to the two of clubs (or any other fixed card), so the two events are equally likely.

Example 5.5. Suppose that n people enter a restaurant and each leaves their hat at the reception desk. The receptionist is lazy and simply selects a random hat to return to each person after the meal. What is the probability that no one is given their original hat?

For $i = 1, ..., n$ let E_i be the event that person i receives their original hat. Then the probability that at least one person gets their hat back is

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + ... + (-1)^{k+1} \sum_{i_1<...<i_k} P(E_{i_1} \cap \cdots \cap E_{i_k}) + ... + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right)$$

by inclusion-exclusion.
Now for any $k = 1, ..., n$ and any $1 \leq i_1 < ... < i_k \leq n$, $P\left(E_{i_1} \cap \cdots \cap E_{i_k}\right) = \frac{(n-k)!}{n!}$ because there are $(n-k)!$ ways to distribute the remaining hats amongst the other $n-k$ people. Since the number of terms in the sum over $1 \leq i_1 < ... < i_k \leq n$ is $\binom{n}{k}$, we see that

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} E_{i_j}\right) \\ &= \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k!(n-k)!} \frac{(n-k)!}{n!} \\ &= -\sum_{k=1}^{n} \frac{(-1)^k}{k!} = 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!}. \end{aligned}$$

Thus the probability that no one receives their hat is

$$1 - P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=0}^{n} \frac{(-1)^k}{k!} \to e^{-1}.$$

Example 5.6. Suppose that a coin is tossed n times. If we know that the coin came up heads m times, what can we say about the number of head runs? (A head run is sequence of consecutive tosses resulting in heads, so if the outcome was $HHHTHTTTHHT$, then we have three head runs of lengths 3, 1, and 2 respectively.)

Let $1 \leq r \leq m$ be given. We want to know what the probability that a sequence of n coin tosses resulting in m heads contains r head runs. Assuming that all sequences of m heads and $n-m$ tails are equally likely, this is just the number of sequences of m heads and $n-m$ tails consisting of r head runs divided by $\binom{n}{m}$.

We can represent such sequences by $y_1 x_1 y_2 x_2 \ldots y_r x_r y_{r+1}$ where x_i is the length of the i^{th} head run and y_i is the length of the tail run immediately preceding the i^{th} head run for $i = 1, ..., r$, and y_{r+1} is the length of the tail run following the last head run.

By assumption, $x_1, ..., x_r$ are positive integers with $x_1 + ... + x_r = m$, and $y_1, ..., y_{r+1}$ are integers with $y_1, y_{r+1} \geq 0$, $y_2, ..., y_r > 0$, and $y_1 + ... + y_{r+1} = n - m$.

By previous arguments, the number of such $(x_1, ..., x_r)$ is $\binom{m-1}{r-1}$ (determined by placement of $r-1$ right parentheses in the $m-1$ positions between m ones), and the number of $(y_1, ..., y_{r+1})$ is the same as the number of positive integer vectors $(z_1, ..., z_{r+1})$ with $z_1 + ... + z_{r+1} = n - m + 2$ (taking $z_1 = y_1 + 1$, $z_{r+1} = y_{r+1} + 1$, and $z_i = y_i$ for $2 \leq i \leq r$), which is $\binom{n-m+2-1}{r+1-1} = \binom{n-m+1}{r}$.

Thus, by the basic principle of counting

$$P(r \text{ head runs}) = \frac{\binom{m-1}{r-1}\binom{n-m+1}{r}}{\binom{n}{m}}.$$

6. Additional Properties of P

Proposition 6.1. *If $E_1, E_2, \ldots$ is any sequence of events, then $P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$.*

Proof. Define $F_1 = E_1$ and $F_k = E_k \setminus \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Then $F_n \subseteq E_n$ for all $n \in \mathbb{N}$ and $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i$. Moreover, the events $F_1, F_2, \ldots$ are disjoint.

It follows from Axiom 3 and Proposition 4.2 that

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) \leq \sum_{i=1}^{\infty} P(E_i). \qquad \square$$

Proposition 6.2. *If $E_1, E_2, \ldots$ is any sequence of events with $E_1 \subseteq E_2 \subseteq \ldots$, then $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{n\to\infty} P(E_n)$.*

Proof. Define $F_1 = E_1$ and $F_k = E_k \setminus \bigcup_{i=1}^{k-1} E_i$ for $k \geq 2$. Then the F_i's are disjoint, $E_n = \bigcup_{i=1}^{n} F_i$, and $\bigcup_{i=1}^{\infty} E_i = \bigcup_{i=1}^{\infty} F_i$. It follows that

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} F_i\right) = \lim_{n\to\infty} P(E_n). \qquad \square$$

Proposition 6.3. *If $E_1, E_2, \ldots$ is any sequence of events with $E_1 \supseteq E_2 \supseteq \ldots$, then $P\left(\bigcap_{i=1}^{\infty} E_i\right) = \lim_{n\to\infty} P(E_n)$.*

Proof. Define $F_n = E_n^C$ for $n \in \mathbb{N}$. Then $F_1 \subseteq F_2 \subseteq \ldots$ and $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i^C = \left(\bigcap_{i=1}^{\infty} E_i\right)^C$. It follows from Propositions 4.1 and 6.2 that

$$\begin{aligned} P\left(\bigcap_{i=1}^{\infty} E_i\right) &= 1 - P\left(\bigcup_{i=1}^{\infty} F_i\right) = 1 - \lim_{n\to\infty} P(F_n) \\ &= \lim_{n\to\infty} \left(1 - \lim_{n\to\infty} P(F_n)\right) = \lim_{n\to\infty} P\left(F_n^C\right) = \lim_{n\to\infty} P\left(E_n\right). \end{aligned} \qquad \square$$

Propositions 6.2 and 6.3 are known as *continuity from below* and *continuity from above*, respectively.

Example 6.1. Recall the infinite urn experiment from the first lecture:

At $\frac{1}{2}$ hour til midnight we add balls labeled $1, \ldots, 10$ and remove one ball.

At $\frac{1}{4}$ hour til midnight we add balls labeled $11, \ldots, 20$ and remove one ball.

At $\frac{1}{8}$ hour til midnight we add balls labeled $21, \ldots, 30$ and remove one ball...

We saw that if we always removed the ball with the highest label, there would be infinitely many balls in the urn at midnight, whereas if we removed the ball with the lowest label, the urn would be empty at midnight.

Suppose now that at each stage we remove a ball chosen uniformly at random from the urn. We claim that $P(\text{urn empty at midnight}) = 1$.

To see that this is so, it suffices to show that for any $n \in \mathbb{N}$, letting A_n denote the event that ball n remains at midnight, we have $P(A_n) = 0$.

(This will imply $P(\text{urn empty at midnight}) = 1 - P\left(\bigcup_{n=1}^{\infty} A_n\right) \geq 1 - \sum_{n=1}^{\infty} P(A_n) = 1$.)

Now ball n was first added at stage m where $10(m-1) < n \leq 10m$. Let F_k^n be the event that ball n remains at stage $m + k - 1$.

By construction, $F_1^n \supseteq F_2^n \supseteq \ldots$ and $A_n = \bigcap_{k=1}^{\infty} F_k^n$, so continuity from above implies $P(A_n) = \lim_{k\to\infty} P(F_k^n)$.

To complete the argument, we observe that

$$\begin{aligned} P(F_k^n) &= \frac{9m}{9m+1} \cdot \frac{9(m+1)}{9(m+1)+1} \cdots \frac{9(m+k-1)}{9(m+k-1)+1} \\ &= \prod_{j=m}^{m+k-1} \frac{9j}{9j+1} = \prod_{j=m}^{m+k-1} \left(1 - \frac{1}{9j+1}\right) \\ &\leq \prod_{j=m}^{m+k-1} \exp\left(-\frac{1}{9j+1}\right) = \exp\left(-\sum_{j=m}^{m+k-1} \frac{1}{9j+1}\right), \end{aligned}$$

which goes to zero as $k \to \infty$.

* The last expression goes to zero since the harmonic series diverges to infinity.
The bound $1 - x \leq e^{-x}$ follows by observing that $f(x) = e^{-x} - 1 + x$ has a global minimum at $x = 0$.

Suppose that $A_1, A_2, \ldots$ is a sequence of events. We define the limit supremum of the sequence by

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

Often we abbreviate the event $\limsup_{n\to\infty} A_n$ as $\{A_n \text{ i.o.}\}$ where i.o. stands for infinitely often. If we write $B_n = \bigcup_{m=n}^{\infty} A_m$, then B_m is the event that at least one A_m occurs for $m \geq n$. Thus $\{A_n \text{ i.o.}\} = \bigcap_{n=1}^{\infty} B_n$ is the event that infinitely many of the A_n occur: $\omega \in \bigcap_{n=1}^{\infty} B_n$ if and only if for every $n \in \mathbb{N}$, there is an $m \geq n$ with $\omega \in A_m$.

Proposition 6.4. *Let $A_1, A_2, \ldots$ be a sequence of events. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.*

Proof. Writing $B_n = \bigcup_{m=n}^{\infty} A_m$, we see that $B_1 \supseteq B_2 \supseteq \ldots$, so Proposition 6.3 implies

$$P\left(A_n \text{ i.o.}\right) = P\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} P(B_n).$$

Also, it follows from Proposition 6.1 that

$$P\left(B_n\right) = P\left(\bigcup_{m=n}^{\infty} A_m\right) \leq \sum_{m=n}^{\infty} P(A_m).$$

Finally, if $\sum_{n=1}^{\infty} P(A_n)$ is convergent, then for every $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$\sum_{m=n+1}^{\infty} P(A_m) = \left| \sum_{m=1}^{\infty} P(A_m) - \sum_{m=1}^{n} P(A_m) \right| < \varepsilon$$

whenever $n \geq N$, hence $\lim_{n\to\infty} \sum_{m=n}^{\infty} P(A_m) = 0$.

Combining these observations shows that if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then

$$P\left(A_n \text{ i.o.}\right) = \lim_{n\to\infty} P(B_n) \leq \lim_{n\to\infty} \sum_{m=n}^{\infty} P(A_m) = 0. \qquad \square$$

7. Conditional Probability

Conditional probability addresses the problem of updating one's assessment of probabilities when presented with additional information. For example, suppose that I toss two fair coins in another room. From your point of view, the probability that both coins land on heads is $\frac{1}{4}$. However, if I tell you that one of the coins landed on tails, then you would know that the outcome (H, H) did not occur. Your evaluation of the probability changes when given additional partial information. Similarly, if I told you that the first coin landed on heads, then the only two possible outcomes are (H, H) and (H, T). Since these outcomes were originally equally likely, they should remain so after you learned that the first coin landed on heads. Thus you should now evaluate the odds of two heads as $\frac{1}{2}$.

In general, we are interested in computing $P(E\,|F)$, the probability that the event E occurs given the information that F has occurred.

Essentially, knowing that F has occurred means that we need to shrink our original sample space Ω to only include outcomes in F. The possible events are thus of the form $A \cap F$: In order for A to have occurred, the experiment must have resulted in an outcome in $A \cap F$. Because we want our updated probability to satisfy the necessary axioms, it must assign F probability 1. Since we don't want to change the relative probabilities of events contained in F (as our information about their relative likelihoods has not changed), we have to normalize the original probability by dividing by $P(F)$. Of course, for this to work, we need to assume that F had positive probability to begin with.

To summarize, suppose that a random experiment is initially described by a probability space $(\Omega, \mathcal{F}, P)$. If F is an event with $P(F) > 0$, then after learning that F has occurred, we need to update our probability space to $(\Omega', \mathcal{F}', P')$ where $\Omega' = F$, $\mathcal{F}' = \{A \cap F : A \in \mathcal{F}\}$, and $P'(E) = \dfrac{P(E \cap F)}{P(F)}$.

We write the updated probability as $P'(E) = P(E\,|F)$, which we call the *conditional probability of E given F*.

It is worth observing that for any event F with $P(F) > 0$, $P(\cdot\,|F)$ is actually a probability on the original sample space and collection of events. Indeed, for any event E, $0 \le P(E \cap F) \le P(F)$, so $0 \le \frac{P(E \cap F)}{P(F)} \le 1$. Also, $P(\Omega\,|F) = \frac{P(\Omega \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1$. Finally, if $E_1, E_2, \ldots$ are disjoint, then so are $E_1 \cap F, E_2 \cap F, \ldots$, hence

$$P\left(\bigcup_{i=1}^{\infty} E_i\,|F\right) = \frac{P\left(\left(\bigcup_{i=1}^{\infty} E_i\right) \cap F\right)}{P(F)} = \frac{1}{P(F)} P\left(\bigcup_{i=1}^{\infty} (E_i \cap F)\right)$$
$$= \frac{1}{P(F)} \sum_{i=1}^{\infty} P(E_i \cap F) = \sum_{i=1}^{\infty} \frac{P(E_i \cap F)}{P(F)} = \sum_{i=1}^{\infty} P(E_i\,|F).$$

In light of this observation, we need not (and typically don't) update the sample space or σ-field when confronted with new information. It is completely encoded in the conditional probability measure.

Example 7.1. Suppose that a missing key is in my right pocket with probability $\frac{2}{5}$ and is in my left pocket with probability $\frac{1}{5}$. If I check my right pocket and do not find the key, what is the probability that it is in my left pocket?

Letting L denote the event that the key is in my left pocket and R that the key is in my right pocket, we compute

$$P\left(L \,\middle|\, R^C\right) = \frac{P\left(L \cap R^C\right)}{P(R^C)} = \frac{P(L)}{1 - P(R)} = \frac{1}{3}.$$

Example 7.2. Recall the introductory example about the town of two sibling households. If you meet a parent at random who you know has a daughter, what is the probability that they have two girls? What if you know that their oldest child is a girl?

Let $E = \{(G, G)\}$ be the event that the person has two girls, $F = \{(G, G), (G, B), (B, G)\}$ be the event that they have at least one girl, and $H = \{(G, G), (G, B)\}$ be the event that their oldest child is a girl. Then the probability that they have two girls given that they have at least one girl is

$$P(E \,|\, F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)}{P(F)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3},$$

and the probability that they have two girls given that their oldest child is a girl is

$$P(E \,|\, H) = \frac{P(E \cap H)}{P(H)} = \frac{P(E)}{P(H)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

The case where you know that they have a girl born on Tuesday can be worked out similarly.

When we initially considered this sibling paradox, we did not need to resort to conditional probabilities since we could just enumerate outcomes in the "reduced sample space." In many cases, it is easier to take this more direct route. An advantage to having the two approaches is that one can sometimes simplify probability computations by introducing conditional probabilities as an intermediary step. To see how this works, note that the definition of conditional probability gives the "multiplication rule"

$$P(E \cap F) = P(F)P(E \,|\, F).$$

The interpretation is that for E and F to occur, F must occur and, given that F occurs, E must occur. The following examples are intended to clarify this point.

Example 7.3. Suppose that you are dealt two cards from a thoroughly shuffled deck of 52 cards. What is the probability of receiving two aces?

We could do this without conditional probability by noting that of the $\binom{52}{2} = \frac{52 \cdot 51}{2}$ possible two-card hands, there are $\binom{4}{2} = \frac{12}{2}$ consisting solely of aces, thus $P(2 \text{ aces}) = \frac{12}{52 \cdot 51}$.

Alternatively, in order to receive two aces, the first card must be an ace and, conditional on the first card being an ace, the second must as well. That is

$$P(2 \text{ aces}) = P(1\text{st card ace})P(2\text{nd card ace} \,|\, 1\text{st card ace}) = \frac{4}{52} \cdot \frac{3}{51} = \frac{12}{52 \cdot 51}$$

where the second probability was computed by noting that in the reduced sample space there are 3 remaining aces and 51 remaining cards.

Example 7.4. Recall the hat check problem where n people drop off their hats at a reception desk and random hats are returned. We saw that the probability that no one receives their original hat is $p_n = \sum_{j=0}^{n} \frac{(-1)^j}{j!}$. What is the probability that exactly k people receive their original hat?

For any specific group of k people, let E be the event that each of them received their hats, and let F be the event that the other $n-k$ all receive someone else's hat. The probability that only the k specified people receive their hats is $P(E \cap F) = P(E)P(F\,|E)$.

Out of the $n!$ ways of distributing the n hats $(n-k)!$ result in the k "special" people receiving their own (since there are $(n-k)!$ ways of distributing the remaining hats to the remaining people), hence $P(E) = \frac{(n-k)!}{n!}$.

Given that E occurs, the probability of F is the same as the probability that in a group of $n-k$ people, none receive their hat, so $P(F\,|E) = p_{n-k}$. Since this is true for any of the $\binom{n}{k}$ specific groups of k people, we see that

$$P(\text{exactly } k \text{ receive original hat}) = \binom{n}{k} P(E \cap F) = \frac{n!}{k!(n-k)!} \cdot \frac{(n-k)!}{n!} p_{n-k} = \frac{1}{k!} \sum_{j=0}^{n-k} \frac{(-1)^j}{j!}.$$

Thus when n is large relative to k, the probability that exactly k people receive their hat is approximately $\frac{e^{-1}}{k!}$.

8. Bayes' Formula

Example 8.1. Suppose k balls are drawn from an urn containing m red balls and n blue balls. Given that j of the balls drawn were red, what is the probability that the first ball drawn was red? We are assuming that $j \leq k \leq m$.

Let R_1 be the event that the sample consists of j red balls, and let R_2 be the event that the first ball sampled is red. We are interested in computing $P(R_2\,|R_1\,) = \frac{P(R_1 \cap R_2)}{P(R_1)}$.
By previous arguments,

$$P(R_1) = \frac{\binom{m}{j}\binom{n}{k-j}}{\binom{m+n}{k}}.$$

To evaluate the numerator, note that the "multiplication rule" for intersections gives $P(R_1 \cap R_2) = P(R_2)P(R_1\,|R_2\,)$. Clearly $P(R_2) = \frac{m}{m+n}$. Given that the first ball was red, the probability that j reds are drawn in total is the same as the probability that a sample of size $k-1$ from an urn containing n blue balls and $m-1$ red balls contains $j-1$ red balls. It follows that

$$P(R_1\,|R_2\,) = \frac{\binom{m-1}{j-1}\binom{n}{k-j}}{\binom{m+n-1}{k-1}}.$$

Therefore,

$$P(R_2\,|R_1\,) = \frac{P(R_1 \cap R_2)}{P(R_1)} = \frac{P(R_2)P(R_1\,|R_2\,)}{P(R_1)} = \frac{\frac{m}{m+n}\frac{\binom{m-1}{j-1}\binom{n}{k-j}}{\binom{m+n-1}{k-1}}}{\frac{\binom{m}{j}\binom{n}{k-j}}{\binom{m+n}{k}}}$$

$$= \frac{m}{m+n} \cdot \frac{\binom{m-1}{j-1}}{\binom{m}{j}} \cdot \frac{\binom{m+n}{k}}{\binom{m+n-1}{k-1}} = \frac{m}{m+n} \cdot \frac{j}{m} \cdot \frac{m+n}{k} = \frac{j}{k}.$$

Note that this does not depend on the proportion of red and blue balls!

The real beauty of this example is the trick of expressing $P(R_1\,|R_2\,)$ in terms of $P(R_2\,|R_1\,)$. This inversion of conditioning lies at the heart of the celebrated *Bayes formula*.

Proposition 8.1. *Let E and F be events having positive probability. Then*

$$P(E\,|F\,) = \frac{P(E)P(F\,|E\,)}{P(F)}.$$

The proof is just the definition of conditional probability and the multiplication rule for intersections.

The interpretation is that $\frac{P(F|E\,)}{P(F)}$ is a measure of the evidence that F provides for or against E. To obtain the *posterior probability*, $P(E\,|F\,)$, we multiply the *prior probability*, $P(E)$, by this factor.

Observe that $\frac{P(F|E\,)}{P(F)} = \frac{P(E \cap F)}{P(E)P(F)} = \frac{P(E|F\,)}{P(E)}$, so the factor by which the knowledge that F occurs changes the prior probability of E is the same as that by which the knowledge that E occurs changes the prior probability of F. Thus the "evidentiary" term is really telling you something about how much the two events are entangled.

Often, one writes $F = (E \cap F) \sqcup (E^C \cap F)$ and uses the multiplication rule to express

$$P(F) = P(E \cap F) + P(E^C \cap F) = P(F\,|E)P(E) + P(F\,\big|E^C)P(E^C),$$

so that Bayes' formula reads

$$P(E\,|F) = \frac{P(F\,|E)P(E)}{P(F\,|E)P(E) + P(F\,|E^C)P(E^C)}.$$

Here we are thinking of E and E^C as competing hypotheses and F as evidence.

The posterior probability of E given F exceeds the prior probability $P(E)$ precisely when

$$\frac{P(F\,|E)}{P(F\,|E)P(E) + P(F\,|E^C)P(E^C)} > 1,$$

which is equivalent to

$$P(F\,\big|E^C)P(E^C) < P(F\,|E) - P(F\,|E)P(E) = P(F\,|E)P(E^C),$$

or

$$P(F\,|E) > P(F\,\big|E^C).$$

Finally, note that the same basic reasoning applies with more than two hypotheses:

Theorem 8.1. *If $E_1, ..., E_n$ is a partition of the sample space, then*

$$P(E_j\,|F) = \frac{P(F\,|E_j)P(E_j)}{\sum_{i=1}^{n} P(F\,|E_i)P(E_i)}.$$

One of the most common examples of the use of Bayes' formula concerns false positives in medical testing.

Example 8.2. Assume that 1% of women between the ages of 40 and 50 have breast cancer. Assume also that a woman with breast cancer has a 90% chance of a positive test from a mammogram, while a woman without has a 10% chance of testing positive. What is the probability that a woman aged $40-50$ has breast cancer given that she tested positive?

Let A be the event "tests positive" and let B be the event "has breast cancer." Bayes' formula gives

$$P(B\,|A) = \frac{P(A\,|B)P(B)}{P(A\,|B)P(B) + P(A\,|B^C)P(B^C)} = \frac{\frac{9}{10}\cdot\frac{1}{100}}{\frac{9}{10}\cdot\frac{1}{100} + \frac{1}{10}\cdot\frac{99}{100}} = \frac{9}{108} \approx 0.0833.$$

In a study in which 95 physicians were asked this question, their average response was 0.75.

This study also showed that the physicians were better able to understand the answer when it was explained in terms of frequency: Out of $1,000$ women, we expect about 10 to have breast cancer, and 9 of these 10 to test positive. But of the 990 without breast cancer, we expect about 99 to test positive. Thus of the $9 + 99 = 108$ positives, 9 are true positives.

Example 8.3. A jar contains m coins of type 1 and n coins of type 2, where coins of type 1 are k times as likely to land on heads as are those of type 2. Suppose that we select a random coin from the jar, toss it, and observe that it lands on heads. What is the probability that we chose a coin of type 1?

Let A be the event "coin lands on heads" and let B be the event "coin of type 1." Write p for the heads probability of type 2 coins. Then

$$P(B\,|A) = \frac{P(A\,|B)P(B)}{P(A\,|B)P(B) + P(A\,|B^C)P(B^C)} = \frac{kp \cdot \frac{m}{m+n}}{kp \cdot \frac{m}{m+n} + p \cdot \frac{n}{m+n}} = \frac{km}{km+n}.$$

Our final example demonstrates the relevance of Bayes' formula to legal proceedings.

Example 8.4. The overall murder rate is estimated to be 1 in $10,000$. In a murder trial in which the accused has been previously convicted of abusing the victim, the defense argues that this information should not be admitted since it is estimated that the percentage of batterers who go on to murder their spouse is 0.1%. Is this a reasonable request? Can we conclude that the probability that the defendant is guilty of murder is 0.001?

Restricting the sample space to include only individuals with abusive spouses, let H be the event that a person is murdered by their abusive spouse, and let M be the event that a person who has been abused by their spouse is murdered. Assuming that the overall murder rate is an accurate estimate for the probability that an abused person is murdered given that they were not murdered by their abusive spouse, Bayes's formula and the given statistics yield

$$\begin{aligned} P(H\,|M) &= \frac{P(M\,|H)P(H)}{P(M\,|H)P(H) + P(M\,|H^C)P(H^C)} \\ &= \frac{1 \cdot \frac{1}{1{,}000}}{1 \cdot \frac{1}{1{,}000} + \frac{1}{10{,}000} \cdot \frac{999}{1{,}000}} = \frac{1}{1 + \frac{999}{10{,}000}} \approx 0.9091. \end{aligned}$$

The history of abuse is certainly relevant information. The probability of an abuser murdering their spouse is drastically higher once we know that the spouse has been murdered.

9. Independent Events

We have previously been considering cases in which $P(E\,|F)$ is different from $P(E)$. When the two agree, the interpretation is that knowledge of whether or not F occurs has no impact on your assessment of the probability that E occurs. In this case, we say that E and F are *independent*.

Since $P(E\,|F) = \frac{P(E\cap F)}{P(F)}$, we see that if $P(E\,|F) = P(E)$, then

$$P(E \cap F) = P(E)P(F).$$

We will take the above equation as our definition of independence.

Some advantages of this formulation over one involving conditional probabilities include the fact that it does not require the events to have positive probability and it is obviously symmetric in E and F. It is also easier to generalize to more than two events and is often much better suited for proving theorems.

For example, it's straightforward that if E and F are independent, then so are E^C and F, E and F^C, and E^C and F^C.

Indeed, if $P(E \cap F) = P(E)P(F)$, then

$$P(E^C \cap F) = P(F) - P(E \cap F) = P(F) - P(E)P(F) = P(F)\,(1 - P(E)) = P(F)P(E^C).$$

(The first equality follows from the fact that $P(F) = P(E \cap F) + P(E^C \cap F)$.)

Interchanging E and F in the above shows that E and F^C are independent as well, and applying the first assertion to E and F^C establishes the third.

Another useful observation is that any event with probability zero or one is independent of all other events:

If $P(E) = 0$, then for any event F, $0 \le P(E \cap F) \le P(E) = 0$, so $P(E \cap F) = 0 = P(E)P(F)$. Since the complement of an event with probability one has probability zero, the previous remark shows that events with probability one are independent of all else.

Example 9.1. In the two fair coins experiment, suppose that A is the event that the first coin lands on heads and B is the event that the second coin lands on tails. Are A and B independent?

If C is the event that the two coins land on different sides, are A and C independent? What about B and C? Is C independent of $A \cap B$?

$A = \{(H,H),(H,T)\}$ and $B = \{(H,T),(T,T)\}$ are independent because $P(A \cap B) = P(\{H,T)\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B)$.

Similarly, since $C = \{(H,T),(T,H)\}$, we have $A \cap C = \{(H,T)\}$, thus $P(A \cap C) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(C)$, so A and C are independent. A nearly identical argument shows that B and C are independent.

However, $P\left((A \cap B) \cap C\right) = P(A \cap B) \neq P(A \cap B)P(C)$, so C is not independent of $A \cap B$.

In the previous example, it would not make sense to say that A, B, and C are independent because if we know that A and B have occurred, then we know that C has occurred. Thus a collection of three events can be dependent even if any two of the events are independent.

Definition. Events E, F, and G are independent if

$$\begin{aligned} P(E \cap F) &= P(E)P(F) \\ P(E \cap G) &= P(E)P(G) \\ P(F \cap G) &= P(F)P(G) \\ P(E \cap F \cap G) &= P(E)P(F)P(G). \end{aligned}$$

Note that if E, F, and G are independent, then E is independent of any event formed from F and G. For example,

$$\begin{aligned} P\left(E \cap (F \cup G)\right) &= P\left((E \cap F) \cup (E \cap G)\right) = P(E \cap F) + P(E \cap G) - P(E \cap F \cap G) \\ &= P(E)P(F) + P(E)P(G) - P(E)P(F)P(G) \\ &= P(E)\left(P(F) + P(G) - P(F)P(G)\right) \\ &= P(E)\left(P(F) + P(G) - P(F \cap G)\right) = P(E)P(F \cup G). \end{aligned}$$

We can extend this idea to collections of more than three events as follows.

Definition. A collection of events $\{E_i\}_{i \in I}$ is independent if for any finite subset $S \subseteq I$, we have

$$P\left(\bigcap_{i \in S} E_i\right) = \prod_{i \in S} P(E_i).$$

In many cases we are interested in experiments consisting of a finite or infinite sequence of sub-experiments which can be regarded as independent.

Example 9.2. Suppose I flip a coin n times. If the coin has probability p of landing on heads, what is the probability that it lands on heads at least once?
What is the probability that it lands on heads exactly k times?

We may regard the successive flips as independent so that any particular sequence of m heads and $n-m$ tails has probability $p^m(1-p)^{n-m}$. The event that there is at least one heads is the complement of the event that every toss landed on tails. Thus $P(\text{at least one } H) = 1 - (1-p)^n$.
By assumption, every sequence containing exactly k heads has probability $p^k(1-p)^{n-k}$. Since there are $\binom{n}{k}$ such sequences (determined by specifying when the heads occur), we have
$P(k \text{ heads}) = \binom{n}{k} p^k (1-p)^{n-k}$.

Example 9.3. An experiment consists of repeatedly rolling a pair of dice and recording their sum. What is the probability that a sum of 5 appears before a sum of 7?

If E_n is the event that neither a 5 nor a 7 is rolled in the first $n-1$ trials and a 5 is rolled on the nth trial, then the event that a 5 is rolled before a 7 is $\bigcup_{n=1}^{\infty} E_n$. On each particular trial, the probability of a 5 is $\frac{4}{36}$ and the probability of a 7 is $\frac{6}{36}$, so the probability that neither a 5 nor a 7 is rolled is $P(\{5,7\}^C) = 1 - \frac{10}{36} = \frac{13}{18}$.

Since we may assume that the rolls are independent, we have $P(E_n) = \frac{4}{36}\left(\frac{13}{18}\right)^{n-1}$.
Accordingly,

$$P(5 \text{ before } 7) = P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \frac{4}{36}\left(\frac{13}{18}\right)^{n-1} = \frac{1}{9}\sum_{m=0}^{\infty}\left(\frac{13}{18}\right)^{m} = \frac{1}{9}\cdot\frac{1}{1-\frac{13}{18}} = \frac{2}{5}.$$

The exact same argument shows that if E and F are disjoint events in an experiment, then when independent trials of the experiment are repeatedly performed, the probability that E occurs before F is

$$P(E \text{ before } F) = \sum_{n=1}^{\infty} P(E)P\left((E\cup F)^C\right)^{n-1} = P(E)\sum_{m=0}^{\infty}\left(1-P(E)-P(F)\right)^m = \frac{P(E)}{P(E)+P(F)}.$$

Our next example, often referred to as the problem of the points, is of great significance in the history of probability. The question was posed to Blaise Pascal in the mid seventeenth century by a professional gambler. This inspired Pascal to begin thinking about mathematical analyses of games of chance and to initiate a correspondence with Pierre de Fermat on the subject. Many consider the ensuing exchange to mark the birth of probability.

Example 9.4. Suppose that two players are engaged in a game consisting of multiple rounds. In each round, a single point is awarded to one of the players, with player 1 getting the point with probability p. At the beginning of the game, the players put up stakes which are to be awarded to the first to receive N of points. If the game is interrupted when player 1 has i points and player 2 has j points, $i, j < N$, how should the stakes be divided?

The idea is to award the players values proportional to their chances of winning if the game was allowed to continue. Thus we need to consider how many more points each player needs to win.

We can think of the problem in terms of independent trials where each trial has success probability p, and ask what is the probability that $n = N - i$ successes occur before $m = N - j$ failures.

The solution Pascal proposed is to let $P_{n,m}$ denote this probability. By conditioning on the outcome of the first trial, we see that

$$P_{n,m} = pP_{n-1,m} + (1-p)P_{n,m-1}$$

for $m, n \in \mathbb{N}$. Using the boundary conditions $P_{n,0} = 0$ and $P_{0,m} = 1$, this recurrence can be solved for any fixed m and n.

A more elegant solution was proposed by Fermat, who noted that this was equivalent to asking for the chance of at least n successes in $m+n-1$ trials. (Even though the game may have ended before that many rounds were played, there is no harm in assuming that additional trials were performed.)

In this case, there are at most $m-1$ failures, so n successes occurred before m failures. If less than n successes occurred, then there must have been at least $m = (m+n-1)-(n-1)$ failures, so m failures occurred before n successes.

Since the probability of exactly k successes in $m+n-1$ trials is $\binom{m+n-1}{k}p^k(1-p)^{m+n-1-k}$, we have

$$P(\text{Player 1 wins}) = P(\text{at least } n \text{ successes in } m+n-1 \text{ trials}) = \sum_{k=n}^{m+n-1}\binom{m+n-1}{k}p^k(1-p)^{m+n-1-k}.$$

Example 9.5. Consider a "serve and rally" match (such as volleyball or tennis) between players A and B. The game consists of a series of rallies in which one player is the server. Players receive a point for each rally won and the first player to earn n points wins. Suppose that A has probability p_A of winning the rally when they serve, and winning probability p_B when B serves. Assuming that A always serves the first game of the rally, under what conditions on p_A and p_B is the better protocol for player A to alternate serves or to let the winner of a rally serve for the next round?

As in the previous example, we can suppose that $2n-1$ games are played even if the match would have ended in fewer games. Player A wins the game if they win at least n of the $2n-1$ rallies.

If the players alternated serving, then A would have n serves in total and player B would have $n-1$.
By casing out according to the number of games A wins as the server and the receiver, we see that

$$P(A \text{ wins}) = \sum_{k=1}^{n} \binom{n}{k} p_A^k (1-p_A)^{n-k} \sum_{m=n-k}^{n-1} \binom{n-1}{m} p_B^m (1-p_B)^{n-1-m}.$$

In particular, for any fixed p_A and p_B, the probability only depends on the number of times A serves.

On the other hand, in the winner serves protocol, we may always assume that the loser serves every round after the match is over. After all, those are just friendly games that have no effect on the outcome of the match.
Thus in the event that A wins, they must have served exactly n of the rounds: the first round and each round following their i^{th} point for $i = 1, ..., n-1$.
If A loses, then they also serve n rounds: all rounds other than those following B's i^{th} point for $i = 1, ..., n-1$.

It follows that for any p_A, p_B, A has the same probability of winning with either protocol!

Example 9.6. Suppose there are initially r players with player i having $n_i \in \mathbb{N}$ dollars at the outset. At each stage, two of the players are chosen to play a game, with the loser paying one dollar to the winner. Players are eliminated once their fortune drops to zero, and the games are assumed to be independent with each contestant equally likely to win. What is the probability that player i is the victor?

Write $n = \sum_{i=1}^{r} n_i$, and consider the situation in which there are n players with \$1 each. In this case, symmetry dictates that all players are equally likely to triumph, so the probability that any given player is victorious is $\frac{1}{n}$.

Now suppose that these n players are divided into r teams with team i consisting of n_i players. Then the probability that the winner belongs to team i is $\frac{n_i}{n}$. Since the event that a member of team i wins is precisely the event that, as a whole, team i amasses n dollars (since then the members of the other teams have all been eliminated); team i begins with n_i dollars; and the overall fortune of team i only changes when a player from i is pitted against a member of another team (in which case the team's fortune is equally likely to increase or decrease by \$1), the probability that a member of team i wins is the same as the probability that person i wins in the original setup. That is, $P(i \text{ wins}) = \frac{n_i}{n}$.

The *probabilistic method* is a technique pioneered by Paul Erdős for establishing the existence of some mathematical object with prescribed properties. The idea is to choose an element at random from some collection of interest in such a way that each element has positive probability. If the probability that an object chosen thusly has the desired property is zero, then no such object exists. Equivalently, if the probability that the object chosen does not have the property is less than one, then there must exist an object with the desired property. The following example is a nice illustration of this method.

Example 9.7. The complete graph on n vertices consists of n points with each of the $\binom{n}{2}$ pairs connected by an edge. Suppose that each edge can be colored either red or blue. For a fixed integer k, is there such a coloring for which no set of k vertices has all $\binom{k}{2}$ connecting edges the same color?

If the vertices are people, a red edge indicates acquaintance, and a blue edge non-acquaintance, then we are asking if it's possible for every group of k people to contain both friends and strangers.

We will use the probabilistic method to show that if n satisfies $\binom{n}{k} < 2^{\frac{k(k-1)}{2}-1}$, then such an edge coloring exists.

To see this is so, suppose that each edge is independently colored either red or blue with equal probability. Enumerate the $\binom{n}{k}$ distinct sets of k vertices and let E_i be the event that all connecting edges in the i^{th} set of k vertices have the same color, $i = 1, ..., \binom{n}{k}$. Since each of the $\binom{k}{2}$ edges in the i^{th} set is equally likely to be red or blue, independently of the other edge colors, we have

$$P(E_i) = 2\left(\frac{1}{2}\right)^{\binom{k}{2}} = \left(\frac{1}{2}\right)^{\frac{k(k-1)}{2}-1}.$$

It follows from Boole's inequality that the probability that some k-set of vertices has monochromatic connecting edges is

$$P\left(\bigcup_{1\le i\le\binom{n}{k}} E_i\right) \le \sum_{1\le i\le\binom{n}{k}} P(E_i) = \binom{n}{k}\left(\frac{1}{2}\right)^{\frac{k(k-1)}{2}-1}.$$

Thus when $\binom{n}{k} < 2^{\frac{k(k-1)}{2}-1}$, the probability that at least one set of k vertices has all edges the same color is less that 1, so there must exist some coloring in which every k-set of vertices has both red and blue connecting edges.

Conditional Independence.

We conclude our discussion of conditional probability and independence by recalling that the conditional probability $P(\cdot\,|F)$ defines a probability on the events in our original sample space when $P(F) > 0$. Thus, writing $Q(E) = P(E\,|F)$, we see that all of our previous results also apply to Q.

For example,

$$P(E\cup G\,|F) = Q(E\cup G) = Q(E) + Q(G) - Q(E\cap G) = P(E\,|F) + P(G\,|F) - P(E\cap G\,|F).$$

Likewise, we can define the conditional probability $Q(E\,|G) = \frac{Q(E\cap G)}{Q(G)}$, provided that $P(G\,|F) = Q(G) > 0$.

In this case,

$$Q(E\,|G\,) = \frac{Q(E\cap G)}{Q(G)} = \frac{P(E\cap G\,|F\,)}{P(G\,|F)} = \frac{\frac{P(E\cap G\cap F)}{P(F)}}{\frac{P(G\cap F)}{P(F)}} = \frac{P\,(E\cap (G\cap F))}{P(G\cap F)} = P(E\,|G\cap F\,).$$

Thus if $G_1, ..., G_n$ is a partition of the sample space with $P(G_i\,|F\,) > 0$ for all i, then

$$P(E\,|F\,) = Q(E) = \sum_{i=1}^{n} Q(E\cap G_i) = \sum_{i=1}^{n} Q(E\,|G_i)Q(G_i) = \sum_{i=1}^{n} P\,(E\,|G_i\cap F\,)\,P\,(G_i\,|F\,)\,.$$

This perspective also allows us to define the notion of *conditional independence*. Namely, E and G are conditionally independent given F if

$$P(E\cap G\,|F\,) = Q(E\cap G) = Q(E)Q(G) = P(E\,|F\,)P(G\,|F\,).$$

Equivalently, if $P(G\cap F) > 0$, then E and G are conditionally independent given F if

$$P(E\,|G\cap F\,) = \frac{P(E\cap G\cap F)}{P(G\cap F)} = \frac{P(E\cap G\,|F\,)P(F)}{P(G\,|F)P(F)} = \frac{P(E\,|F)P(G\,|F\,)}{P(G\,|F\,)} = P(E\,|F).$$

Example 9.8. Suppose there are $k+1$ coins in a jar with the i^{th} coin having heads probability $\frac{i}{k}$, $i = 0, 1, ..., k$. A coin is drawn uniformly at random from the jar and repeatedly flipped. If the first n tosses result in heads, what is the probability that the outcome of the next toss is heads?

Letting C_i be the event that the i^{th} coin is selected, F_n that the first n flips resulted in heads, and H that the next outcome is heads, we have

$$P(H\,|F_n\,) = \sum_{i=0}^{k} P(H\,|C_i\cap F_n\,)P(C_i\,|F_n\,).$$

Given that the i^{th} coin is selected, we may assume that the successive trials are conditionally independent, so

$$P(H\,|C_i\cap F_n\,) = P(H\,|C_i\,) = \frac{i}{k}.$$

Also, Bayes' formula gives

$$P(C_i\,|F_n\,) = \frac{P(F_n\,|C_i)P(C_i)}{\sum_{j=0}^{k} P(F_n\,|C_j\,)P(C_j)} = \frac{\left(\frac{i}{k}\right)^n \frac{1}{k+1}}{\sum_{j=0}^{k}\left(\frac{j}{k}\right)^n \frac{1}{k+1}}.$$

It follows that

$$P(H\,|F_n\,) = \sum_{i=0}^{k}\left(\frac{i}{k}\right)\frac{\left(\frac{i}{k}\right)^n}{\sum_{j=0}^{k}\left(\frac{j}{k}\right)^n} = \frac{\sum_{i=0}^{k}\left(\frac{i}{k}\right)^{n+1}}{\sum_{j=0}^{k}\left(\frac{j}{k}\right)^n}.$$

When k is large, we can use the integral approximations

$$\frac{1}{k}\sum_{i=0}^{k}\left(\frac{i}{k}\right)^{n+1} \approx \int_0^1 x^{n+1}dx = \frac{1}{n+2},$$

$$\frac{1}{k}\sum_{j=0}^{k}\left(\frac{j}{k}\right)^{n} \approx \int_0^1 x^{n}dx = \frac{1}{n+1},$$

to obtain $P(H\,|F_n\,) \approx \dfrac{n+1}{n+2}$.

10. Random Variables

At this point, we have gone about as far as possible without introducing random variables, a concept that will be central to all of our discussions from here on out.

Definition. A *random variable* X on a probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$.

The idea is that we often do not have direct access to the actual outcomes of a random experiment, but only have information concerning some particular measurement.

Also, in many cases we are only interested in some specific facet of the outcomes and do not want to burden ourselves with extraneous details.

Random variables represent the values of the measurement or the information of interest in these scenarios.

For example, in the two fair dice experiment, we may only be interested in the sum of the values rather than the particular pair of values. In this case, the sample space is $\Omega = \{(i,j)\}_{i,j=1}^{6}$ and the sum of the values is the random variable $X((i,j)) = i + j$.

Similarly, I might toss a coin n times, but only tell you the outcome of the third toss. The sample space can be represented by the set of bit strings of length n, and the outcome of the third toss is the random variable giving the projection onto the third coordinate - e.g. $Y(11010001) = 0$.

We can think of random variables as specifying events: For $B \subseteq \mathbb{R}$, $\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}$.

From this perspective, X induces a probability on $\mathbb{R}$ (called the distribution of X) defined by
$\mu_X(B) = P(X \in B)$ for $B \subseteq \mathbb{R}$.

Note that $0 \leq \mu_X(B) \leq 1$ for all B since P takes values in $[0,1]$.

Also, $\mu_X(\mathbb{R}) = P(X \in \mathbb{R}) = P(\Omega) = 1$ since every outcome in ω maps to some real number under X.

Finally, since X is a function, if $B_1, B_2, ...$ are disjoint subsets of $\mathbb{R}$, then $\{X \in B_1\}, \{X \in B_2\}, ...$ are also disjoint, hence

$$\mu_X\left(\bigcup_{i=1}^{\infty} B_i\right) = P\left(X \in \bigcup_{i=1}^{\infty} B_i\right) = P\left(\bigcup_{i=1}^{\infty} \{X \in B_i\}\right) = \sum_{i=1}^{\infty} P(X \in B_i) = \sum_{i=1}^{\infty} \mu_X(B_i).$$

* As usual, the foregoing ignores some measurability issues, but this can be easily remedied.

Example 10.1. An urn contains balls labeled $1, 2, ..., 20$. Four balls are randomly selected without replacement and their highest value, X, is recorded. The possible values that X can take are $4, 5, ..., 20$. For $i = 4, ..., 20$, what is $P(X = i)$? What about $P(X > 10)$?

Since each of the $\binom{20}{4}$ possible selections of 4 balls is equally likely and $X = i$ whenever ball i is selected along with 3 other balls having labels less than i, we see that for $i = 4, ..., 20$

$$P(X = i) = \frac{\binom{i-1}{3}}{\binom{20}{4}}.$$

(Of course, $P(X = x) = 0$ if $x \notin \{4, ..., 20\}$.)

We can compute the probability of the event $\{X > 10\}$ by summing over the associated values:

$$P(X > 10) = \sum_{i=11}^{20} P(X = i) = \binom{20}{4}^{-1} \sum_{j=10}^{19} \binom{j}{3}.$$

A more direct approach is to note that $X \leq 10$ if and only if all balls chosen have labels $1, ..., 10$.
Thus

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \frac{\binom{10}{4}}{\binom{20}{4}}.$$

(I guess you could also compute the probability that at least one of the balls labeled $11, ..., 20$ is chosen using inclusion-exclusion, but that's a bit messy for my taste.)

Example 10.2. A coin with heads probability p is flipped until either a heads appears or a total of n flips is made. If X denotes the number of flips, then X is a random variable with

$$\begin{aligned}
P(X = 1) &= P(\{H\}) = p, \\
P(X = 2) &= P(\{T, H\}) = (1-p)p, \\
&\vdots \\
P(X = n-1) &= P(\{\underbrace{T, ..., T}_{n-2}, H\}) = (1-p)^{n-2}p \\
P(X = n) &= P(\{\underbrace{T, ..., T}_{n-1}, H\}) + P(\{\underbrace{T, ..., T}_{n}\}) = P(\{\underbrace{T, ..., T}_{n-1}\}) = (1-p)^{n-1}.
\end{aligned}$$

To check that this makes sense, note that

$$\begin{aligned}
\sum_{i=1}^{n} P(X = i) &= \sum_{i=1}^{n-1} p(1-p)^{i-1} + (1-p)^{n-1} \\
&= p \sum_{j=0}^{n-2} (1-p)^j + (1-p)^{n-1} \\
&= p \left[\frac{1 - (1-p)^{n-1}}{1 - (1-p)} \right] + (1-p)^{n-1} = 1.
\end{aligned}$$

Example 10.3. Suppose there are n distinct types of coupons and that each time a coupon is collected it is equally likely to be any of the n types, independently of previous selections. Let T be the random variable which records the number of coupons needed to obtain a complete set. Then T takes values in $n, n+1, ...$
For $k \geq n$, what is $P(T = k)$?

We will begin by computing $P(T > k)$ for arbitrary $k \in \mathbb{N}$.
To this end, define the events $A_1, ..., A_n$ by $A_i = \{\text{coupon } i \text{ not selected in first } k \text{ rounds}\}$.
Then $\{T > k\} = \bigcup_{i=1}^{n} A_i$, so the principle of inclusion-exclusion gives

$$P(T > k) = P\left(\bigcup_{i=1}^{n} A_i \right) = \sum_{m=1}^{n} (-1)^{m+1} \sum_{i_1 < ... < i_m} P\left(\bigcap_{j=1}^{m} A_{i_j} \right).$$

For any particular set of m coupon types $\{i_1, ..., i_m\}$, the event that none have been collected in the first k stages is the same as the event that the first k coupons were all of the $n - m$ other types, hence
$P\left(\bigcap_{j=1}^{m} A_{i_j} \right) = \left(\frac{n-m}{n} \right)^k$.

Accordingly, we have

$$P(T > k) = \sum_{m=1}^{n} (-1)^{m+1} \sum_{i_1 < \ldots < i_m} P\left(\bigcap_{j=1}^{m} A_{i_j} \right) = \sum_{m=1}^{n-1} (-1)^{m+1} \binom{n}{m} \left(\frac{n-m}{n} \right)^k .$$

Finally, we can compute the desired probability as

$$\begin{aligned} P(T = k) &= P(T > k-1) - P(T > k) \\ &= \sum_{m=1}^{n-1} (-1)^{m+1} \binom{n}{m} \left(\frac{n-m}{n} \right)^{k-1} - \sum_{m=1}^{n-1} (-1)^{m+1} \binom{n}{m} \left(\frac{n-m}{n} \right)^k \\ &= \sum_{m=1}^{n-1} (-1)^{m+1} \binom{n}{m} \left(\frac{n-m}{n} \right)^{k-1} \left(1 - \frac{n-m}{n} \right) \\ &= \sum_{m=1}^{n-1} (-1)^{m+1} \binom{n}{m} \left(\frac{n-m}{n} \right)^{k-1} \left(\frac{m}{n} \right). \end{aligned}$$

* It is interesting to note that since $P(T > k) = 1$ for $1 \leq k < n$, the preceding shows that for such k we have

$$\sum_{m=1}^{n-1} (-1)^{m+1} \binom{n}{m} \left(\frac{n-m}{n} \right)^k = 1,$$

which is equivalent to

$$\sum_{m=0}^{n-1} (-1)^{m} \binom{n}{m} \left(\frac{n-m}{n} \right)^k = 0.$$

Multiplying by $(-1)^n n^k$ and setting $l = n - m$ gives the identity

$$\sum_{l=1}^{n} \binom{n}{l} (-1)^l l^k = 0, \ 1 \leq k < n.$$

Another random variable that we may care about in this problem is the number of distinct types of coupons collected by time $t \in \mathbb{N}$, which we denote N_t.

To compute the probability that $N_t = m$ for $m = 1, ..., n$, we first consider a particular set of m coupon types $\{i_1, ..., i_m\}$.

In order for the types collected by time t to be precisely this set, both of the following events must occur:

$$A = \{\text{each of the } t \text{ coupons are of these } m \text{ types}\},$$
$$B = \{\text{at least one coupon is of type } i_j \text{ for } j = 1, ..., m\}.$$

Reasoning as before, $P(A) = \left(\frac{m}{n}\right)^t$. Given that all coupons collected are of types $i_1, ..., i_m$, the probability that each type is represented is the same as the probability that a complete set of m types has been collected by time t, which is the complement of the event that it takes more than t trials to obtain one of each of the m types.

Thus, letting T_m be the number of trials to complete the set of m types, we see that

$$\begin{aligned} P(B\,|A) &= 1 - P(T_m > t) = 1 - \sum_{r=1}^{m-1} (-1)^{r+1} \binom{m}{r} \left(\frac{m-r}{m}\right)^t \\ &= \sum_{r=0}^{m-1} (-1)^r \binom{m}{r} \left(\frac{m-r}{m}\right)^t \end{aligned}$$

Because there are $\binom{n}{m}$ possible choices for the set of m types, we have

$$\begin{aligned} P(N_t = m) &= \binom{n}{m} P(A \cap B) = \binom{n}{m} P(A) P(B\,|A\,) \\ &= \binom{n}{m} \left(\frac{m}{n}\right)^t \left[\sum_{r=0}^{m-1} (-1)^r \binom{m}{r} \left(\frac{m-r}{m}\right)^t\right] \\ &= \binom{n}{m} \sum_{r=0}^{m-1} (-1)^r \binom{m}{r} \left(\frac{m-r}{n}\right)^t. \end{aligned}$$

11. Discrete Random Variables and Expectation

A random variable which can take only countably many values is called *discrete*.
That is, X is discrete if there is a countable set S such that $P(X \in S) = 1$.

If X is a discrete random variable, we define the *probability mass function*, or p.m.f., of X by $p(x) = P(X = x)$ for $x \in \mathbb{R}$.

Thus $p : \mathbb{R} \to [0, 1]$ and $S = \{x : p(x) > 0\}$ is countable with $\sum_{x \in S} p(x) = P(X \in S) = 1$.

Countable additivity implies that the distribution of X is completely determined by p - for any $E \subseteq \mathbb{R}$, $P(X \in E) = P(X \in E \cap S) = \sum_{x \in E \cap S} p(x)$.

Conversely, every function $q : \mathbb{R} \to [0, \infty)$ with $R = \{x : q(x) > 0\}$ countable and $\sum_{x \in R} q(x) = 1$ is the p.m.f. of some discrete random variable Y.

Example 11.1. Suppose that W is a random variable taking values in $\mathbb{N}_0$ whose p.m.f. is given by $p(i) = \frac{c\lambda^i}{i!}$ for $i = 0, 1, ...$ where $\lambda > 0$ and c is some unknown constant.
What is $P(W = 0)$? What about $P(W > 2)$?

Since $1 = \sum_{i=0}^{\infty} p(i) = c \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = ce^{\lambda}$, we must have $c = e^{-\lambda}$.

Consequently, $P(W = 0) = p(0) = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda}$.

Similarly, $P(W > 2) = 1 - P(W \leq 2) = 1 - p(0) - p(1) - p(2) = 1 - e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2}\right)$.

If X is a discrete random variable with possible values $x_1, x_2, ...$, then we define its *cumulative distribution function*, or c.d.f., by

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i).$$

If $x_1 < x_2 < ...$, then we have $p(x_1) = F(x_1)$ and $p(x_i) = F(x_i) - F(x_{i-1})$ for $i > 1$.

In the previous example, the c.d.f. of W is

$$F(x) = \begin{cases} 0, & x < 0 \\ e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!}, & x \geq 0 \end{cases}$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

Even if X is not discrete, it still has a well-defined c.d.f. $F : \mathbb{R} \to [0, 1]$ given by by $F(x) = P(X \leq x)$.
(In contrast, the p.m.f. is only defined for discrete random variables.)

Theorem 11.1. *If X is a random variable with c.d.f. $F(x) = P(X \leq x)$, then F satisfies*

(1) *F is nondecreasing: $F(a) \leq F(b)$ whenever $a \leq b$*
(2) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$
(3) *F is right continuous:* $\lim_{x \to x_0^+} F(x) = F(x_0)$

Proof. For the first claim, if $a \leq b$, then $\{\omega : X(\omega) \leq a\} \subseteq \{\omega : X(\omega) \leq b\}$, so

$$F(a) = P\left(X \leq a\right) \leq P(X \leq b) = F(b).$$

For the second claim, if x_n is any sequence which decreases to $-\infty$, then $\{X \leq x_1\} \supseteq \{X \leq x_2\} \supseteq ...$ and $\bigcap_{n=1}^{\infty}\{X \leq x_n\} = \emptyset$, so continuity from above implies that

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P(X \leq x_n) = P\left(\bigcap_{n=1}^{\infty}\{X \leq x_n\}\right) = P(\emptyset) = 0.$$

Similarly, if $x_n \nearrow \infty$, then $\{X \leq x_n\}$ is a nested increasing sequence converging to Ω, so

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P(X \leq x_n) = P\left(\bigcup_{n=1}^{\infty}\{X \leq x_n\}\right) = P(\Omega) = 1.$$

Finally, if $x_n \searrow x_0$, then $\{X \leq x_n\} \searrow \{X \leq x_0\}$, so continuity from above implies $\lim_{n\to\infty} F(x_n) = F(x_0)$.

* Note that if $P(X = x_0) > 0$, the same basic argument shows that

$$\lim_{x\to x_0^-} F(x) = P(X < x) \neq P(X \leq x) = F(x),$$

so c.d.f.s are not necessarily continuous (though the set of discontinuity points is always countable).

□

The c.d.f. completely describes the distribution of X. For example, given real numbers $a < b$, we can compute

$$P(a < X < b) = P\left(X < b\right) - P(X \leq a) = \lim_{n\to\infty} P\left(X \leq b - \frac{1}{n}\right) - P(X \leq a) = \lim_{n\to\infty} F\left(b - \frac{1}{n}\right) - F(a).$$

If X and Y have the same c.d.f., then $P(X \in A) = P(Y \in A)$ for all sets A. In this case, we say that X and Y are equal in distribution (and write $X =_d Y$).
X and Y can be equal in distribution even if they are not defined on the same probability space.

One of the most important quantities associated with a random variable is its expectation.

The *expectation*, or expected value, of a random variable X (when it exists) is its probability-weighted average, and is often interpreted as the "best guess" for X.

For X discrete with p.m.f. p, the expected value of X is defined as

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

provided that this sum exists in $[-\infty, \infty]$.

Example 11.2. If X takes the values $0, 1, 2, 3, ..., n$ with equal probability, then

$$E[X] = \sum_{i=0}^{n} \frac{i}{n+1} = \frac{1}{n+1}\sum_{i=1}^{n} i = \frac{1}{n+1}\frac{n(n+1)}{2} = \frac{n}{2}.$$

Example 11.3. If X has p.m.f. $p(i) = e^{-\lambda}\frac{\lambda^i}{i!}$, $i = 0, 1, 2, ...$, then

$$E[X] = \sum_{i=0}^{\infty} ie^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^i}{(i-1)!} = \lambda e^{-\lambda}\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!} = \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda.$$

Example 11.4. If X has p.m.f. $p(n) = \frac{6}{\pi^2 n^2}$, $n = 1, 2, ...$ (which is a p.m.f. since $\sum_{n=1}^{\infty} n^{-2} = \frac{\pi^2}{6}$), then

$$E[X] = \sum_{n=1}^{\infty} n \frac{6}{\pi^2 n^2} = \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Example 11.5. If X has p.m.f. $p(n) = \frac{3}{\pi^2 n^2}$, $n = \pm 1, \pm 2, ...$, then

$$E[X] = \sum_{n \in \mathbb{Z} \setminus \{0\}} n \frac{6}{\pi^2 n^2} = \frac{6}{\pi^2} \sum_{n \in \mathbb{Z} \setminus \{0\}} \frac{1}{n}$$

does not exist.

Example 11.6. If $(\Omega, \mathcal{F}, P)$ is any probability space and A is an event, we define the *indicator* random variable

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}.$$

1_A is discrete since it takes values in the finite set $\{0, 1\}$.

Its p.m.f. is given by $p(1) = P(1_A = 1) = P(A)$ and $p(0) = P(1_A = 0) = P(A^C) = 1 - P(A)$, so

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

The utility of this simple example cannot be overemphasized.

Note that $1_{A^C} = 1$ if and only if $1_A = 0$, so $1_{A^C} = 1 - 1_A$.

If A and B are events, then $1_{A \cap B} = 1$ implies that A and B occurred and thus $1_A, 1_B = 1$. Conversely, $1_{A \cap B} = 0$ implies that at last one of A or B did not occur, so $1_A = 0$ or $1_B = 0$. It follows that $1_{A \cap B} = 1_A 1_B$.

Similarly, one can verify that $1_{A \cup B} = 1_A + 1_B - 1_A 1_B$ by checking that equality holds on each of the events $A \setminus B$, $A \cap B$, $B \setminus A$, $A^C \cap B^C$.

Example 11.7. In American Roulette, a wheel is divided into 18 red spaces, 18 black spaces, and 2 green spaces. The wheel is spun and a ball is dropped so that it is equally likely to come to rest on any of the 38 spaces. Suppose that a person places a \$1 bet on red with even odds – they win a dollar if the ball lands on red and lose a dollar otherwise. If X denotes the gambler's winnings after a single play, then

$$E[X] = 1 \cdot \frac{18}{38} + (-1) \frac{20}{38} = -\frac{1}{19},$$

so the gambler loses about 5 cents a play on average.

Proposition 11.1. *If X is a random variable taking values in $\mathbb{N}_0$, then $E[X] = \sum_{n=0}^{\infty} P(X > n)$.*

Proof.

$$\begin{aligned} \sum_{n=0}^{\infty} P(X > n) &= \sum_{n=1}^{\infty} P(X \geq n) = \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} p(m) \\ &= \sum_{n=1}^{\infty} np(n) = \sum_{n=0}^{\infty} np(n) = E[X] \end{aligned}$$

since $p(n)$ appears in n of the inner sums in the third expression, once for each lower limit $1 \leq i \leq n$. □

Ultimately, a random variable is a function $X : \Omega \to \mathbb{R}$, so if $f : \mathbb{R} \to \mathbb{R}$ is a function (e.g. $f(x) = x^2$), then $Y = f(X)$ is a random variable: $Y = f \circ X : \Omega \to \mathbb{R}$.

When X is discrete, $Y = f(X)$ can only take countably many values and thus is discrete as well. If p_X is the p.m.f. for X, then Y has p.m.f.

$$p_Y(y) = P(Y = y) = P\left(X \in f^{-1}(y)\right) = \sum_{x \in f^{-1}(y)} p_X(x).$$

Since $\{y : p_Y(y) > 0\} = \{f(x) : p_X(x) > 0\}$, we have

$$\begin{aligned} E[f(X)] &= \sum_{y : p_Y(y) > 0} y p_Y(y) = \sum_{y : p_Y(y) > 0} y \sum_{x \in f^{-1}(y)} p_X(x) \\ &= \sum_{y : p_Y(y) > 0} \sum_{x \in f^{-1}(y)} f(x) p_X(x) = \sum_{x : p_X(x) > 0} f(x) p_X(x). \end{aligned}$$

Example 11.8. A product which is sold seasonally yields a net profit of b dollars for each unit sold and a net loss of l dollars for each unit unsold by the end of the season. The number of units ordered is a random variable X with p.m.f. $p(n)$, $n \in \mathbb{N}_0$. How many units should the store stock in order to maximize its profit.

If s units are stocked, then the profit is given by $P_s(X) = \begin{cases} bX - l(s - X), & X \leq s \\ bs, & X > s \end{cases}$.

The expected profit when s units are stocked is thus

$$\begin{aligned} E[P_s(X)] &= \sum_{n=0}^{s} [bn - l(s-n)]\, p(n) + \sum_{n=s+1}^{\infty} bs p(n) \\ &= (b+l) \sum_{n=0}^{s} n p(n) - ls \sum_{n=0}^{s} p(n) + bs \left(1 - \sum_{n=0}^{s} p(n)\right) \\ &= bs + (b+l) \sum_{n=0}^{s} n p(n) - (b+l) s \sum_{n=0}^{s} p(n) \\ &= bs + (b+l) \sum_{n=0}^{s} (n-s) p(n). \end{aligned}$$

It follows that

$$\begin{aligned} E[P_{s+1}(X)] &= b(s+1) + (b+l) \sum_{n=0}^{s+1} (n-s-1) p(n) \\ &= b(s+1) + (b+l) \sum_{n=0}^{s} (n-s-1) p(n) \\ &= b + bs + (b+l) \sum_{n=0}^{s} (n-s) p(n) - (b+l) \sum_{n=0}^{s} p(n) \\ &= b - (b+l) \sum_{n=0}^{s} p(n) + E[P_s(X)], \end{aligned}$$

hence stocking $s+1$ units is better than stocking s units precisely when

$$\frac{b}{b+l} > \sum_{n=0}^{s} p(n).$$

Since the right-hand side is increasing in s, the optimal number of units is

$$s^* = \min\left\{ s \in \mathbb{N}_0 : \sum_{n=0}^{s} p(n) \geq \frac{b}{b+l} \right\}.$$

Proposition 11.2. *Let X be a discrete random variable with p.m.f. p, let $a, b \in \mathbb{R}$, and let $f, g : \mathbb{R} \to \mathbb{R}$ satisfy $f(x) \leq g(x)$ for all $x \in \mathbb{R}$. Then*

(1) $E[aX + b] = aE[X] + b$
(2) $E[f(X)] \leq E[g(X)]$

Proof. For the first assertion, we have

$$E[aX + b] = \sum_{x:p(x)>0} (ax + b)p(x) = a \sum_{x:p(x)>0} xp(x) + b \sum_{x:p(x)>0} p(x) = aE[X] + b,$$

and for the second,

$$E[f(X)] = \sum_{x:p(x)>0} f(x)p(x) \leq \sum_{x:p(x)>0} g(x)p(x) = E[g(X)]. \qquad \square$$

The expectation of X is often called its *mean*. Expectations of higher powers of X are often of interest as well. For $n \in \mathbb{N}$, we define the *nth moment* of X as

$$E[X^n] = \sum_{x:p(x)>0} x^n p(x).$$

When X has finite mean $\mu = E[X] \in \mathbb{R}$, we define the *nth central moment* of X as

$$E[(X - \mu)^n] = \sum_{x:p(x)>0} (x - \mu)^n p(x).$$

Of particular interest is the second central moment. This quantity measures how "spread out" the distribution of X is.

Definition. If X has mean μ, the *variance* of X is defined as $\mathrm{Var}(X) = E[(X - \mu)^2]$.

Proposition 11.3. *If X has finite mean, then*

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

Proof. Since $\mu = E[X] = \sum_{x:p(x)>0} xp(x)$, we have

$$\begin{aligned}
\mathrm{Var}(X) = E[(X - \mu)^2] &= \sum_{x:p(x)>0} (x - \mu)^2 p(x) \\
&= \sum_{x:p(x)>0} (x^2 - 2\mu x + \mu^2)p(x) \\
&= \sum_{x:p(x)>0} x^2 p(x) - 2\mu \sum_{x:p(x)>0} xp(x) + \mu^2 \sum_{x:p(x)>0} p(x) \\
&= E[X^2] - 2\mu \cdot \mu + \mu^2 \cdot 1 = E[X^2] - \mu^2. \qquad \square
\end{aligned}$$

Example 11.9. If X is a random variable with $P(X=-1)=P(X=1)=\frac{1}{2}$, then $E[X]=-1\cdot\frac{1}{2}+1\cdot\frac{1}{2}=0$ and $E[X^2]=(-1)^2\cdot\frac{1}{2}+1^2\cdot\frac{1}{2}=1$, so $\mathrm{Var}(X)=E[X^2]-E[X]^2=1$.

If Y is a random variable which takes values $-n,...,-1,0,1,...,n$ with equal probability, then $E[Y]=\sum_{k=-n}^{n}\frac{k}{2n+1}=0$ and

$$E[Y^2]=\sum_{k=-n}^{n}\frac{k^2}{2n+1}=\frac{2}{2n+1}\sum_{j=1}^{n}j^2=\frac{2}{2n+1}\cdot\frac{n(n+1)(2n+1)}{6}=\frac{n(n+1)}{3}.$$

(The "sum of squares formula" in the penultimate equality can be verified by induction.)
It follows that $\mathrm{Var}(Y)=E[Y^2]-E[Y]=\frac{n(n+1)}{3}$.

X and Y both have mean zero, but Y has a higher variance since it is "spread out" over more values.

Proposition 11.4. *For any $a,b\in\mathbb{R}$,*

$$\mathrm{Var}(aX+b)=a^2\mathrm{Var}(X).$$

Proof. Writing $\mu=E[X]$, Proposition 11.2 gives $E[aX+b]=a\mu+b$, so the definition of variance and Proposition 11.2 imply

$$\begin{aligned}\mathrm{Var}(aX+b)&=E\left[(aX+b-a\mu-b)^2\right]=E\left[a^2(X-\mu)^2\right]\\&=a^2E\left[(X-\mu)^2\right]=a^2\mathrm{Var}(X).\end{aligned}$$

□

Before moving on to discuss some of the more common families of discrete random variables, we observe that our definition of expectation is not really the most natural or the easiest to work with.

For example, using our definitions, it requires a lot of work to establish the simple and useful facts

Theorem 11.2. *If X and Y are random variables with finite means, then $E[X+Y]=E[X]+E[Y]$.*

Corollary 11.1. *If $X_1,...,X_n$ are random variables with finite means, then $E\left[\sum_{i=1}^{n}X_i\right]=\sum_{i=1}^{n}E[X_i]$.*

Theorem 11.3. *If X and Y are random variables with $P(X\leq Y)=1$, then $E[X]\leq E[Y]$ whenever the expectations exist.*

Moreover, the above results are true in general, but we are forced to treat separately the rather artificial (and incomplete) classes of discrete and (absolutely) continuous random variables.

To give a feel for the general technique, we will prove Theorem 11.2 under the assumption that the underlying sample space is countable. This can also be proved in terms of p.m.f.s without making this assumption, but we leave that as a homework exercise.

Note that Corollary 11.1 follows from Theorem 11.2 by an induction argument:
$X_1+....+X_{n+1}=(X_1+...+X_n)+X_{n+1}$.

We begin with what is essentially the correct way to define expectation

Proposition 11.5. *If X is a random variable on $(\Omega, \mathcal{F}, P)$ with Ω countable, then*

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}).$$

Proof. Let p be the p.m.f. of X, let $x_1, x_2, ...$ denote the possible values of X, and let $E_i = X^{-1}(x_i) = \{\omega : X(\omega) = x_i\}$. Then $\Omega = N \sqcup (\bigsqcup_i E_i)$ where $P(N) = 0$.
Since Ω is countable, the definitions imply

$$\begin{aligned}
\sum_{\omega \in \Omega} X(\omega) P(\{\omega\}) &= \sum_{\omega \in N} X(\omega) P(\{\omega\}) + \sum_i \sum_{\omega \in E_i} X(\omega) P(\{\omega\}) \\
&= \sum_{\omega \in N} X(\omega) \cdot 0 + \sum_i \sum_{\omega \in E_i} x_i P(\{\omega\}) \\
&= \sum_i x_i \sum_{\omega \in E_i} P(\{\omega\}) = \sum_i x_i P(E_i) \\
&= \sum_i x_i p(x_i) = E[X].
\end{aligned}$$

□

When expectation is defined in this manner, results such as Theorems 11.2 and 11.3 are almost immediate. For example,

Proof of Theorem 11.2 (countable sample space). If X and Y are random variables on $(\Omega, \mathcal{F}, P)$, then so is $Z(\omega) = X(\omega) + Y(\omega)$. If Ω is countable, then Proposition 11.5 implies

$$\begin{aligned}
E[Z] &= \sum_\omega Z(\omega) P(\{\omega\}) = \sum_\omega \left(X(\omega) + Y(\omega)\right) P(\{\omega\}) \\
&= \sum_\omega X(\omega) P(\{\omega\}) + \sum_\omega Y(\omega) P(\{\omega\}) = E[X] + E[Y].
\end{aligned}$$

□

Example 11.10. Suppose that n dice are rolled. Let X_i denote the value of the i^{th} die, and let X denote the sum of the dice. We have

$$E[X_i] = \sum_{j=1}^{6} \frac{j}{6} = \frac{7}{2}$$

for $i = 1, ..., n$, so the expected sum of the n dice is

$$E[X] = E\left[\sum_i^n X_i\right] = \sum_{i=1}^n E[X_i] = \frac{7n}{2}.$$

12. Common Discrete Distributions

Now that we understand the basic ideas associated with discrete random variables, we will consider a couple of the more famous discrete distributions. The main purpose of doing so is to practice the techniques we have learned so far. Also, these distributions often arise in applications so it's good to have some familiarity with them.

We will compute various quantities associated with each distribution, but it is not very useful to memorize the corresponding p.m.f.s, means, variances, etc... Rather, one should associate a story with each name, so that the p.m.f. is a simple consequence and other quantities can then be derived.

Bernoulli and Binomial.

Suppose that the outcome of an experiment can be considered either a success or a failure. If A is the set of outcomes which count as a success, then the random variable $X = 1_A$ takes the value 1 if the experiment results in a success and 0 if failure.

That is, $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $p = P(A)$. We say that X has the Bernoulli(p) distribution.

Clearly, $E[X] = 1 \cdot p + 0 \cdot (1-p) = p$. Also, $E[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$, so $\text{Var}(X) = E[X^2] - E[X]^2 - p - p^2 = p(1-p)$.

What is more interesting is if n independent trials are performed, each having success probability p (and failure probability $1 - p$). The random variable, X, which records the total number of successes is said to have the Binomial(n, p) distribution.

(An alternative description is $X = \sum_{k=1}^{n} X_k$ where the $X_k's$ are independent and identically distributed Bernoulli(p) random variables: X_k is the indicator of the event that the kth trial resulted in a success. At present, we will stick to the original description.)

The associated p.m.f. is given by

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

(Any particular sequence of k successes and $n - k$ failures has probability $p^k(1-p)^{n-k}$ by independence, and the number of such sequences is $\binom{n}{k}$ since a sequence is determined by specifying when the successes occur.)

The binomial theorem shows that p is indeed a p.m.f. since

$$\sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$$

A useful metaphor is that X gives the number of heads in n flips of a coin with heads probability p. Various examples of random variables having binomial distributions can be found in the text and homework.

To compute the moments of $X \sim \text{Binomial}(n, p)$, we recall the identity $i\binom{n}{i} = n\binom{n-1}{i-1}$.
(This can be derived algebraically or by counting the ways to select a committee of size i, with one member designated committee leader, from a group of n people.)

It follows that

$$\begin{aligned}E[X^k] &= \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} i^{k-1} \cdot i \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=1}^{n} i^{k-1} n \binom{n-1}{i-1} p^i (1-p)^{n-i} = np \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{(n-1)-(i-1)} \\ &= np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} = npE\left[(Y+1)^{k-1}\right]\end{aligned}$$

where $Y \sim \text{Binomial}(n-1, p)$.

Setting $k = 1$ shows that X has mean $E[X] = np$.

Setting $k = 2$ gives

$$\begin{aligned}E[X^2] &= npE[Y+1] = np\left(E[Y]+1\right) \\ &= np\left[(n-1)p+1\right] = (np)^2 + np - np^2,\end{aligned}$$

hence

$$\text{Var}(X) = E[X^2] - E[X]^2 = (np)^2 + np - np^2 - (np)^2 = np(1-p).$$

The following proposition shows that the Binomial(n, p) distribution is unimodal with mode $\lfloor (n+1)p \rfloor$.

Proposition 12.1. *If $X \sim$ Binomial(n, p), $n \in \mathbb{N}$, $p \in (0, 1)$, then the p.m.f. of X increases as k goes from 0 to $\lfloor (n+1)p \rfloor$ and then decreases as k goes from $\lfloor (n+1)p \rfloor$ to n.*

Proof. For $k = 1, ..., n$,

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

hence $p(k) \geq p(k-1)$ if and only if $(n-k+1)p \geq k(1-p)$, which happens if and only if $k \leq (n+1)p$. $\square$

Geometric and Negative Binomial.

Another quantity of interest in repeated Bernoulli trials is $X = \#$ trials until first success.

If the trials are i.i.d. with success probability p, then we say that X has a Geometric(p) distribution.

The p.m.f. of X is

$$p(k) = p(1-p)^{k-1}, \ k = 1, 2, ...$$

p is indeed a p.m.f. since it's nonnegative and has

$$\sum_{k=1}^{\infty} p(k) = p \sum_{k=1}^{\infty} (1-p)^{k-1} = p \sum_{j=0}^{\infty} (1-p)^j = \frac{p}{1-(1-p)} = 1.$$

To compute the c.d.f. of X, we first recall that for all integers $i < n$ and all real numbers $a \neq 1$,

$$S_{i,n}(a) = \sum_{k=i}^{n} a^k = \frac{a^i - a^{n+1}}{1-a}$$

(which follows by observing that $S_{i,n}(a) - aS_{i,n}(a) = a^i - a^{n+1}$).

Accordingly, for any $n \in \mathbb{N}$,

$$F(n) = P(X \leq n) = \sum_{k=1}^{n} p(1-p)^{k-1} = p\sum_{j=0}^{n-1}(1-p)^j = p\frac{1-(1-p)^n}{1-(1-p)} = 1-(1-p)^n.$$

Since X only takes positive integer values, we have

$$F(x) = \begin{cases} 0, & x < 1 \\ 1-(1-p)^{\lfloor x \rfloor}, & x \geq 1 \end{cases}.$$

The mean of X can be computed easily if we recall that power series can be differentiated termwise in the interior of their intervals of convergence:

$$\begin{aligned} E[X] &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p\sum_{k=0}^{\infty} k(1-p)^{k-1} \\ &= p\sum_{k=0}^{\infty} -\frac{d}{dp}(1-p)^k = -p\frac{d}{dp}\sum_{k=0}^{\infty}(1-p)^k \\ &= -p\frac{d}{dp}\frac{1}{p} = \frac{1}{p}. \end{aligned}$$

The second moment can be similarly computed, but for the sake of variety we observe that

$$\begin{aligned} E[X^2] &= \sum_{k=1}^{\infty} k^2 p(1-p)^{k-1} = \sum_{k=1}^{\infty}(k-1+1)^2 p(1-p)^{k-1} \\ &= \sum_{k=1}^{\infty}(k-1)^2 p(1-p)^{k-1} + 2\sum_{k=1}^{\infty}(k-1)p(1-p)^{k-1} + \sum_{k=1}^{\infty} p(1-p)^{k-1} \\ &= \sum_{j=0}^{\infty} j^2 p(1-p)^j + 2\sum_{j=0}^{\infty} jp(1-p)^j + 1 \\ &= (1-p)\sum_{j=1}^{\infty} j^2 p(1-p)^{j-1} + 2(1-p)\sum_{j=1}^{\infty} jp(1-p)^{j-1} + 1 \\ &= (1-p)E[X^2] + 2(1-p)E[X] + 1, \end{aligned}$$

so $pE[X^2] = \dfrac{2(1-p)}{p} + 1$.

Consequently,

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2-2p+p-1}{p^2} = \frac{1-p}{p^2}.$$

An obvious generalization of the geometric distribution is to let Y be the number of independent Bernoulli(p) trials until the r^{th} success.

In this case, we say that Y is a negative binomial random variable with parameters (r, p).

Since the r^{th} success occurs on trial n if and only if there are $r-1$ successes in the first $n-1$ trials and the n^{th} trial is a success, we see that Y has p.m.f.

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

To see that p is a p.m.f., we write

$$\sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} = \sum_{m=0}^{\infty} \binom{r+m-1}{m} p^r (1-p)^m$$

and note that

$$\begin{aligned}\binom{r+m-1}{m} &= \frac{(r+m-1)(r+m-2)\cdots r}{m!} = (-1)^m \frac{(-r-(m-1))\,(-r-(m-2))\cdots(-r)}{m!} \\ &= (-1)^m \frac{(-r)(-r-1)\cdots(-r-(m-1))}{m!} := (-1)^m \binom{-r}{m}\end{aligned}$$

where we are using the generalized binomial coefficients that appear in Newton's binomial series

$$(1+x)^{\alpha} = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$$

for $|x| < 1$ and arbitrary $\alpha \in \mathbb{R}$. (The series also converges for other combinations of x and α.)

* This also helps explain the name "negative binomial": If Z is the number of failures before the r^{th} success, then

$$P(Z = m) = P(Y = r + m) = \binom{r+m-1}{m} p^r (1-p)^m = \binom{-r}{m} p^r (p-1)^m.$$

With this identity in hand, we see that

$$\begin{aligned}\sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (1-p)^{n-r} &= p^r \sum_{m=0}^{\infty} \binom{r+m-1}{m} (1-p)^m = p^r \sum_{m=0}^{\infty} \binom{-r}{m} (p-1)^m \\ &= p^r \left[1 + (p-1)\right]^{-r} = p^r p^{-r} = 1.\end{aligned}$$

The easiest way to compute the mean and variance of $Y \sim \text{NegBin}(r, p)$ is to note that we can write $Y = \sum_{i=1}^{r} X_i$ where the X_i's are i.i.d. Geometric(p) random variables: X_1 is the number of trials until the first success, X_2 is the number of additional trials until the second success, and so forth.

Accordingly,

$$E[Y] = \sum_{i=1}^{r} E[X_i] = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}.$$

A similar computation gives $\text{Var}(Y)$ once we learn about expectations of products of independent random variables. For the time being, the reader is referred to the calculations in the text (using the same basic method as with the binomial distribution) which show that $\text{Var}(Y) = \dfrac{r(1-p)}{p^2}$.

Example 12.1. In a sequence of independent Bernoulli trials with success probability p, what is the probability that m successes occur before n failures?

This happens if and only if the m^{th} success occurs by the $(m+n-1)^{\text{st}}$ trial.

Thus, taking $X \sim \text{NegBin}(m,p)$, we see that the desired probability is

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

Example 12.2. A pipe-smoking mathematician always carries two boxes of matches - one is his right pocket and another in his left pocket. Each time he needs a match, he is equally likely to take it from either pocket. Suppose that each box initially contains n matches. What is the probability that once the mathematician discovers that one of the boxes is empty, there are exactly k matches in the other box, $k = 0, 1, ..., n$?

Let E be the event that the mathematician first discovers that the right-hand matchbox is empty and that there are k matches in the left-hand box.

In order for E to occur, the $(n+1)^{\text{st}}$ choice of the right-hand matchbox must occur on the $(n+1+n-k)^{\text{th}}$ trial. This is equivalent to the probability that $X = 2n-k+1$ where $X \sim \text{NegBin}\left(n+1, \frac{1}{2}\right)$: A "success" is choosing from the right pocket.

Thus

$$P(E) = P(X = 2n-k+1) = \binom{2n-k}{n} \left(\frac{1}{2}\right)^{n+1} \left(1-\frac{1}{2}\right)^{n-k}.$$

By the symmetry of the problem, we have

$$P\left(k \text{ matches in other box}\right) = 2P(E) = \binom{2n-k}{n} \left(\frac{1}{2}\right)^{2n-k}.$$

Hypergeometric.

Suppose that a population consists of N items, of which m are distinguished. (Think $N-m$ white balls and m black balls.) If n items are sampled without replacement, then

$X = \#$ of distinguished items in sample is called a hypergeometric random variable with parameters (N, m, n).

It follows from the preceding story that

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \; k = 0, 1, ..., n.$$

(Note that if $k < n-(N-m)$ or $k > m, n$, then $P(X=k) = 0$ by definition of the binomial coefficient, and this also makes sense in terms of the sampling story.)

To compute the mean of X, we recall that $\binom{r}{j} = \frac{r}{j}\binom{r-1}{j-1}$, so if $X \sim \text{Hypergeometric}(N, m, n)$,

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} \frac{k\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \sum_{k=1}^{n} \frac{k\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \\ &= \sum_{k=1}^{n} \frac{m\binom{m-1}{k-1}\binom{N-m}{n-k}}{\frac{N}{n}\binom{N-1}{n-1}} = \frac{mn}{N} \sum_{k=1}^{n} \frac{\binom{m-1}{k-1}\binom{(N-1)-(m-1)}{(n-1)-(k-1)}}{\binom{N-1}{n-1}} \\ &= \frac{mn}{N} \sum_{i=0}^{n-1} \frac{\binom{m-1}{i}\binom{(N-1)-(m-1)}{(n-1)-i}}{\binom{N-1}{n-1}} = \frac{mn}{N} \end{aligned}$$

since $\frac{\binom{m-1}{i}\binom{(N-1)-(m-1)}{(n-1)-i}}{\binom{N-1}{n-1}} = P(Y=i)$ for $Y \sim \text{Hypergeometric}(N-1, m-1, n-1)$.

Higher moments can be computed similarly. See the text for details.

Example 12.3. Suppose that we wish to estimate the number N of animals inhabiting a certain region. The capture-recapture method proceeds by first capturing and marking some number, m, of the animals. The animals are then released and after some time, they capture a set of n animals. Assuming that all subsets of the N animals of a given size are equally likely to be caught each time (which is a bit of a stretch), the number of marked animals in the second catch is $X \sim \text{Hypergeometric}(N, m, n)$.

If k animals are observed in the second catch, then we have seen the realization of an outcome with probability

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} := P_k(N).$$

The *maximum likelihood estimate* for N is the value which maximizes $P_k(N)$. To find the maximum value, we note that

$$\frac{P_k(N)}{P_k(N-1)} = \frac{\binom{N-1}{n}}{\binom{N}{n}} \frac{\binom{N-m}{n-k}}{\binom{N-1-m}{n-k}} = \frac{N-n}{N} \cdot \frac{N-m}{N-m-n+k} \geq 1$$

if and only if

$$1 - \frac{n}{N} = \frac{N-n}{N} \geq \frac{N-m-n-k}{N-m} = 1 - \frac{n-k}{N-m}$$

if and only if

$$nN - kN = N(n-k) \geq n(N-m) = nN - nm$$

if and only if

$$N \leq \frac{mn}{k}.$$

Therefore, if k marked animals appear in the second sample, the population is estimated to be of size $\left\lfloor \frac{mn}{k} \right\rfloor$.

Note that if $n \ll N, m$, then we should get roughly the same answers whether we sample with replacement or without replacement.
Specifically, if $p = \frac{m}{N}$ is the proportion of distinguished items in the population, then we the number of items in the sample, X, should be approximately $\text{Binomial}(n, p)$.

To check this intuition, we compute

$$\begin{aligned}
P(X = k) &= \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} = \frac{m!}{(m-k)!k!} \frac{(N-m)!}{(N-m-n+k)!(n-k)!} \frac{(N-n)!n!}{N!} \\
&= \binom{n}{k} \frac{m!}{(m-k)!} \frac{(N-m)!}{(N-m-n+k)!} \frac{(N-n)!}{N!} \\
&= \binom{n}{k} \frac{\frac{m!}{(m-k)!}}{\frac{N!}{(N-k)!}} \cdot \frac{(N-m)!}{(N-k)!} \cdot \frac{(N-n)!}{(N-m-(n-k))!} \\
&= \binom{n}{k} \frac{\frac{m!}{(m-k)!}}{\frac{N!}{(N-k)!}} \cdot \frac{\frac{(N-m)!}{(N-m-(n-k))!}}{\frac{(N-n+(n-k))!}{(N-n)!}} = \binom{n}{k} \prod_{i=1}^{k} \frac{m-k+i}{N-k+i} \prod_{j=1}^{n-k} \frac{N-m-(n-k)+j}{N-n+j} \\
&\approx \binom{n}{k} \left(\frac{m}{N}\right)^k \left(\frac{(N-n)-(m-k)}{N-n}\right)^{n-k} \approx \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

(This calculation should probably be skipped.)

Poisson.

The final major discrete random variable we will discuss is the Poisson.

If, for some $\lambda > 0$, X is an $\mathbb{N}_0$-valued random variable with p.m.f.

$$p(n) = e^{-\lambda}\frac{\lambda^n}{n!},\ n = 0, 1, 2, ...,$$

then X is said to have the Poisson distribution with rate λ.

We have already shown that p is actually a p.m.f. and computed

$$E[X] = \sum_{n=0}^{\infty} ne^{-\lambda}\frac{\lambda^n}{n!} = \lambda e^{-\lambda}\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{(n-1)!} = \lambda e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!} = \lambda.$$

Similarly, one finds that

$$\begin{aligned} E[X^2] &= \sum_{n=0}^{\infty} n^2 e^{-\lambda}\frac{\lambda^n}{n!} = \sum_{n=0}^{\infty}\left[n(n-1)+n\right]e^{-\lambda}\frac{\lambda^n}{n!} \\ &= \sum_{n=2}^{\infty} n(n-1)e^{-\lambda}\frac{\lambda^n}{n!} + \sum_{n=0}^{\infty} ne^{-\lambda}\frac{\lambda^n}{n!} \\ &= \lambda^2 e^{-\lambda}\sum_{n=2}^{\infty}\frac{\lambda^{n-2}}{(n-2)!} + E[X] \\ &= \lambda^2 e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!} + \lambda = \lambda^2 + \lambda, \end{aligned}$$

so

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

as well.

One of the reasons that the Poisson distribution is so important is that when n is large and $\lambda = np$ is of moderate size, the Poisson(λ) distribution is a good approximation to the Binomial(n, p) distribution.

Indeed, suppose that $X \sim \text{Binomial}(n, p)$. Then

$$\begin{aligned} P(X = k) &= \binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \cdot \frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

If λ and k are much smaller than n, then

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \quad \frac{n(n-1)\cdots(n-k+1)}{n^k} \approx 1, \quad \left(1-\frac{\lambda}{n}\right)^{-k} \approx 1,$$

so $P(X = k) \approx e^{-\lambda}\dfrac{\lambda^k}{k!}$.

Many real world phenomena have been shown to be well approximated by a Poisson distribution, such as

- The number of soldiers in the Prussian cavalry killed per year by horse kicks in the late 1800's
- The spatial distribution of yeast cells used in brewing Guinness beer
- The number of goals scored in soccer games
- The number of mutations in a given segment of DNA after radiation exposure
- The number of photons per second arriving on a particular photodetector

In general, the *Poisson paradigm* says that if events $E_1, ..., E_n$ are "weakly dependent" with n "large" and $P(E_i) = p_i$ "small" for $i = 1, ..., n$, then the number of events which occur is approximately Poisson with mean $\sum_{i=1}^{n} p_i$. This heuristic can be made precise in terms of the "law of small numbers."

Thus in addition to being more computationally tractable than the binomial distribution, the Poisson also has broader applicability.

Example 12.4. On the first day of class, we saw that in a room of 23 or more people, it is more likely than not that at least 2 share a birthday (using the slightly inaccurate assumptions that each person is equally likely to be born on any given day of the year independently of the rest, and excluding leap year birthdays).

This is because the probability that no 2 people share a birthday in a room of n people is $\frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-n+1}{365}$, which is approximately 0.4927 for $n = 23$. However, such calculations are rather tedious and numerically unstable, especially for similar problems where the numbers involved are larger.

To approximate the probability of no common birthdays in a room of n people using the Poisson distribution, we note that for each of the $\binom{n}{2}$ pairs of people, the probability that they share a birthday is $\frac{1}{365}$.

Though the events $E_{i,j} = \{\text{persons } i \text{ and } j \text{ share a birthday}\}$, $1 \leq i < j \leq n$, are pairwise independent, they are not independent as a whole. Still the dependence is rather weak, so we expect that the number of $E_{i,j}$'S which occur should be approximately Poisson(λ) with $\lambda = \binom{n}{2}\frac{1}{365} = \frac{n(n-1)}{730}$. It follows that

$$P(\text{no shared birthdays}) = P(0 \text{ successes}) \approx \exp\left(-\frac{n(n-1)}{730}\right).$$

To determine the smallest integer n for which this probability is less than $\frac{1}{2}$, we solve

$$\exp\left(-\frac{n(n-1)}{730}\right) \leq \frac{1}{2},$$

or

$$\exp\left(\frac{n(n-1)}{730}\right) \geq 2.$$

Taking logarithms, this reduces to the quadratic inequality

$$n(n-1) \geq 730 \log(2).$$

The smallest integer solution is $n = 23$.

If $X \sim \text{Poisson}\left(\frac{23 \cdot 22}{730}\right)$, then $P(X = 0) \approx 0.4952$, which matches the true answer to two decimal places.

Simple variants of the birthday problem (involving near matches, for example) are much more difficult to compute exactly, but the Poisson paradigm offers easy approximations.

For instance, suppose that we wanted to known how many people one needs to have better than even odds that at least 3 people share a birthday.

In a room of n people, there are $\binom{n}{3}$ different groups of 3, and the probability that any 3 share a birthday is $\frac{1}{365^2}$, so $X = \#$ triple matches is approximately Poisson $\left(\frac{n(n-1)(n-2)}{6\cdot 365^2}\right)$, hence

$$P(\text{no triple match}) \approx \exp\left(-\frac{n(n-1)(n-2)}{6\cdot 365^2}\right),$$

which is less than $\frac{1}{2}$ precisely when $n(n-1)(n-2) > 6\cdot 365^2 \log(2)$.

Solving the cubic equation shows that the odds of three people having a common birthday are better than 50% when $n \geq 84$.

In addition to arising as a limit of other probability distributions, the Poisson distribution is also a natural model for the number of occurrences of "rare" events in a given time interval.

Typical examples are the number of arrivals per minute at an eatery or the number of major earthquakes per year near a fault line.

Specifically, let $N(s,t)$ denote the number of occurrences in the time interval $(s,t]$ for $0 \leq s < t$. If

(1) The number of occurrences in disjoint time intervals are independent
(2) The distribution of $N(s,t)$ depends only on $t-s$
(3) $P\left(N(0,h)=1\right) = \lambda h + o(h)$
(4) $P\left(N(0,h)\geq 2\right) = o(h)$,

then $N(t) := N(0,t) \sim \text{Poisson}(\lambda t)$.

* $o(h)$ refers to any function f such that $\lim_{h\to 0}\frac{f(h)}{h} = 0$.

To see that this is so, fix $k \in \mathbb{N}_0$ and break up the interval $(0,t]$ into $n \geq k$ disjoint pieces of equal length: $\left(0,\frac{t}{n}\right], \left(\frac{t}{n},\frac{2t}{n}\right], \ldots, \left(\frac{t(n-1)}{n}, t\right]$.

Set

$$A_{n,k} = \{k \text{ intervals contain a single occurrence and } n-k \text{ contain no occurrences}\},$$
$$B_{n,k} = \{N(0,t) = k \text{ and at least one interval contains multiple occurrences}\}.$$

Then

$$P\left(N(0,t)=k\right) = P(A_{n,k}) + P(B_{n,k})$$

for all $n \geq k$.

Now

$$\begin{aligned} P(B_{n,k}) &\leq P\left(\text{at least one interval contains multiple occurrences}\right) \\ &= P\left(\bigcup_{i=1}^{n}\{\text{interval } i \text{ contains multiple occurrences}\}\right) \\ &\leq \sum_{i=1}^{n} P\left(\text{interval } i \text{ contains multiple occurrences}\right) \\ &= \sum_{i=1}^{n} o\left(\frac{t}{n}\right) = no\left(\frac{t}{n}\right) = t\cdot\frac{o\left(\frac{t}{n}\right)}{\frac{t}{n}} \to 0 \text{ as } n\to\infty. \end{aligned}$$

Also, assumptions $2-4$ imply that for any subinterval I of length $\frac{t}{n}$,

$$P\left(1 \text{ event occurs in } I\right) = \frac{\lambda t}{n} + o\left(\frac{t}{n}\right),$$

$$P\left(0 \text{ events occur in } I\right) = 1 - \left(\lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right) - o\left(\frac{t}{n}\right) = 1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right).$$

It follows that

$$\begin{aligned} P(A_{k,n}) &= \binom{n}{k}\left(\frac{\lambda t}{n} + o\left(\frac{t}{n}\right)\right)^k\left(1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right)\right)^{n-k} \\ &= \frac{1}{k!}\cdot\frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot n^k\left(\frac{\lambda t}{n} + o\left(\frac{t}{n}\right)\right)^k\cdot\left(1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right)\right)^n\cdot\left(1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right)\right)^{-k}. \end{aligned}$$

Because

$$\begin{gathered} \frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1, \\ n^k\left(\frac{\lambda t}{n} + o\left(\frac{t}{n}\right)\right)^k = \left(\lambda t + t\cdot\frac{o\left(\frac{t}{n}\right)}{\frac{t}{n}}\right)^k \to (\lambda t)^k, \\ \left(1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right)\right)^n \to e^{-\lambda t}, \\ \left(1 - \frac{\lambda t}{n} - o\left(\frac{t}{n}\right)\right)^{-k} \to 1 \end{gathered}$$

for fixed k, t, λ as $n \to \infty$, we see that

$$\begin{aligned} P\left(N(0,t) = k\right) &= \lim_{n\to\infty}\left(P(A_{n,k}) + P(B_{n,k})\right) \\ &= \lim_{n\to\infty} P(A_{k,n}) + \lim_{n\to\infty} P(B_{n,k}) \\ &= \frac{1}{k!}\cdot 1\cdot(\lambda t)^k\cdot e^{-\lambda t}\cdot 1 + 0 = e^{-\lambda t}\frac{(\lambda t)^k}{k!}. \end{aligned}$$

Example 12.5. Suppose that earthquakes occur in the western U.S. in accordance with assumptions $1-4$ at a rate of $\lambda = 2$ per week.

a): What is the probability that at least 3 earthquakes occur in the next two weeks?

b): What is the distribution of the time between successive earthquakes?

The number of earthquakes in the next two weeks is given by $N(2) \sim \text{Poisson}(4)$, hence

$$\begin{aligned} P\left(N(2) \geq 3\right) &= 1 - P(N(2) = 0) - P(N(2) = 1) - P(N(2) = 2) \\ &= 1 - e^{-4} - 4e^{-4} - \frac{4^2}{2}e^{-4} = 1 - 13e^{-4}. \end{aligned}$$

Let T be the length of time in weeks between successive earthquakes. Then

$$P(T > t) = P\left(N(t) = 0\right) = e^{-2t},$$

so T has distribution function $F(t) = P(T \leq t) = \begin{cases} 1 - e^{-2t}, & t > 0 \\ 0, & t \leq 0 \end{cases}$.

(The time between successive earthquakes is the same as the time until the next earthquake by properties 1 and 2.)

13. Absolutely Continuous Random Variables

So far, we have been discussing variables which take on only countably many values.

However, many quantities of interest, such as times, temperatures, or distances, can be thought of as taking values in all of $\mathbb{R}$ or some uncountable subset thereof.

A random variable $X : \Omega \to \mathbb{R}$ with the property that $P(X = x) = 0$ for all $x \in \mathbb{R}$ is called *continuous*.

Continuous random variables are not discrete (and conversely) since if $S \subseteq \mathbb{R}$ is countable and $P(X = x) = 0$ for all x, then $P(X \in S) = \sum_{x \in S} P(X = s) = 0$.

Of course, it is possible for random variables to be neither continuous nor discrete. For example, X might satisfy $P(X = 0) = \frac{1}{2}$ and $P(X \in (a, b)) = \frac{b^2 - a^2}{2}$ for all $0 < a < b \leq 1$.

Most continuous random variables of practical interest are *absolutely continuous*, which means that they have a *density*, f, which satisfies $P(X \in A) = \int_A f(x)dx$ for all (measurable) $A \subseteq \mathbb{R}$.

Thus f specifies the distribution of X in that the probability that X is in some set (a, b), for example, is given by the area under the graph of f between points a and b.

* Sometimes we will call densities p.d.f.s (for probability density function).

Absolutely continuous random variables are clearly continuous since if X has density f, then for any $a \in \mathbb{R}$, $P(X = a) = \int_{\{a\}} f(x)dx = \int_a^a f(x)dx = 0$.

In particular, for real numbers $a \leq b$,

$$P(X \in (a, b)) = P(X \in [a, b]) = P(X \in (a, b]) = P(X \in [a, b)) = \int_a^b f(x)dx.$$

Henceforth, we will just say "continuous" instead of "absolutely continuous" as we will only consider continuous random variables which have densities.

Note that if f is the density of some random variable X, then we must have $f(x) \geq 0$ for all x (as probabilities are nonnegative) and $\int_{-\infty}^{\infty} f(x)dx = P(X \in \mathbb{R}) = 1$. The latter condition implies that $\lim_{y \to \pm\infty} f(y) = 0$.

Indeed, any nonnegative function which integrates to 1 is the density of some continuous random variable.

In many regards, everything is exactly the same as with discrete random variables except that densities take the place of mass functions and integrals take the place of sums.

For example, if X is discrete with p.m.f. p and support $\{x_1, x_2, ...\}$, then

$$P(a \leq X \leq b) = \sum_{a \leq x_i \leq b} p(x_i)$$

and if Y is continuous with density f, then

$$P(a \leq Y \leq b) = \int_a^b f(y)dy.$$

In fact, it is often easier to work with continuous random variables because we can use calculus.

Continuing with the previous example, X has c.d.f.

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

and Y has c.d.f.

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f(z)dz.$$

If it happens that $x_1 < x_2 < ...$, then we can recover the p.m.f. of X from its c.d.f. by

$$p(x_i) = F_X(x_i) - F_X(x_{i-1}).$$

We can always recover p.d.f.s from c.d.f.s since the fundamental theorem of calculus shows that

$$f(y) = \frac{d}{dy}\int_{-\infty}^{y} f(z)dz = F_Y'(y).$$

Example 13.1. Suppose that X is continuous with p.d.f. $f(x) = \begin{cases} C(2x - x^2), & 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$ where C is some unknown constant.

What is the probability that X is greater than 1?

The first step is to determine C. Since f is a density, we must have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)dx = C\int_0^2 (2x - x^2)dx \\ &= C\left(x^2 - \frac{x^3}{3}\Big|_0^2\right) = C\left(4 - \frac{8}{3}\right) = \frac{4}{3}C. \end{aligned}$$

It follows that

$$\begin{aligned} P(X > 1) &= \int_1^{\infty} f(x)dx = \frac{3}{4}\int_1^2 (2x - x^2)dx = \frac{3}{4}\left(x^2 - \frac{x^3}{3}\Big|_1^2\right) \\ &= \frac{3}{4}\left[\left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right)\right] = \frac{1}{2}. \end{aligned}$$

Example 13.2. The lifetime (in hours) of a certain type of battery is a random variable having density $f(x) = \frac{50}{x^2}$, $x \geq 50$.

An electronic device requires 4 such batteries. Assuming that the battery lifetimes are independent, what is the probability that the device will function at least 80 hours before any batteries need to be changed?

Letting X_i denote the lifetime of battery i, we have

$$P(X_i > 100) = \int_{80}^{\infty} \frac{50}{x^2}dx = \frac{-50}{x}\Big|_{80}^{\infty} = 0 - \left(-\frac{50}{80}\right) = \frac{5}{8}.$$

By independence,

$$P(\text{all batteries last at least 80 hours}) = \left(\frac{5}{8}\right)^4.$$

Recall that we defined the expected value of a discrete random variable W with p.m.f. p as

$$E[W] = \sum_{x:p(x)>0} xp(x)$$

whenever this sum exists.

In light of the parallels between discrete and continuous random variables, it is natural to define the expectation of a continuous random variable X with density f as

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

whenever the integral exists.

One way to see that this makes sense is to observe that if f is continuous, then for ε sufficiently small,

$$P\left(x \leq X \leq x+\varepsilon\right) = \int_{x}^{x+\varepsilon} f(y)dy \approx \varepsilon f(x).$$

If f is supported on $[a,b]$, then breaking $[a,b]$ into intervals $\left(a+i\frac{b-a}{n}, a+(i+1)\frac{b-a}{n}\right]$, $i = 0, 1, ..., n-1$, and approximating X by the discrete random variable X_n with p.m.f.

$$p_n\left(a+i\frac{b-a}{n}\right) = P\left(a+i\frac{b-a}{n} < X \leq a+(i+1)\frac{b-a}{n}\right) \approx \frac{1}{n} f\left(a+i\frac{b-a}{n}\right),$$

we get the estimate

$$E[X] \approx E[X_n] = \sum_{i=0}^{n-1} \left(a+i\frac{b-a}{n}\right) p_n\left(a+i\frac{b-a}{n}\right) \approx \sum_{i=0}^{n-1} \left(a+i\frac{b-a}{n}\right) f\left(a+i\frac{b-a}{n}\right) \frac{1}{n}.$$

Intuitively, the approximations are better for larger values of n, and the right-hand side is a Riemann sum which converges to $\int_a^b xf(x)dx$ as $n \to \infty$ (since densities are assumed to be integrable).

If f is supported on an infinite interval (such as all of $\mathbb{R}$), we first approximate X by the random variable $Y_{M,N} = \begin{cases} X, & -M \leq X \leq N \\ 0, & \text{otherwise} \end{cases}$ and apply the above procedure to $Y_{M,N}$, and then we send $M, N \to \infty$.

Of course, the preceding is not a rigorous proof of anything, but serves only to show that the definition is plausibly compatible with the definition in the discrete case.

Example 13.3.

Radioactive decay is the process by which the nucleus of an unstable atom loses energy by emitting radiation.

The time it takes for this decay to occur is modeled as a random variable T with p.d.f. $f(t) = \lambda e^{-\lambda t}$, $t \geq 0$ for some $\lambda > 0$ depending on the substance.

The value of λ is determined empirically by measuring the fraction of unstable atoms in a sample that undergo decay in a given time period. This is typically reported in terms of the half-life, which is the amount of time needed for the activity of a sample to decay to half its original value.

What is the expected length of time for an atom with half-life τ to decay into a stable isotope?

We first compute

$$E[T] = \int_{-\infty}^{\infty} tf(t)dt = \int_0^{\infty} \lambda t e^{-\lambda t} dt = \lambda \left[\frac{-t}{\lambda} e^{-\lambda t} \Big|_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda t} dt \right] = \int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda}.$$

To find λ, we solve

$$\frac{1}{2} = P(T \leq \tau) = \lambda \int_0^{\tau} e^{-\lambda t} dt = -e^{-\lambda t}\big|_0^{\tau} = 1 - e^{-\lambda \tau},$$

Rearranging, taking logarithms, and dividing by the half-life gives $\lambda = \frac{\log(2)}{\tau}$, hence

$$E[T] = \frac{\tau}{\log(2)}.$$

Example 13.4. A random variable X is said to have the standard Cauchy distribution if it has density $f(x) = \dfrac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$.

Taking the antiderivative, we see that X has c.d.f. $F(x) = \frac{1}{\pi} \arctan(x) + C$ for some $C \in \mathbb{R}$. Because $1 = \lim\limits_{x\to\infty} F(x) = \dfrac{1}{\pi} \cdot \dfrac{\pi}{2} + C$, we must have $C = \frac{1}{2}$.

F is nondecreasing, continuous (and thus right-continuous), and satisfies $\lim\limits_{x\to\infty} F(x) = 1$ and $\lim\limits_{x\to-\infty} F(x) = 0$, so it is the c.d.f. of a random variable, hence $f = F'$ is indeed a p.d.f.

* Alternatively, f is nonnegative and

$$\int_{-\infty}^{\infty} f(x)dx = \frac{1}{\pi} \lim_{\substack{a\to-\infty \\ b\to\infty}} \int_a^b \frac{dx}{1+x^2} = \frac{1}{\pi} \lim_{\substack{a\to-\infty \\ b\to\infty}} (\arctan(b) - \arctan(a)) = \frac{1}{\pi}\left(\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right) = 1.$$

f is clearly symmetric about zero, but $xf(x) = \dfrac{x}{\pi(1+x^2)} \approx \dfrac{1}{\pi x}$ for $|x|$ large, hence $E[X]$ does not exist.

This also provides a nice example of how the *Cauchy principal value* $\lim\limits_{a\to\infty} \int_{-a}^{a} g(x)dx$ is not necessarily the same as

$$\int_{-\infty}^{\infty} g(x)dx := \lim_{\substack{b\to\infty \\ a\to-\infty}} \int_a^b g(x)dx.$$

For instance, $\lim\limits_{a\to\infty} \int_{-a}^{a} \dfrac{x}{\pi(1+x^2)} dx = 0$ since the integrand is odd, but

$$\lim_{a\to\infty} \int_{-2a}^{a} \frac{x}{\pi(1+x^2)} dx = \frac{1}{2\pi} \lim_{a\to\infty} \log\left(1+x^2\right)\Big|_{-2a}^{a} = \frac{1}{2\pi} \lim_{a\to\infty} \log\left(\frac{1+a^2}{1+4a^2}\right) = -\frac{\log(4)}{2\pi},$$

so the improper Riemann integral $\displaystyle\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} dx$ is not defined.

Based on our experience with discrete random variables, we expect that if $g : \mathbb{R} \to \mathbb{R}$ is a function, then $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ whenever the integral exists.

To prove this assertion from the definition $E[X] = \int_{-\infty}^{\infty} xf(x)dx$, we first need the following lemma.

Lemma 13.1. *If Y is a random variable with density f_Y and either Y is nonnegative or $E|Y| < \infty$, then*

$$E[Y] = \int_0^{\infty} P(Y > y)dy - \int_0^{\infty} P(Y < -y)dy.$$

Proof. Since one can interchange the order of integration for nonnegative integrands, we have

$$\begin{aligned}\int_0^\infty P(Y>y)dy &= \int_0^\infty \left(\int_y^\infty f_Y(x)dx\right)dy = \int_0^\infty \left(\int_0^x f_Y(x)dy\right)dx \\ &= \int_0^\infty f_Y(x)\left(\int_0^x dy\right)dx = \int_0^\infty x f_Y(x)dx.\end{aligned}$$

Similarly,

$$\int_0^\infty P(Y<-y)dy = \int_0^\infty \int_{-\infty}^{-y} f_Y(x)dxdy = \int_{-\infty}^0 \int_0^{-x} f_Y(x)dydx = -\int_{-\infty}^0 x f_Y(x)dx.$$

Thus if Y is nonnegative, then

$$\begin{aligned}E[Y] &= \int_0^\infty x f_Y(x)dx = \int_0^\infty P(Y>y)dy \\ &= \int_0^\infty P(Y>y)dy - \int_0^\infty P(Y<-y)dy,\end{aligned}$$

and if $E|Y|<\infty$, then

$$\begin{aligned}E[Y] &= \int_{-\infty}^\infty x f_Y(x)dx = \int_0^\infty x f_Y(x)dx + \int_{-\infty}^0 x f_Y(x)dx \\ &= \int_0^\infty P(Y>y)dy - \int_0^\infty P(Y<-y)dy.\end{aligned}$$

The condition $E|Y|<\infty$ ensures that the integral can be decomposed as above. □

Using Lemma 13.1, we can prove

Theorem 13.1. *If X has density f_X and $g:\mathbb{R}\to\mathbb{R}$ is a function, then*

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

whenever g is nonnegative or $E|g(X)|<\infty$.

Proof. If g is nonnegative, then

$$\begin{aligned}E[g(X)] &= \int_0^\infty P(g(X)\geq y)dy = \int_0^\infty \int_{x:g(x)>y} f_X(x)dxdy \\ &= \int_{x:g(x)>0}\int_0^{g(x)} f_X(x)dydx = \int_{x:g(x)>0} g(x)f_X(x)dx = \int_{-\infty}^\infty g(x)f_X(x)dx.\end{aligned}$$

If $E|g(X)|<\infty$, then

$$\begin{aligned}E[g(X)] &= \int_0^\infty P(g(X)>y)dy - \int_0^\infty P(g(X)<-y)dy \\ &= \int_{x:g(x)>0}\int_0^{g(x)} f_X(x)dydx - \int_{x:g(x)<0}\int_0^{-g(x)} f_X(x)dydx \\ &= \int_{-\infty}^\infty g(x)f_X(x)dx.\end{aligned}$$

□

Example 13.5. A stick of length 1 is split at a point U, which is uniformly distributed over $[0,1]$. That is, U has p.d.f. $f(u) = 1$, $0 \le u \le 1$.

For $p \in [0,1]$, determine the length of the piece containing the point p.

The length of the piece containing p is $L_p(U) = \begin{cases} 1-U, & U < p \\ U, & U > p \end{cases}$. (We don't have to worry about $U = p$ since this happens with probability 0.)

Accordingly,

$$\begin{aligned} E\left[L_p(U)\right] &= \int_0^1 L_p(u)du = \int_0^p (1-u)du + \int_p^1 u du \\ &= \left(u - \frac{u^2}{2}\Big|_0^p\right) + \left(\frac{u^2}{2}\Big|_p^1\right) = p - \frac{p^2}{2} + \frac{1}{2} - \frac{p^2}{2} = \frac{1}{2} + p - p^2. \end{aligned}$$

* Note that the longer of the two pieces must contain the point $p = \frac{1}{2}$, so the expected length of the longest piece is $\frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$.

By paralleling the proofs in the discrete case, one can derive many familiar properties such as $E[aX+b] = aE[X] + b$.

Similarly, the variance $\mathrm{Var}(X) = E\left[(X - E[X])^2\right]$ satisfies $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$ and can be computed as $\mathrm{Var}(X) = E[X^2] - E[X]^2$.

It is also the case that $E[X+Y] = E[X] + E[Y]$ whenever these expectations exist, but we need the concept of joint distributions to prove this fact.

(Really, the proof is just that $E[X] = \int_\Omega X(\omega)dP(\omega)$ for any random variable X on $(\Omega, \mathcal{F}, P)$, so

$$E[X+Y] = \int (X+Y)dP = \int X dP + \int Y dP = E[X] + E[Y],$$

but this requires a more sophisticated theory of integration than we have at our disposal.)

Before moving on to consider specific examples, we note that in many cases one can recover the distribution of $g(X)$ from that of X.

Example 13.6. Suppose that X is nonnegative with p.d.f. f and c.d.f. F. To find the c.d.f. of X^n, we note that for any $x \ge 0$,

$$F_n(x) = P(X^n \le x) = P\left(X \le x^{\frac{1}{n}}\right) = F\left(x^{\frac{1}{n}}\right).$$

Consequently, X^n has density

$$f_n(x) = F_n'(x) = \frac{1}{n}x^{\frac{1}{n}-1}F'\left(x^{\frac{1}{n}}\right) = \frac{x^{\frac{1-n}{n}}}{n}f\left(x^{\frac{1}{n}}\right).$$

Example 13.7. Suppose that X has p.d.f. f_X and c.d.f. F_X. Let $Y = X^2$. For $y \geq 0$,

$$\begin{aligned} F_Y(y) &= P(X^2 \leq y) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) \\ &= P\left(-\sqrt{y} < X \leq \sqrt{y}\right) = F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right). \end{aligned}$$

Differentiation gives

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left(f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right).$$

More generally, we have

Theorem 13.2. *If X has density f_X and $g : \mathbb{R} \to \mathbb{R}$ is differentiable and strictly monotone, then $Y = g(X)$ has density*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|, & y \in \text{Range}(g) \\ 0, & y \notin \text{Range}(g) \end{cases}.$$

Proof. We first note that the assumptions ensure that g has a differentiable inverse.

If $y \notin \text{Range}(g)$, then $F_Y(y) = P\left(g(X) \leq y\right)$ is identically 0 or 1 in some neighborhood of y since g is monotone, so $f_Y(y) = F_Y'(y) = 0$.
(This may not hold at the boundary of Range(g) and in fact F_Y may not be differentiable at certain points, but the number of such points is countable and thus may be disregarded.)
Accordingly, we may henceforth restrict our attention to $y \in \text{Range}(g)$.

If g is strictly increasing, then so is g^{-1}, thus

$$F_Y(y) = P\left(g(X) \leq y\right) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right),$$

hence

$$f_Y(y) = F_Y'(y) = F_X'\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|.$$

If g is strictly decreasing, then so is g^{-1}, thus

$$F_Y(y) = P\left(g(X) \leq y\right) = P\left(X \geq g^{-1}(y)\right) = P\left(X > g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y)\right),$$

hence

$$f_Y(y) = F_Y'(y) = -F_X'\left(g^{-1}(y)\right)\frac{d}{dy}g^{-1}(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$$

(as g^{-1} decreasing implies $\frac{d}{dy}g^{-1}(y) < 0$.)

□

Of course, the basic logic of Theorem 13.2 extends to functions g which are only piecewise monotone (such as polynomials) by treating separately the various intervals of monotonicity.

Corollary 13.1. *If X has density $f(x)$, then $aX + b$ has density* $\frac{1}{|a|}f\left(\frac{x-b}{a}\right)$.

14. Common Continuous Distributions

Uniform.

A simple but important continuous random variable is $U \sim \text{Unif}(a, b)$, the uniform distribution on (a, b).

Intuitively, U is chosen uniformly from the interval (a, b), in that the probability that U lies in any subinterval is proportional to the length of the subinterval.

In particular, the distribution of U is invariant under translations (which may "wrap around" (a, b)).

The p.d.f. of U is thus $f_U(u) = \dfrac{1}{b-a} 1_{(a,b)}(u)$.

In general, it suffices to consider $U \sim \text{Unif}(0, 1)$ since $V = (b-a)U + a$ is then uniform on (a, b).

To see this, note that $g(x) = (b-a)x + a$ is strictly monotone with inverse $g^{-1}(y) = \frac{y-a}{b-a}$, so $V = g(U)$ has density

$$f_V(v) = f_U\left(g^{-1}(v)\right) \left| \frac{d}{dv} g^{-1}(v) \right| = 1_{(0,1)}\left(\frac{v-a}{b-a}\right) \frac{1}{b-a} = \frac{1}{b-a} 1_{(a,b)}(v).$$

To compute the mean and variance of $V \sim \text{Unif}(a, b)$, we note that if $U \sim \text{Unif}(0, 1)$, then

$$E[U] = \int_0^1 u du = \frac{1}{2} \text{ and } E[U^2] = \int_0^1 u^2 du = \frac{1}{3},$$

so $\text{Var}(U) = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$.

It follows that

$$E[V] = E\left[(b-a)U + a\right] = (b-a)E[U] + a = a + \frac{b-a}{2} = \frac{a+b}{2},$$

the midpoint of a and b.

Similarly, $\text{Var}(V) = (b-a)^2 \text{Var}(U) = \dfrac{(b-a)^2}{12}$.

Of course these quantities could also be computed directly from the density of V.

Example 14.1. One important use of uniform random variables is to simulate random variables having other continuous distributions.

To see how this works, we first note that $U \sim \text{Unif}(0, 1)$ has c.d.f.

$$F_U(u) = \begin{cases} 0, & u \le 0 \\ \int_0^u dx, & 0 < u < 1 \\ 1, & u \ge 1 \end{cases} = \begin{cases} 0, & u \le 0 \\ u, & 0 < u < 1 \\ 1, & u \ge 1 \end{cases}.$$

Now suppose that suppose that F is the c.d.f. of a continuous random variable and is strictly increasing.

Then $Y = F^{-1}(U)$ has c.d.f. F since for any $y \in \mathbb{R}$,

$$P(Y \leq y) = P\left(F^{-1}(U) \leq y\right) = P\left(U \leq F(y)\right) = F(y).$$

* The condition that F is strictly increasing can be circumvented by defining

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}.$$

The utility of this observation is that one can sample from any continuous distribution by generating a "random" number between 0 and 1, and there are algorithms for doing this pretty well.

Normal.

The normal distribution is extremely important in probability and statistics because of its starring role in the central limit theorem, which says that the c.d.f. of the suitably scaled sum of n i.i.d. random variables (from virtually any distribution) converges to that of a normal random variable as $n \to \infty$.

We say that $W \sim \mathcal{N}(\mu, \sigma^2)$ if its density is given by

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

The graph of f is the "bell curve" that is symmetric about μ.

Our first order of business is to show that f integrates to 1.

As usual, we will be somewhat cavalier in our treatment of improper integrals, but everything is easily justified in the case at hand.

To begin, we make the change of variables $u = \frac{x-\mu}{\sigma\sqrt{2}}$ to get

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

so we need to show that

$$I := \int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\pi}.$$

Since I is clearly nonnegative, this will follow if we can show that $I^2 = \pi$.

Writing an iterated integral as a double integral and changing to polar coordinates gives

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-u^2} du\right)\left(\int_{-\infty}^{\infty} e^{-v^2} dv\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-u^2} e^{-v^2} du dv \\ &= \int\int_{\mathbb{R}^2} e^{-(u^2+v^2)} dA = \int_0^{2\pi}\int_0^{\infty} e^{-r^2} r dr d\theta = 2\pi \int_0^{\infty} r e^{-r^2} dr \\ &= 2\pi \int_0^{\infty} \frac{1}{2} e^{-s} ds = \pi \int_0^{\infty} e^{-s} ds = \pi \end{aligned}$$

where the last line used the change of variables $s = r^2$.

To compute the mean and variance, we note that if $Z \sim \mathcal{N}(0,1)$, then for all $\mu \in \mathbb{R}$, $\sigma > 0$, $X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2)$ since

$$F_X(x) = P(X \leq x) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

hence

$$f_X(x) = F_X'(x) = \Phi'\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where φ and Φ are the p.d.f. and c.d.f., respectively, of Z.

Now

$$E[Z] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ze^{-\frac{z^2}{2}}dz = 0$$

since $ze^{-\frac{z^2}{2}}$ is odd and integrable.

Using integration by parts with $u = z$, $dv = ze^{-\frac{z^2}{2}}$,

$$\begin{aligned} E[Z^2] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2e^{-\frac{z^2}{2}}dz = \frac{1}{\sqrt{2\pi}}\left[-ze^{-\frac{z^2}{2}}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}dz\right] \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}dz = 1 \end{aligned}$$

since $\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} = \varphi(z)$.

Thus Z has mean $E[Z] = 0$ and variance $\mathrm{Var}(Z) = E[Z^2] - E[Z]^2 = 1$, so

$$\begin{aligned} E[X] &= E[\sigma Z + \mu] = \sigma E[Z] + \mu = \mu, \\ \mathrm{Var}(X) &= \mathrm{Var}(\sigma Z + \mu) = \sigma^2\mathrm{Var}(Z) = \sigma^2. \end{aligned}$$

If $Z \sim \mathcal{N}(0,1)$, we say that Z has the standard normal distribution. Typically normal probabilities are computed by "standardizing" to express $X \sim \mathcal{N}(\mu, \sigma^2)$ as $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.

For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then for any $a < b$,

$$\begin{aligned} P(a \le X \le b) = P(a < X \le b) &= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} < Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

where Φ denotes the c.d.f. of $Z \sim \mathcal{N}(0,1)$.

The reason for doing this is that

$$P(a \le X \le b) = \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

does not have a closed form expression in terms of elementary functions, but must be evaluated numerically.

Before the advent of small powerful computers, the way around this problem was to compute a single table of $\Phi(z)$ for Z a standard normal and then transform the question about X to one involving Z.

Because of the symmetry of φ, we have $\Phi(-x) = 1 - \Phi(x)$, so one only has to give a table of "z-scores" for $z \ge 0$.

Usually, $\Phi(z)$ is given for $0 \le z \le 3.5$ (or so) at increments of 0.01. Since $\Phi(z) \ge 0.9998$ for $z \ge 3.5$, this is generally sufficient.

Often the normal distribution is used to approximate binomial probabilities which are difficult to calculate. Basically, if $X \sim \mathrm{Binom}(n,p)$, then $X =_d \sum_{i=1}^n X_i$ where $X_1, ..., X_n$ are i.i.d. Bernoulli(p).

The Central Limit Theorem (which we will discuss later) says that if $Y_1, Y_2, ...$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then the c.d.f. of

$$\frac{\sum_{i=1}^{n} Y_i - n\mu}{\sigma\sqrt{n}}$$

converges pointwise to Φ as $n \to \infty$.

Thus for large n, the c.d.f. of $X \sim \text{Binom}(n,p)$ is approximately equal to that of $W \sim \mathcal{N}(np, np(1-p))$.

To account for the fact that X is discrete, one can apply the "continuity correction"

$$P(X = k) = P\left(k - \frac{1}{2} \leq X \leq k + \frac{1}{2}\right) \approx P\left(k - \frac{1}{2} \leq W \leq k + \frac{1}{2}\right).$$

There are various rules of thumb about when the normal can be used to approximate the binomial.

The *Berry-Esseen Theorem* (which we will not discuss) ensures that if $X \sim \text{Binom}(n,p)$, then

$$\left| P\left(X \leq x\right) - \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)\right| \leq \frac{1 - 2p(1-p)}{2\sqrt{np(1-p)}}$$

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

Exponential and Gamma.

If, for some $\lambda > 0$, X has density $f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$, then X is said to be exponential with rate λ.

The c.d.f. of $X \sim \text{Exponential}(\lambda)$ is given by

$$F(x) = \int_0^x \lambda e^{-\lambda y} dy = -e^{-\lambda y}\Big|_0^x = 1 - e^{-\lambda x}$$

for $x > 0$ and $F(x) = 0$ for $x \leq 0$.

Using integration by parts, we see that for any $n \in \mathbb{N}$,

$$\begin{aligned} E[X^n] &= \int_0^\infty \lambda x^n e^{-\lambda x} dx = -x^n e^{-\lambda x}\Big|_0^\infty + \int_0^\infty n x^{n-1} e^{-\lambda x} dx \\ &= \frac{n}{\lambda} \int_0^\infty \lambda x^{n-1} e^{-\lambda x} dx = \frac{n}{\lambda} E[X^{n-1}]. \end{aligned}$$

In particular, $E[X] = \frac{1}{\lambda}$ and $E[X^2] = \frac{2}{\lambda^2}$, so $\text{Var}(X) = \frac{1}{\lambda^2}$.

A key feature of the exponential distribution is *memorylessness*.

That is, for any $s, t \geq 0$,

$$\begin{aligned} P\left(X > s + t \,|X > s\right) &= \frac{P(X > s+t)}{P(X > s)} = \frac{1 - F(s+t)}{1 - F(s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = 1 - F(t) = P(X > t). \end{aligned}$$

If we think of X as representing the lifetime of some electrical component, say, the memoryless property says that it does not exhibit wear and tear (or toughening): No matter how long it has survived, its chance of surviving t more days is the same as it was when brand new.

In fact, this memorylessness property uniquely characterizes the exponential amongst random variables taking values in $[0, \infty)$.

To see this, define the survival function $G(x) = 1 - F(x) = P(X > x)$. The memorylessness property is equivalent to the statement that $G(s+t) = G(s)G(t)$.

Thus for any $m, n \in \mathbb{N}$, $G\left(\frac{m}{n}\right) = G(\frac{1}{n})^m$. Since $G(\frac{1}{n})^n = G(1)$ by the exact same reasoning, we have $G\left(\frac{m}{n}\right) = G(1)^{\frac{m}{n}}$.

Now, since F is nondecreasing, $G = 1 - F$ is nonincreasing. Thus for any $x \in \mathbb{R}$ if r_n and q_n are sequences of rational numbers with $q_n \leq x \leq r_n$ for all n and $r_n, q_n \to x$, we must have

$$G(1)^{q_n} = G(q_n) \geq G(x) \geq G(r_n) = G(1)^{r_n}$$

for all n, hence, taking $n \to \infty$, $G(x) = G(1)^x$.

Writing $\lambda = -\log(G(1))$ shows that $G(x) = G(1)^x = e^{-\lambda x}$, so $F(x) = 1 - e^{-\lambda x}$.

A similar argument shows that the only memoryless random variables taking values in $\mathbb{N}$ are geometric.

Like geometric random variables, exponentials are good at describing waiting times.

Just as $W \sim$ Geometric(p) represents the time between successive occurrences in a Bernoulli(p) process, $T \sim$ Exponential(λ) gives the time between occurrences in Poisson(λ) process.

Recall that we showed that if an arrivals process has stationary independent increments with the probability of an arrival in a small time interval proportional to its length and the probability of multiple arrivals in a small time interval negligible, then $N(t)$, the number of arrivals between time 0 and time t, has a Poisson(λt) distribution where λ is the proportionality constant in the assumptions.

If T is the time until the first arrival, then for any $t > 0$

$$P(T > t) = P(N(t) = 0) = e^{-\lambda t},$$

hence $T \sim$ Exponential(λ), and the stationary independent increments assumption implies that the interarrival times are i.i.d.

The exponential distribution is a special case of a more general family of probability distributions.

We say that X has the gamma distribution with shape $\alpha > 0$ and scale $\beta > 0$ if its density is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}},\ x \geq 0$$

where the *gamma function* is defined by

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

(The gamma density is sometimes stated in terms of the rate $\lambda = \frac{1}{\beta}$.)

The gamma function is an extension of the factorial function: Integration by parts gives

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx = -x^{t-1}e^{-x}\big|_0^\infty + (t-1)\int_0^\infty x^{t-2}e^{-x}dx = (t-1)\Gamma(t-1).$$

Thus, since $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$, it follows by induction that $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$.

To see that f defines a density, we use the substitution $u = \frac{x}{\beta}$ to get

$$\begin{aligned}\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}}dx &= \frac{1}{\Gamma(\alpha)}\int_0^\infty \frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}e^{-\frac{x}{\beta}}dx \\ &= \frac{1}{\Gamma(\alpha)}\int_0^\infty u^{\alpha-1}e^{-u}du = \frac{\Gamma(\alpha)}{\Gamma(\alpha)} = 1.\end{aligned}$$

If $X \sim \text{Gamma}(\alpha, \beta)$, then

$$\begin{aligned}E[X^n] &= \int_0^\infty x^n \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}}dx = \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{(\alpha+n)-1}e^{-\frac{x}{\beta}}dx \\ &= \frac{\Gamma(\alpha+n)\beta^{\alpha+n}}{\Gamma(\alpha)\beta^\alpha}\cdot\frac{1}{\Gamma(\alpha+n)\beta^{\alpha+n}}\int_0^\infty x^{(\alpha+n)-1}e^{-\frac{x}{\beta}}dx \\ &= \frac{\Gamma(\alpha+n)\beta^{\alpha+n}}{\Gamma(\alpha)\beta^\alpha} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\beta^n.\end{aligned}$$

Thus

$$E[X] = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\beta = \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)}\beta = \alpha\beta$$

and

$$E[X^2] = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}\beta^2 = \frac{(\alpha+1)\alpha\Gamma(\alpha)}{\Gamma(\alpha)}\beta^2,$$

so

$$\text{Var}(X) = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

When $\alpha = 1$ and $\beta = \frac{1}{\lambda}$, we recover the Exponential(λ) distribution.

When $\alpha = \frac{n}{2}$ and $\beta = 2$, we get the Chi-Squared distribution with n degrees of freedom.

The latter is the distribution of a sum of squares of n independent standard normals and is used often in statistics for goodness of fit tests.

It is also related to Student's t-distribution for estimating means when the variance is unknown
(If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(n)$ are independent, then $\frac{Z}{\sqrt{\frac{V}{n}}} \sim t(n-1)$)
and Snedcor's F-distribution used in comparing variances
(If $U \sim \chi^2(n-1)$ and $W \sim \chi^2(m-1)$ are independent, then $\frac{\left(\frac{U}{n-1}\right)}{\left(\frac{W}{m-1}\right)} \sim F(n-1, m-1)$).

Ultimately, this is because the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ from a $\mathcal{N}(\mu, \sigma^2)$ population satisfies $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ and is independent of the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

15. Joint Distributions

Thus far we have only focused on one random variable at a time. However, we often want to make statements about several random variables.

As is done in a course on multivariate calculus, we will primarily concentrate on statements about two random variables as this captures most of the essential ideas without being too notationally cumbersome. It is straightforward to generalize all concepts we will discuss to any finite number of random variables.

Perhaps the first thing to observe is that if X_1 and X_2 are $\mathbb{R}$-valued random variables on $(\Omega, \mathcal{F}, P)$, then statements concerning both X_1 and X_2 are equivalent to statements about the *random vector*

$$X = (X_1, X_2)^T \in \mathbb{R}^2.$$

Thus, broadly speaking, the only thing that has changed is the codomain in our definition of random variables.

Also, statements about random variables can be regarded as statements about events and we already have tools for describing probabilities of composite events.

For example,

$$P(X_1 \in A, X_2 \in B) = P\left(\{X_1 \in A\} \cap \{X_2 \in B\}\right) = P\left(\{\omega \in \Omega : X_1(\omega) \in A \text{ and } X_2(\omega) \in B\}\right).$$

More concretely, suppose that the random experiment consists of tossing a fair coin n times
(so that $\Omega = \{0,1\}^n$, $\mathcal{F} = 2^\Omega$, and $P(\{\omega\}) = \frac{1}{2^n}$ for all $\omega \in \Omega$).
Let X_1 denote the total number of heads and let X_2 be 1 if the first outcome is heads and 0 otherwise.
Then

$$P(X_1 = 4, X_2 = 1) = \frac{1}{2^n}\binom{n-1}{3}$$

since the sequences which contain exactly four heads and begin with a head can be enumerated by specifying which 3 of the last $n-1$ tosses resulted in heads.

In general, it is more convenient to work with multivariate analogs of c.d.f.s, p.m.f.s, and p.d.f.s.

For any random variables X and Y, we define the *joint c.d.f.* of X and Y by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y),\ x, y \in \mathbb{R}.$$

Note that the c.d.f. of X can be recovered from the joint c.d.f. since for any $x \in \mathbb{R}$,

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(X \leq x, Y < \infty) \\ &= P\left(\lim_{y\to\infty} \{X \leq x, Y \leq y\}\right) \\ &= \lim_{y\to\infty} P\left(X \leq x, Y \leq y\right) = \lim_{y\to\infty} F_{X,Y}(x,y) \end{aligned}$$

because $\{X \leq x, Y \leq y\} \nearrow \{X \leq x, Y < \infty\}$ as $y \nearrow \infty$.

The exact same argument shows that $F_Y(y) = \lim_{x\to\infty} F_{X,Y}(x,y)$.

In theory, all statements about X and Y can be answered in terms of $F_{X,Y}$.

For example,

$$\begin{aligned} P(X > x, Y > y) &= 1 - P\left(\{X > x, Y > y\}^C\right) \\ &= 1 - P\left(\{X > x\}^C \cup \{Y > y\}^C\right) \\ &= 1 - P\left(\{X \leq x\} \cup \{Y \leq y\}\right) \\ &= 1 - [P(X \leq x) + P(Y \leq y) - P(X \leq x, Y \leq y)] \\ &= 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y). \end{aligned}$$

If X and Y are discrete, so that $\{x \in \mathbb{R} : P(X = x) > 0\} = \{x_i\}_{i \in I}$ and $\{y \in \mathbb{R} : P(Y = y) > 0\} = \{y_j\}_{j \in J}$ with I, J countable, then $\left\{(x, y) \in \mathbb{R}^2 : P(X = x, Y = y) > 0\right\} \subseteq \{(x_i, y_j)\}_{(i,j) \in I \times J}$ is countable.

In this case, we define the *joint p.m.f.* of X and Y by

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

You have already worked with joint p.m.f.s in the homework exercise where you proved that $E[X + Y] = E[X] + E[Y]$ when X and Y are discrete.

Note that

$$F_{X,Y}(a, b) = P(X \leq a, Y \leq b) = \sum_{\substack{(x,y): x \leq a, y \leq b \\ p_{X,Y}(x,y) > 0}} p_{X,Y}(x, y).$$

Also,

$$\begin{aligned} p_X(x) = P(X = x) &= P\left(\bigcup_{y: p_Y(y) > 0} \{X = x, Y = y\}\right) \\ &= \sum_{y: p_Y(y) > 0} P(X = x, Y = y) = \sum_{y: p_Y(y) > 0} p_{X,Y}(x, y) \end{aligned}$$

and similarly for $p_Y(y)$.

We say that p_X and p_Y are the *marginal mass functions* for X and Y, respectively.

Finally, we say that X and Y are *jointly (absolutely) continuous* if there exists a function $f : \mathbb{R}^2 \to [0, \infty)$ such that for all (measurable) $C \subseteq \mathbb{R}^2$,

$$P\left((X, Y) \in C\right) = \int \int_C f(x, y) dx dy.$$

$f = f_{X,Y}$ is called the *joint density* of X and Y.

If $A, B \subseteq \mathbb{R}$, setting $C = A \times B = \{(x, y) \in \mathbb{R}^2 : x \in A, y \in B\}$ in the above equation and noting that $f \geq 0$ shows that

$$P(X \in A, Y \in B) = \int_B \int_A f(x, y) dx dy.$$

In particular, if $A = (-\infty, a]$ and $B = (-\infty, b]$, then we have

$$F_{X,Y}(a,b) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy.$$

Accordingly,

$$\begin{aligned}\frac{\partial^2}{\partial a \partial b}F_{X,Y}(a,b) &= \frac{\partial^2}{\partial a \partial b}\int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy \\ &= \frac{\partial}{\partial a}\left[\frac{\partial}{\partial b}\int_{-\infty}^{b}\left(\int_{-\infty}^{a} f(x,y)dx\right)dy\right] = \frac{\partial}{\partial a}\int_{-\infty}^{a} f(x,b)dx = f(a,b).\end{aligned}$$

As in the discrete case, the marginal density of X, say, can be recovered from the joint density of X and Y by integrating out the y term:

$$P(X \in A) = P(X \in A, Y \in \mathbb{R}) = \int_A \int_{-\infty}^{\infty} f(x,y)dydx = \int_A \left(\int_{-\infty}^{\infty} f(x,y)dy\right)dx$$

for all A, so

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy,$$

and similarly for f_Y.

This shows that if X and Y are jointly continuous, then X and Y are each continuous. Taking $X = Y$ shows that the converse need not hold.

(It is important to note that we are using continuous to mean absolutely continuous here. The statement is not true using the real definition of continuous.)

Of course, all of this applies to more than two random variables as well.

For example, the joint distribution function of $X_1, ..., X_n$ is given by $F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$, and we say that $X_1, ..., X_n$ are jointly (absolutely) continuous if there is a function $f : \mathbb{R}^n \to [0, \infty)$ with $P((X_1, ..., X_n) \in C) = \int \cdots \int_C f(x_1, ..., x_n)dx_1 \cdots dx_n$.

Example 15.1. Suppose that a point is chosen uniformly at random from a disc of radius R centered at the origin. Letting X and Y denote the x and y coordinates of the point, we see that the joint density function of X and Y has the form

$$f(x,y) = \begin{cases} c, & x^2 + y^2 \leq R \\ 0, & \text{otherwise} \end{cases}$$

for some $c > 0$.

(1) Determine c.
(2) Find the marginal densities of X and Y.
(3) Compute the probability that D, the distance to the origin, is less than or equal to a.
(4) Find $E[D]$.

Since

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = c\iint_{x^2+y^2\leq R^2} dxdxy,$$

we have that $c = \frac{1}{\pi R^2}$ (as the rightmost integral gives the area of the disc $\left\{(x,y) : x^2 + y^2 \leq R^2\right\}$).

The marginal density of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\pi R^2}\int_{x^2+y^2\le R^2} dy = \frac{1}{\pi R^2}\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dy = \frac{2}{\pi R^2}\sqrt{R^2-x^2},\ -R \le x \le R.$$

since for any fixed x, $f(x,y) = 0$ if $|y| > \sqrt{R^2-x^2}$.

By symmetry, Y has density $f_Y(y) = \frac{2}{\pi R^2}\sqrt{R^2-y^2}$, $-R \le y \le R$.

To compute the distribution function of $D = \sqrt{X^2+Y^2}$, we note that if $0 \le a \le R$, then

$$\begin{aligned} F_D(a) &= P\left(\sqrt{X^2+Y^2} \le a\right) = P\left(X^2+Y^2 \le a^2\right) \\ &= \iint_{x^2+y^2\le a} f(x,y)dxdy = \frac{1}{\pi R^2}\iint_{x^2+y^2\le a} dxdy \\ &= \frac{\pi a^2}{\pi R^2} = \frac{a^2}{R^2}. \end{aligned}$$

Of course, $F_D(a) = 1$ if $a > R$ and $F_D(a) = 0$ if $a < 0$.

D has density $f_D(a) = \dfrac{d}{da}\dfrac{a^2}{R^2} = \dfrac{2a}{R^2}$, $0 \le a \le R$, so

$$E[D] = \int_0^R a \cdot \frac{2a}{R^2}da = \frac{2}{R^2}\int_0^R a^2 da = \frac{2}{R^2}\cdot\frac{R^3}{3} = \frac{2}{3}R.$$

* As you might expect at this point, if $g : \mathbb{R}^2 \to \mathbb{R}$ is a function and (X,Y) has joint density f, then $E[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y)f(x,y)dxdy$ whenever the expectation exists.

Thus an alternative solution is

$$\begin{aligned} E[D] &= \iint_{\mathbb{R}^2} \sqrt{x^2+y^2}f(x,y)dxdy = \frac{1}{\pi R^2}\iint_{x^2+y^2\le R^2} \sqrt{x^2+y^2}dxdy \\ &= \frac{1}{\pi R^2}\int_0^{2\pi}\int_0^R r\cdot r dr d\theta = \frac{2}{R^2}\int_0^R r^2 dr = \frac{2}{3}R. \end{aligned}$$

Example 15.2. Suppose that X and Y have joint density $f(x,y) = e^{-(x+y)}1\{x>0, y>0\}$. Find the density of $W = X/Y$.

For $a > 0$,

$$\begin{aligned} F_W(a) &= P\left(\frac{X}{Y} \le a\right) = \iint_{x\le ay} f(x,y)dxdy = \int_0^\infty\int_0^{ay} e^{-(x+y)}dxdy \\ &= \int_0^\infty e^{-y}\int_0^{ay} e^{-x}dxdy = \int_0^\infty e^{-y}(1-e^{-ay})dy \\ &= \int_0^\infty e^{-y} - e^{-(a+1)y}dy = -e^{-y} + \frac{1}{a+1}e^{-(a+1)y}\Big|_0^\infty = 1 - \frac{1}{a+1}. \end{aligned}$$

Thus W has density $f_W(a) = \dfrac{d}{da}\left(1-\dfrac{1}{a+1}\right) = \dfrac{1}{(a+1)^2}$, $a > 0$.

16. Independent Random Variables

Recall that events E and F are said to be independent if $P(E \cap F) = P(E)P(F)$.

We say that random variables X and Y are independent if for all sets $A, B \subseteq \mathbb{R}$, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent. That is,

Definition. X and Y are independent if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all $A, B \subseteq \mathbb{R}$.

Thus if X and Y are independent, we have

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b) = F_X(a)F_Y(b).$$

In fact, this is an equivalent characterization of independence:

Theorem 16.1. *X and Y are independent if and only if $F_{X,Y}(a,b) = F_X(a)F_Y(b)$ for all $a, b \in \mathbb{R}$.*

The proof that the factoring condition implies independence requires a technical result known as the $\pi - \lambda$ theorem which is beyond the scope of this class, so we'll just accept it as a known fact.

(Basically, it's the same reason that c.d.f.s completely determine distributions: they tell you what happens on sets of the form $(-\infty, a]$ and all other sets of interest can be built out of these.)

In the case where X and Y are discrete with joint p.m.f. $p_{X,Y}$ and marginals p_X, p_Y, if X and Y are independent, then for all x, y,

$$p_{X,Y}(x,y) = P(X = x, Y = y) = P(X = x)P(Y = y) = p_X(x)p_Y(y).$$

Conversely, if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$, then for any $A, B \subseteq \mathbb{R}$,

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{\substack{(x,y) \in A \times B: \\ p_{X,Y}(x,y) > 0}} p_{X,Y}(x,y) = \sum_{\substack{(x,y) \in A \times B: \\ p_X(x), p_Y(y) > 0}} p_X(x)p_Y(y) \\ &= \sum_{\substack{x \in A: \\ p_X(x) > 0}} \sum_{\substack{y \in B: \\ p_Y(y) > 0}} p_X(x)p_Y(y) = \sum_{\substack{x \in A: \\ p_X(x) > 0}} p_X(x) \sum_{\substack{y \in B: \\ p_Y(y) > 0}} p_Y(y) = P(X \in A)P(Y \in B), \end{aligned}$$

so X and Y are independent.

We record this observation as

Theorem 16.2. *Discrete random variables X and Y are independent if and only if the joint p.m.f. factors as the product of the marginal p.m.f.s.*

Similarly, we have

Theorem 16.3. *Jointly continuous random variables X and Y are independent if and only if the joint p.d.f. factors as the product of the marginal p.d.f.s.*

Proof. Denoting joint/marginal c.d.f.s/p.d.f.s in the usual fashion, it follows from Theorem 16.1 that X and Y are independent if and only if

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_X(x)F_Y(y) = f_X(x)f_Y(y). \qquad \square$$

A more convenient way to check independence is given by

Theorem 16.4. *Jointly continuous (respectively, discrete) random variables X and Y are independent if and only if their joint density (respectively, mass function) can be written as $f_{X,Y}(x,y) = g(x)h(y)$.*

Proof.

We will only prove the continuous case as the discrete case is completely analogous.

If X and Y are independent, then Theorem 16.3 shows that $f_{X,Y}(x,y) = g(x)h(y)$ where $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.

Now suppose that $f_{X,Y}(x,y) = g(x)h(y)$ for some g, h. Then

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)dxdy \\ &= \left(\int_{-\infty}^{\infty} g(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)dy\right) = C_1C_2 \end{aligned}$$

where $C_1 = \int_{-\infty}^{\infty} g(x)dx$ and $C_2 = \int_{-\infty}^{\infty} h(y)dy$.

Also,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} g(x)h(y)dy = g(x)C_2, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_{-\infty}^{\infty} g(x)h(y)dx = h(y)C_1, \end{aligned}$$

so, since $C_1C_2 = 1$, we have

$$f_{X,Y}(x,y) = g(x)h(y) = g(x)C_2h(y)C_1 = f_X(x)f_Y(y),$$

which shows that X and Y are independent.

□

Example 16.1. Suppose that the number of people entering the post office on a given day is a Poisson random variable with rate λ, and that each person who enters is a male with probability p and a female with probability $1-p$. Let X denote the number of males who enter and Y the number of females. What is the joint distribution of X and Y?

For any $i, j \in \mathbb{N}_0$, we have

$$P(X = i, Y = j) = P(X = i, Y = j \,|X + Y = i + y)P(X + Y = i + j)$$

since $P(X = i, Y = j \,|X + Y \neq i + y) = 0$.

By assumption,

$$P(X + Y = i + j) = e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!}$$

and

$$P(X = i, Y = j \,|X + Y = i + y) = \binom{i+j}{i} p^i(1-p)^j,$$

so

$$\begin{aligned}
P(X=i,Y=j) &= P(X=i,Y=j\,|X+Y=i+y)P(X+Y=i+j) \\
&= \binom{i+j}{i} p^i(1-p)^j e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!} \\
&= \frac{(i+j)!}{i!j!}p^i(1-p)^j e^{-p\lambda}e^{-(1-p)\lambda}\frac{\lambda^i\lambda^j}{(i+j)!} \\
&= e^{-p\lambda}\frac{(p\lambda)^i}{i!}\cdot e^{-(1-p)\lambda}\frac{[(1-p)\lambda]^j}{j!}.
\end{aligned}$$

Thus X and Y are independent Poissons with rates $p\lambda$ and $(1-p)\lambda$, respectively.

Example 16.2. A man and woman decide to meet at a certain location. If each independently arrives at a time uniformly distributed between 12:00 and 1:00, what is the probability that the first to arrive has to wait at least 10 minutes for the other.

Let X and Y denote the number of minutes past 12:00 that the man and woman arrive.

Then X and Y are independent $U(0,60)$ random variables and the desired probability is

$$P(X+10<Y)+P(Y+10<X) = 2P(X+10<Y)$$

by symmetry.

Since

$$\begin{aligned}
P(X+10<Y) &= \int\int_{x+10<y} f_{X,Y}(x,y)dydx = \int\int_{x+10<y} f_X(x)f_Y(y)dydx \\
&= \frac{1}{60^2}\int_0^{50}\int_{x+10}^{60} dydx = \frac{1}{60^2}\int_0^{50}(50-x)dx \\
&= \frac{1}{60^2}\left(50x - \frac{x^2}{2}\Big|_0^{50}\right) = \left(\frac{1}{60}\right)^2\cdot\frac{50^2}{2} = \frac{1}{2}\cdot\frac{25}{36},
\end{aligned}$$

the probability that one has to wait at least 10 minutes is $2P(X+10<Y) = \frac{25}{36}$.

Our next example is known as *Buffon's needle* (after its originator), and is the earliest problem in geometric probability. It provides an interesting way to approximate π using Monte Carlo methods.

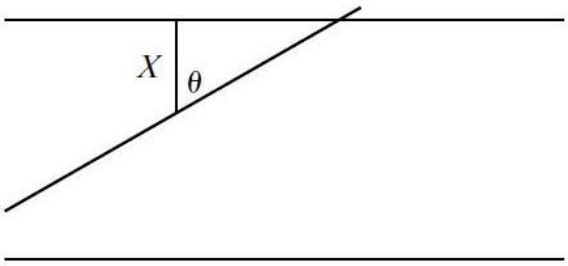
Example 16.3. Suppose we have a floor made of parallel strips of wood, each of width D, and we drop a needle of length $L \le D$ onto the floor. What is the probability that the needle will lie across a line between two strips?

We can represent the position of the needle by specifying

$$X = \text{distance from midpoint of needle to nearest parallel line}$$

and

$$\theta = \text{acute angle between needle and line connecting midpoint to nearest parallel line.}$$



The needle will intersect a line if $\frac{X}{\cos(\theta)}$, the hypotenuse of the triangle formed from the nearest line, the perpendicular line running through the needle's midpoint, and the line on which the needle lies, is less than $\frac{L}{2}$.

Now, from the problem statement it is reasonable to assume that X is uniform on $\left[0, \frac{D}{2}\right]$ and θ is uniform on $\left[0, \frac{\pi}{2}\right]$.

Accordingly,

$$\begin{aligned} P(\text{needle crosses a line}) &= P\left(\frac{X}{\cos(\theta)} < \frac{L}{2}\right) = \iint_{x<\frac{L}{2}\cos(y)} f_{X,\theta}(x,y)dxdy \\ &= \int_0^{\frac{\pi}{2}} \int_0^{\frac{L}{2}\cos(y)} \frac{2}{D} \cdot \frac{2}{\pi} dxdy = \frac{4}{\pi D} \int_0^{\frac{\pi}{2}} \frac{L}{2}\cos(y)dy \\ &= \frac{2L}{\pi D}\left(L\sin(y)|_0^{\frac{\pi}{2}}\right) = \frac{2L}{\pi D}. \end{aligned}$$

One can estimate π by repeatedly dropping a needle of length 1 on a table painted with parallel lines at distance 2 from one another.

If the needle crosses a line on n out of N trials for N large, then one expects that

$$\frac{n}{N} \approx P(\text{needle crosses a line}) = \frac{1}{\pi},$$

so

$$\pi \approx \frac{N}{n}.$$

Since the needle either crosses the line once or not at all, the above analysis shows that the expected number of crossings is $\frac{2L}{\pi D}$.

If we don't require the needle to be straight, but instead allow any rigid plane curve of length L (Buffon's noodle), then the expected number of crossings turns out to be the same.

To see that this is so, we approximate the curve with a sequence of piecewise linear curves C_i and let $X_{i,j}$ denote the number of line crossings by the j^{th} segment of C_i (whose length is denoted $L_{i,j}$). The expected number of line crossings of C_n is

$$E[X_{n,1} + ... + X_{n,n}] = E[X_{n,1}] + ... + E[X_{n,n}] = \sum_{j=1}^{n} \frac{2L_{n,j}}{\pi D} = \frac{2}{\pi D} \sum_{i=1}^{n} L_{n,j}.$$

Sending n to infinity gives $E[X] = \frac{2L}{\pi D}$.

(As long as we use approximations with each $L_{i,j} \leq D$ in the above argument, we see that the requirement $L \leq D$ is unnecessary)

17. Sums and Convolution

If X and Y are discrete random variables, then $X+Y$ is discrete with p.m.f.

$$p_{X+Y}(k) = \sum_{\substack{(i,j):i+j=k,\\ p_{X,Y}(i,j)>0}} p_{X,Y}(i,j) = \sum_{j:p_{X,Y}(k-j,j)>0} p_{X,Y}(k-j,j).$$

The continuous analog takes a little more work in general, but there is a nice answer if the variates are independent.

We begin with some facts from analysis.

Definition. The *convolution* of functions f and g is defined as

$$(f*g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy$$

provided that the integral exists.

If f and g are both integrable - as is the case with two density functions - then one can show that the integral defining $(f*g)(x)$ exists (for a.e. x).

Also, note that the substitution $u = x-y$ shows that

$$(f*g)(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy = -\int_{\infty}^{-\infty} f(u)g(x-u)du = \int_{-\infty}^{\infty} g(x-u)f(u)du = (g*f)(x).$$

Finally, it can be shown that $f*g$ is differentiable with

$$(f*g)'(t) = (f'*g)(t) = (f*g')(t).$$

Now if X and Y are independent with densities f_X and f_Y, the c.d.f. of $X+Y$ is given by

$$\begin{aligned} F_{X+Y}(a) = P(X+Y \le a) &= \iint_{x+y\le a} f_X(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} f_X(x)dx\right) f_Y(y)dy \\ &= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = (F_X * f_Y)(a). \end{aligned}$$

Of course, we could have integrated with respect to y first to get $F_{X+Y} = F_Y * f_X = f_X * F_Y$ instead.

Differentiating shows that $X+Y$ has density

$$f_{X+Y}(a) = F'_{X+Y}(a) = (F_X * f_Y)'(a) = (F'_X * f_Y)(a) = (f_X * f_Y)(a).$$

In the discrete case, the independence assumption gives

$$p_{X+Y}(k) = \sum_{j:p_{X,Y}(k-j,j)>0} p_{X,Y}(k-j,j) = \sum_{j:p_X(k-j),p_Y(j)>0} p_X(k-j)p_Y(j),$$

which is also a kind of convolution.

Summing over $k \le l$ gives

$$F_{X+Y}(l) = \sum_{k\le l}\sum_j p_Y(k-j)p_X(j) = \sum_j \sum_{k\le l-j} p_X(k)p_X(j) = \sum_i F_X(l-j)p_Y(j).$$

Example 17.1. If $X \sim \text{Gamma}(\alpha_1, \beta)$ and $Y \sim \text{Gamma}(\alpha_2, \beta)$ are independent, then $X + Y$ has density

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}\Gamma(\alpha_2)\beta^{\alpha_2}} \int_0^z (z-y)^{\alpha_1-1} y^{\alpha_2-1} e^{-\beta(z-y)} e^{-\beta y} dy \\ &= \frac{e^{-\beta z}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \int_0^z (z-y)^{\alpha_1-1} y^{\alpha_2-1} dy = \frac{e^{-\beta z} z^{\alpha_1+\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \int_0^1 (1-x)^{\alpha_1-1} x^{\alpha_2-1} dx \end{aligned}$$

where the final equality uses the substitution $x = \dfrac{y}{z}$.

Since the last integral is just some constant, we can write

$$f_{X+Y}(z) = C z^{\alpha_1+\alpha_2-1} e^{-\beta z}.$$

Recognizing that $z^{\alpha_1+\alpha_2-1} e^{-\beta z}$ is the kernel of the Gamma$(\alpha_1 + \alpha_2, \beta)$ distribution and noting that f_{X+Y} is a density supported on $[0, \infty)$, we must have $C = \dfrac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}}$, hence $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)$.

(If we had discussed the beta family in class, we would have known that $B \sim \text{Beta}(\alpha_2, \alpha_1)$ has density

$$f_B(x) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_2-1}(1-x)^{\alpha_1-1}, \ 0 \le x \le 1,$$

which would also allow us to derive the constant.)

It follows by induction that if $X_1, ..., X_n$ are independent with $X_i \sim \text{Gamma}(\alpha_i, \beta)$, then

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right).$$

In particular, if $X_1, ..., X_n$ are i.i.d. Exponential$(\lambda) = \text{Gamma}\left(1, \frac{1}{\lambda}\right)$, then

$$\sum_{i=1}^n X_i \sim \text{Gamma}\left(n, \frac{1}{\lambda}\right) = \text{Erlang}(n, \lambda).$$

Similarly, if $Y_1, ..., Y_n$ are independent with $Y_i \sim \chi^2(d_i) = \text{Gamma}\left(\frac{d_i}{2}, 2\right)$, then

$$\sum_{i=1}^n Y_i \sim \text{Gamma}\left(\frac{1}{2}\sum_{i=1}^n d_i, 2\right) = \chi^2\left(\sum_{i=1}^n d_i\right).$$

Now if $Z_1, ..., Z_n$ are i.i.d. $\mathcal{N}(0,1)$, then Z_i^2 has c.d.f.

$$\begin{aligned} F_{Z_i^2}(z) &= P(Z_i^2 \le z) = P\left(-\sqrt{z} \le Z_i \le \sqrt{z}\right) \\ &= P\left(Z_i \le \sqrt{z}\right) - P\left(Z_i \le -\sqrt{z}\right) = \Phi\left(\sqrt{z}\right) - \Phi\left(-\sqrt{z}\right), \ z \ge 0, \end{aligned}$$

so its density is

$$\begin{aligned} f_{Z_i^2}(z) = F'_{Z_i^2}(z) &= \frac{1}{2\sqrt{z}}\left(\varphi\left(\sqrt{z}\right) + \varphi\left(-\sqrt{z}\right)\right) = \frac{1}{2\sqrt{z}}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{z}{2}}\right) \\ &= \frac{1}{\sqrt{\pi} 2^{\frac{1}{2}}} z^{-\frac{1}{2}} e^{-\frac{z}{2}} = \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{\frac{1}{2}}} z^{\frac{1}{2}-1} e^{-\frac{1}{2}z}, \ z \ge 0, \end{aligned}$$

hence $Z_1^2, ..., Z_i^2$ are independent Gamma$(\frac{1}{2}, 2) = \chi^2(1)$ random variables.

It follows that $\displaystyle\sum_{i=1}^n Z_i^2 \sim \chi^2(n)$.

Example 17.2. Suppose that $X_1 \sim \mathcal{N}(0,1)$ and $X_2 \sim \mathcal{N}(0,\sigma^2)$ are independent.

To compute the density of $X_1 + X_2$, we first note that if $c = \frac{\sigma^2+1}{\sigma^2}$, then

$$\begin{aligned}(x-y)^2 + \frac{y^2}{\sigma^2} &= x^2 - 2xy + cy^2 = x^2 + c\left[\left(y - \frac{x}{c}\right)^2 - \frac{x^2}{c^2}\right] \\ &= \left(1 - \frac{1}{c}\right)x^2 + c\left(y - \frac{x}{c}\right)^2 = \frac{x^2}{\sigma^2+1} + c\left(y - \frac{x}{c}\right)^2,\end{aligned}$$

so

$$\begin{aligned}f_{X_1+X_2}(x) &= \int_{-\infty}^{\infty} f_{X_1}(x-y) f_{X_2}(y) dy = \frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{2}} e^{-\frac{y^2}{2\sigma^2}} dy \\ &= \frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left[(x-y)^2 + \frac{y^2}{\sigma^2}\right]\right) dy \\ &= \frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left[\frac{x^2}{\sigma^2+1} + c\left(y - \frac{x}{c}\right)^2\right]\right) dy \\ &= \exp\left(-\frac{x^2}{2(\sigma^2+1)}\right)\frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{c}{2}\left(y - \frac{x}{c}\right)^2\right) dy \\ &= \exp\left(-\frac{x^2}{2(\sigma^2+1)}\right)\frac{1}{2\pi\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{c}{2}z^2\right) dz \\ &= C\exp\left(-\frac{x^2}{2(\sigma^2+1)}\right)\end{aligned}$$

for some constant C.

Since $f_{X_1+X_2}$ must integrate to 1, it follows that $C = \dfrac{1}{\sqrt{2\pi(\sigma^2+1)}}$, so $X_1 + X_2 \sim \mathcal{N}(0, 1+\sigma^2)$.

Now suppose that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent.

Then $\frac{X_1-\mu_1}{\sigma_1} \sim \mathcal{N}(0,1)$ and $\frac{X_2-\mu_2}{\sigma_1} \sim \mathcal{N}\left(0, \frac{\sigma_2^2}{\sigma_1^2}\right)$, so

$$X_1 + X_2 = \sigma_1\left(\frac{X_1-\mu_1}{\sigma_1} + \frac{X_2-\mu_2}{\sigma_1}\right) + \mu_1 + \mu_2 = \sigma_1 W + (\mu_1 + \mu_2)$$

where $W \sim \mathcal{N}\left(0, 1 + \frac{\sigma_2^2}{\sigma_1^2}\right)$.

Therefore, $X_1 + X_2$ is normal with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2\left(1 + \frac{\sigma_2^2}{\sigma_1^2}\right) = \sigma_1^2 + \sigma_2^2$.

An induction argument shows that if $X_1, ..., X_n$ are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

We conclude with an easy example of discrete convolution for which we don't even need to appeal to a general formula.

Example 17.3. If $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ are independent, then for any $n \in \mathbb{N}_0$,

$$\begin{aligned} P(X+Y=n) &= \sum_{k=0}^{n} P(X=k)P(Y=n-k) = \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!}e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!} \\ &= e^{-(\lambda_1+\lambda_2)}\frac{1}{n!}\sum_{k=0}^{n}\frac{n!}{k!(n-k)!}\lambda_1^k\lambda_2^{n-k} = e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^n}{n!} \end{aligned}$$

by the binomial theorem. It follows that $X+Y \sim \text{Poisson}(\lambda_1+\lambda_2)$, and in general, the sum of n independent Poissons is a Poisson with mean equal to the sum of the constituent means.

18. Functions of Several Random Variables

Recall that if X is a random variable with density f_X and $g : \mathbb{R} \to \mathbb{R}$ is strictly monotone, then $Y = g(X)$ has density $f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$ for all $y \in \text{Range}(g)$.
This means that for all $A \subseteq \text{Range}(g)$,

$$\int_A f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|dy = \int_A f_Y(y)dy = P(g(X) \in A) = \int_{g^{-1}(A)} f_X(x)dx.$$

When $A = (a, b)$, this is just taking $u = g^{-1}$ in the change of variables formula

$$\int_a^b f_X\left(u(x)\right)u'(x)dx = \int_{u(a)}^{u(b)} f_X(u)du.$$

The absolute value is because the integral is signed in the sense that $\int_a^b f = -\int_b^a f$, and if u is decreasing, then $u\left((a, b)\right) = (u(b), u(a))$.
The inverse function theorem says that $u = g^{-1}$ exists and is continuously differentiable if g is continuously differentiable with nonzero derivative. Such a condition is needed in order to integrate $(f \circ u)\, u'$, hence the assumption that g is differentiable and strictly monotone.

The change of variables formula gives an alternate proof of Theorem 13.2 since it describes the probability that $g(X)$ belongs to any interval and this determines its distribution.

For functions of several variables, the change of variables formula says that if $A \subseteq \mathbb{R}^n$ and $u : \mathbb{R}^n \to \mathbb{R}^n$ are sufficiently nice (e.g. A is open and u is one-to-one and continuously differentiable with a continuously differentiable inverse), then for any integrable function $f : \mathbb{R}^n \to \mathbb{R}$,

$$\int_A f\left(u(x)\right)\left|\det\left(J_u(x)\right)\right|dx = \int_{u(A)} f(u)du$$

where

$$J_u\left(x\right) = \begin{bmatrix} \frac{\partial u_1}{\partial x_1} & \cdots & \frac{\partial u_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial u_n}{\partial x_1} & \cdots & \frac{\partial u_n}{\partial x_n} \end{bmatrix}$$

is the *Jacobian* matrix of $u(x) = (u_1(x_1, ..., x_n), \ldots, u_n(x_1, ..., x_n))^T$.
The inverse function theorem says that u^{-1} exists and is continuously differentiable at x if $u \in C^1$ and $\det\left(J_u(x)\right) \neq 0$. Moreover, $J_{u^{-1}}\left(u\left(x\right)\right) = J_u(x)^{-1}$. (This is the multivariate version of $\dfrac{d}{dx}f^{-1}(x) = \dfrac{1}{f'(f^{-1}(x))}$.)

Example 18.1. Consider the transformation $u(r, \theta) = (r\cos(\theta), r\sin(\theta)) := (x, y)$.
The Jacobian determinant is

$$\det\left(J_u(r, \theta)\right) = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{vmatrix} = r\cos^2(\theta) + r\sin^2(\theta) = r,$$

so the change of variables formula gives

$$\iint_A f(x, y)dxdy = \iint_{u^{-1}(A)} f\left(u(r, \theta)\right)\left|\det\left(J_u(r, \theta)\right)\right|drd\theta = \iint_{u^{-1}(A)} f(u(r, \theta))rdrd\theta,$$

where $u^{-1}(x, y) = \left(x^2 + y^2, \tan^{-1}\left(\frac{y}{x}\right)\right)$.

This is the familiar rule for changing to polar coordinates!

As in the univariate case, the change of variable formula shows how to define the density of certain functions of continuous random variables.

Theorem 18.1. *If $X_1, ..., X_n$ are random variables with joint density $f_{X_1,...,X_n}$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable with $\det(J_g(x_1, ..., x_n)) \neq 0$ for all $(x_1, ..., x_n)^T \in \mathbb{R}^n$, then the random variables $Y_1, ..., Y_n$ defined by $Y_i = g_i(X_1, ..., X_n)$ have joint density*

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = f_{X_1,...,X_n}\left(g^{-1}(y_1, ..., y_n)\right) \left|\det\left(J_{g^{-1}}(y_1, ..., y_n)\right)\right|, \quad (y_1, ..., y_n) \in \text{Range}(g).$$

Specializing to the two dimensional case and using slightly more concrete language, we have

Corollary 18.1. *Suppose that X and Y are random variables with joint density $f_{X,Y}$ and define $U = g_1(X, Y)$, $V = g_2(X, Y)$ for some $g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}$. Suppose that the transformation $g : (x, y) \mapsto (g_1(x, y), g_2(x, y))$ is invertible in the sense that there exist $h_1, h_2 : \mathbb{R}^2 \to \mathbb{R}$ such that if $u = g_1(x, y)$ and $v = g_2(x, y)$, then $x = h_1(u, v)$ and $y = h_2(u, v)$. Define $J(u, v) = \begin{vmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ \frac{\partial h_2}{\partial u} & \frac{\partial h_2}{\partial v} \end{vmatrix} = \frac{\partial h_1}{\partial u}\frac{\partial h_2}{\partial v} - \frac{\partial h_1}{\partial v}\frac{\partial h_2}{\partial u}$. Then U and V are jointly continuous with density*

$$f_{U,V}(u, v) = f_{X,Y}\left(h_1(u, v), h_2(u, v)\right) |J(u, v)|, \quad (u, v) \in \text{Range}(g).$$

Example 18.2. Suppose that X and Y have joint density $f_{X,Y}$ and define $U = X + Y$, $V = X - Y$. Then $X = \frac{1}{2}(U + V)$ and $Y = \frac{1}{2}(U - V)$. The Jacobian is

$$J(u, v) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix},$$

so $|\det(J(u, v))| = \left|-\frac{1}{2}\right| = \frac{1}{2}$.

Thus Corollary 18.1 shows that (U, V) has density

$$f_{U,V}(u, v) = \frac{1}{2} f_{X,Y}\left(\frac{1}{2}(u + v), \frac{1}{2}(u - v)\right).$$

To find the density of $U = X + Y$, we integrate with respect to v to obtain

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u, v)dv = \frac{1}{2}\int_{-\infty}^{\infty} f_{X,Y}\left(\frac{1}{2}(u + v), \frac{1}{2}(u - v)\right) dv \\ &= \int_{-\infty}^{\infty} f_{X,Y}(u - w, w)dw \end{aligned}$$

where the final equality uses the change of variables $w = \frac{1}{2}(u - v)$.

When X and Y are independent, this gives the convolution formula.

Example 18.3. Suppose that $Z \sim \mathcal{N}(0, 1)$ and $V \sim \chi^2(n)$ are independent.

What is the density of $T = \dfrac{Z}{\sqrt{\frac{V}{n}}}$?

We first note that Z has density $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ and V has density $f_V(v) = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n}{2}}} v^{\frac{n}{2}-1} e^{-\frac{v}{2}} 1_{[0,\infty)}(v)$, so independence gives

$$f_{V,Z}(v,z) = \frac{1}{\Gamma\left(\frac{n}{2}\right)\sqrt{2\pi} 2^{\frac{n}{2}}} v^{\frac{n}{2}-1} e^{-\frac{v}{2}} e^{-\frac{z^2}{2}} 1_{[0,\infty)}(v).$$

Setting $T = \frac{Z}{\sqrt{\frac{V}{n}}}$, $W = V$, we have $Z = T\sqrt{\frac{W}{n}}$, $V = W$.

Since the Jacobian determinant of $(t,w) \mapsto \left(t\sqrt{\frac{w}{n}}, w\right)$ is

$$\det\left(J(t,w)\right) = \begin{vmatrix} \sqrt{\frac{w}{n}} & \frac{t}{2\sqrt{wn}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{n}},$$

Corollary 18.1 shows that (T,W) has density

$$\begin{aligned} f_{T,W}(t,w) &= \frac{1}{\Gamma\left(\frac{n}{2}\right)\sqrt{2\pi} 2^{\frac{n}{2}}} w^{\frac{n}{2}-1} e^{-\frac{w}{2}} \exp\left(-\frac{1}{2}\left(t\sqrt{\frac{w}{n}}\right)^2\right)\sqrt{\frac{w}{n}} 1_{[0,\infty)}(w) \\ &= \frac{1}{\Gamma\left(\frac{n}{2}\right)\sqrt{2n\pi} 2^{\frac{n}{2}}} w^{\frac{n+1}{2}-1} e^{-\frac{w}{2}} e^{-\frac{t^2 w}{2n}} 1_{[0,\infty)}(w) = \frac{1}{\Gamma\left(\frac{n}{2}\right)\sqrt{2n\pi} 2^{\frac{n}{2}}} w^{\frac{n+1}{2}-1} e^{-\frac{w}{2}\left(1+\frac{t^2}{n}\right)} 1_{[0,\infty)}(w). \end{aligned}$$

Recognizing $w^{\frac{n+1}{2}-1} e^{-\frac{w}{2}\left(1+\frac{t^2}{n}\right)}$ as the kernel of the Gamma $\left(\frac{n+1}{2}, \frac{2n}{t^2+n}\right)$ density gives

$$\begin{aligned} f_T(t) &= \frac{1}{\Gamma\left(\frac{n}{2}\right)\sqrt{2n\pi} 2^{\frac{n}{2}}} \int_0^\infty w^{\frac{n}{2}-1} e^{-\frac{w}{2}\left(1+\frac{t^2}{n}\right)} dw \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)\left(\frac{2n}{t^2+n}\right)^{\frac{n+1}{2}}}{\Gamma\left(\frac{n}{2}\right)\sqrt{2n\pi} 2^{\frac{n}{2}}} \cdot \frac{1}{\Gamma\left(\frac{n+1}{2}\right)\left(\frac{2n}{t^2+n}\right)^{\frac{n+1}{2}}} \int_0^\infty w^{\frac{n+1}{2}-1} e^{-\frac{w}{2}\left(1+\frac{t^2}{n}\right)} dw \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(\frac{n}{t^2+n}\right)^{\frac{n+1}{2}}. \end{aligned}$$

We say that T has the t-distribution with n degrees of freedom.

Example 18.4. Suppose that X and Y are independent standard normals. What is the density of $\frac{X}{Y}$?

Let $U = \frac{X}{Y}$, $V = Y$. Then $X = UV$, $Y = V$, and the transformation $(u,v) \mapsto (uv, v)$ has Jacobian determinant $\begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$.

Corollary 18.1 and the assumptions on X, Y show that

$$f_{U,V}(u,v) = f_X(uv) f_Y(v) |v| = \frac{1}{2\pi} e^{-\frac{u^2 v^2}{2}} e^{-\frac{v^2}{2}} |v| = \frac{1}{2\pi} |v| e^{-\frac{v^2}{2}(u^2+1)}.$$

The density of $U = \frac{X}{Y}$ is thus

$$\begin{aligned} f_U(u) &= \int_{-\infty}^\infty f_{U,V}(u,v) dv = -\frac{1}{2\pi}\int_{-\infty}^0 v e^{-\frac{v^2}{2}(u^2+1)} dv + \frac{1}{2\pi}\int_0^\infty v e^{-\frac{v^2}{2}(u^2+1)} dv \\ &= -\frac{1}{2\pi(u^2+1)}\int_\infty^0 e^{-w} dw + \frac{1}{2\pi(u^2+1)}\int_0^\infty e^{-w} dw = \frac{1}{\pi(u^2+1)}\int_0^\infty e^{-w} dw = \frac{1}{\pi(u^2+1)}. \end{aligned}$$

Thus the ratio of two independent standard normals is a standard Cauchy.

19. Conditional Distributions

Recall that the conditional probability of A given B is defined as $P(A\,|B) = \dfrac{P(A \cap B)}{P(B)}$ when $P(B) > 0$.

If X and Y are discrete random variables, it is natural to define the conditional mass function of X given that $Y = y$ by

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

for y such that $p_Y(y) > 0$.

Similarly, the conditional distribution function of X given that $Y = y$ is

$$F_{X|Y}(x\,|y) = P(X \le x\,|Y = y) = \frac{P(X \le x, Y = y)}{P(Y = y)} = \frac{\sum_{w \le x} P(X = w, Y = y)}{P(Y = y)} = \sum_{w \le x} p_{X|Y}(w\,|y).$$

If X and Y are independent, then

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

which gives the interpretation that X and Y are independent if knowing Y does not change your assessment of X.

Example 19.1. Suppose that X and Y are independent Poissons with rates λ and μ. What is the conditional distribution of X given $X + Y = n$?

For $k = 0, 1, ..., n$, we have

$$\begin{aligned} P(X = k\,|X + Y = n) &= \frac{P(X = k, X + Y = n)}{P(X + Y = n)} \\ &= \frac{P(X = k, Y = n - k)}{P(X + Y = n)} = \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)}. \end{aligned}$$

Since $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$, and, by Example 17.3, $X + Y \sim \text{Poisson}(\lambda + \mu)$, we have

$$\begin{aligned} P(X = k\,|X + Y = n) &= \frac{P(X = k)P(Y = n - k)}{P(X + Y = n)} = \frac{e^{-\lambda}\frac{\lambda^k}{k!}e^{-\mu}\frac{\mu^{n-k}}{(n-k)!}}{e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}} \\ &= \frac{e^{-\lambda-\mu}}{e^{-\lambda-\mu}} \cdot \frac{n!}{k!(n-k)!}\frac{\lambda^k\mu^{n-k}}{(\lambda+\mu)^n} = \binom{n}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k}. \end{aligned}$$

Writing $p = \frac{\lambda}{\lambda+\mu}$ gives

$$P(X = k\,|X + Y = n) = \binom{n}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k} = \binom{n}{k}p^k(1-p)^{n-k},$$

so, conditional on $X + Y = n$, $X \sim \text{Binomial}\left(n, \frac{\lambda}{\lambda+\mu}\right)$.

Now suppose that X and Y are continuous with joint density $f_{X,Y}$ and marginals f_X, f_Y.

Motivated by the discrete case, we define the conditional density of X given that $Y = y$ by

$$f_{X|Y}(x\,|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

provided that $f_Y(y) > 0$.

Heuristically, for small $h, k > 0$, multiplying by $h = \frac{hk}{k}$ gives

$$\begin{aligned} f_{X|Y}(x\,|y)h = \frac{f_{X,Y}(x,y)hk}{f_Y(y)k} &\approx \frac{P(x \le X \le x+h, y \le Y \le y+k)}{P(y \le Y \le y+k)} \\ &= P\left(x \le X \le x+h \,|y \le Y \le y+k\right), \end{aligned}$$

so $f_{X|Y}(x\,|y)$ represents the probability that X is in a neighborhood of x given that Y is in a neighborhood of y.

It should be emphasized that our construction of conditional densities is a definition, not a derivation from previous results.

Still it is a natural construct which can be used to assign probabilities to events associated with X when we are given that $Y = y$ (which has probability zero).
Namely, for $A \subseteq \mathbb{R}$, we have

$$P(X \in A\,|Y = y) = \int_A f_{X|Y}(x,y)dx.$$

* Note that this conditional probability is a function of y. A more sophisticated treatment of conditioning would define conditional probabilities as random variables, and one can show that with this approach, we have that for each $A \subseteq \mathbb{R}$, $P(X \in A\,|Y) = h(Y)$ where $h(y) = \int_A \frac{f_{X,Y}(x,y)}{f_Y(y)}dx$.

By taking $A = (-\infty, a]$, we can define the conditional distribution function of X given that $Y = y$ by

$$F_{X|Y}(a\,|y) = P(X \le a\,|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x\,|y)dx.$$

Observe also that, as with the discrete case, if X and Y are independent, then

$$f_{X|Y}(x\,|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$$

and

$$f_{Y|X}(y\,|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y).$$

Example 19.2. Suppose that the joint density of X and Y is given by

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

Find $P(X > 1\,|Y = y)$ for $y > 0$.

To find $f_{X|Y}(x\,|y)$, we first need to compute the marginal density of Y.
Making the obvious change of variables yields

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = e^{-y}\int_0^{\infty} \frac{e^{-\frac{x}{y}}}{y}dx = e^{-y}\int_0^{\infty} e^{-w}dw = e^{-y}.$$

It follows that

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

hence

$$P(X>1\,|Y=y) = \int_1^{\infty} f_{X|Y}(x\,|y)dx = \int_1^{\infty}\frac{1}{y}e^{-\frac{x}{y}}dx = \int_{\frac{1}{y}}^{\infty} e^{-w}dw = e^{-\frac{1}{y}}.$$

Example 19.3. Suppose that X is an exponential random variable with rate $U \sim \text{Unif}(0,1)$. That is, conditional on $U=u$, $X \sim \text{Exponential}(u)$. What is the density of X?

We know that for $u \in (0,1)$,

$$\frac{f_{U,X}(u,x)}{f_U(u)} = f_{X|U}(x\,|u) = ue^{-ux},$$

so

$$f_{U,X}(u,x) = ue^{-ux}f_U(u) = ue^{-ux}1_{(0,1)}(u).$$

Consequently,

$$f_X(x) = \int_0^1 ue^{-ux}du = \left[-\frac{u}{x}e^{-ux}\Big|_0^1 + \frac{1}{x}\int_0^1 e^{-ux}du\right] = \frac{1}{x^2}\left(1-e^{-x}\right) - \frac{e^{-x}}{x},\quad x>0.$$

We now turn to the problem of computing expectations of functions of several random variables (without resorting to the intermediary step of finding the law of the new random variable).

The result is exactly what you would guess from previous cases.

Theorem 20.1.

(1) *If $X_1, ..., X_n$ have joint mass function $p(x_1, ..., x_n)$ and $g : \mathbb{R}^n \to \mathbb{R}$, then*

$$E[g(X_1, ..., X_n)] = \sum_{\substack{(x_1,...,x_n)\in\mathbb{R}^n: \\ p(x_1,...,x_n)>0}} g(x_1, ..., x_n)p(x_1, ..., x_n)$$

if the sum converges absolutely or g is nonnegative.

(2) *If $X_1, ..., X_n$ have joint density $f(x_1, ..., x_n)$ and $g : \mathbb{R}^n \to \mathbb{R}$ is nonnegative or integrable, then*

$$E[g(X_1, ..., X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, ..., x_n)f(x_1, ..., x_n)dx_1 \cdots dx_n.$$

The discrete case is proved by expressing the mass function of $g(X_1, ..., X_n)$ as

$$P\left(g(X_1, ..., X_n) = y\right) = \sum_{\substack{(x_1,...,x_n)\in\mathbb{R}^n: \\ p(x_1,...,x_n)>0 \\ g(x_1,...,x_n)=y}} p(x_1, ..., x_n)$$

and applying the ordinary definition of discrete expectation, and the jointly continuous case is proved using Lemma 13.1.

For example, suppose that X and Y have joint density f and let $g : \mathbb{R}^2 \to [0, \infty)$. Then

$$\begin{aligned} E[g(X,Y)] &= \int_0^{\infty} P\left(g(X,Y) > z\right) dz = \int_0^{\infty} \left(\iint_{(x,y):g(x,y)>z} f(x,y)dxdy \right) dz \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_0^{g(x,y)} f(x,y)dzdxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy. \end{aligned}$$

An especially important application that we have put off until now is

Theorem 20.2. *If X and Y are jointly continuous with $E\left|X\right|, E\left|Y\right| < \infty$, then $E[X + Y] = E[X] + E[Y]$.*

Proof. Denoting the joint density and marginal densities as usual and appealing to Theorem 20.1, we have

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right)dx + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx\right)dy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy = E[X] + E[Y]. \end{aligned}$$

□

Of course, Theorem 20.2 extends to sums of any finite number of random variables by induction.

Theorem 20.2 also allows us to establish the monotonicity of expectation in the jointly continuous case. (Like linearity, this is true in general, but we need more powerful theories of integration to prove it.)

Theorem 20.3. *If X and Y are jointly continuous with $X \geq Y$ (i.e. $X(\omega) \geq Y(\omega)$ for every $\omega \in \Omega$), then $E[X] \geq E[Y]$.*

Proof. Let $Z = X - Y$. Then $Z \geq 0$, so

$$0 \leq \int_0^\infty P(Z > z)dz = E[Z] = E[X + (-Y)] = E[X] + E[-Y] = E[X] - E[Y]. \qquad \square$$

The linearity of expectation can also be used to simplify calculations by expressing random variables as sums of simpler random variables.

Example 20.1. Suppose that n balls are selected at random and without replacement from an urn containing N balls, m of which are white. Let X be the number of white balls in the sample. We know that $X \sim$ Hypergeometric(N, m, n) and we have computed $E[X] = \frac{mn}{N}$ using binomial identities. Here is another way. Enumerate the m white balls and let $X_i = 1\{\text{ball } i \text{ selected}\}$. Then $X = \sum_{i=1}^n X_i$.

Since

$$E[X_i] = P(\text{ball } i \text{ selected}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

we have

$$E[X] = \sum_{i=1}^m E[X_i] = \sum_{i=1}^m \frac{n}{N} = \frac{mn}{N}.$$

Example 20.2. In the coupon collecting problem where one collects one of n types of coupons with equal probability at each stage, it is of interest to estimate $T =$ time to collect all coupons.

To compute $E[T]$, we let T_k be the time until k distinct types have been collected for $k = 1, ..., n$ and set $T_0 = 0$, so that $T = T_n = \sum_{k=1}^n (T_k - T_{k-1})$.

Writing $X_k = T_k - T_{k-1}$, we see that X_k is the time to collect a new type after $k-1$ types have already been collected.

Given that $k-1$ types have been collected, the probability of collecting a new type is $\frac{n-k+1}{n}$, so $X_k \sim$ Geometric $\left(\frac{n-k+1}{n}\right)$. Since $Y \sim$ Geometric(p) has mean $E[Y] = \frac{1}{p}$, we have

$$E[T] = \sum_{k=1}^n E[T_k - T_{k-1}] = \sum_{k=1}^n E[X_k] = \sum_{k=1}^n \frac{n}{n-k+1} = n\sum_{j=1}^n \frac{1}{j}$$

where the final equality used the substitution $j = n - k + 1$.

* Since $\log(n) \leq \sum_{j=1}^n \frac{1}{j} \leq \log(n) + 1$ (by bounding $\log(n) = \int_1^n \frac{dx}{x}$ with the upper Riemann sum $\sum_{j=1}^{n-1} \frac{1}{j} \leq \sum_{j=1}^n \frac{1}{j}$ and the lower Riemann sum $\sum_{j=2}^n \frac{1}{j} = \sum_{j=1}^n \frac{1}{j} - 1$), we have $\frac{E[T(n)]}{n\log(n)} \to 1$ as $n \to \infty$.

Example 20.3. Another interesting quantity in the coupon collector problem is $N_t = \#$ types collected by time t. We can relax the assumptions a little bit and suppose that type i has probability p_i of being drawn at each stage, where $p_1, ..., p_n > 0$ and $\sum_{i=1}^n p_i = 1$. (Setting $p_i = \frac{1}{n}$ gives the previous case.)

Writing $X_{i,t} = 1\{\text{type } i \text{ collected by time } t\}$, we have $N_t = \sum_{i=1}^n X_{i,t}$ and

$$E[X_{i,t}] = P(\text{type } i \text{ collected by time } t) = 1 - (1-p_i)^t.$$

It follows that the expected number of coupon types collected by time t is

$$E[N_t] = \sum_{i=1}^n E[X_{i,t}] = n - \sum_{i=1}^n (1-p_i)^t.$$

In general, many random variables of interest can be expressed as $X = \#$ of A_i's that occur for some collection of events $\{A_1, ..., A_n\}$.

Writing $X = \sum_{i=1}^n 1_{A_i}$, we have

$$E[X] = \sum_{i=1}^n E\left[1_{A_i}\right] = \sum_{i=1}^n P(A_i).$$

To find the variance of X as well, we need to compute the expectation of

$$X^2 = \left(\sum_{i=1}^n 1_{A_i}\right)^2 = \sum_{i=1}^n 1_{A_i}^2 + \sum_{i \neq j} 1_{A_i} 1_{A_j}.$$

Since $1_{A_i}^2 = 1_{A_i}$ and $1_{A_i} 1_{A_j} = 1_{A_i \cap A_j}$, we have

$$E[X^2] = \sum_{i=1}^n E[1_{A_i}] + \sum_{i \neq j} E[1_{A_i \cap A_j}] = \sum_{i=1}^n P(A_i) + 2 \sum_{1 \leq i < j \leq n} P(A_i \cap A_j).$$

Higher moments can be computed similarly using $\prod_{i \in S} 1_{A_i} = 1_{\cap_{i \in S} A_i}$.

The multinomial theorem gives the concise (but not especially useful) expression

$$E\left[\left(\sum_{i=1}^n 1_{A_i}\right)^k\right] = \sum_{m_1 + ... + m_n = k} \binom{k}{m_1, ..., m_n} P\left(\bigcap_{j : m_j > 0} A_j\right).$$

Example 20.4. To compute the variance of $X \sim \text{Hypergeometric}(N, m, n)$, we write $X = \sum_{i=1}^m X_i$ as in Example 20.1. Since

$$E[X_i] = P(\text{ball } i \text{ selected}) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

and

$$E[X_i X_j] = P(\text{balls } i \text{ and } j \text{ selected}) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)},$$

we have

$$E[X^2] = \sum_{i=1}^m E[X_i] + \sum_{i \neq j} E[X_i X_j] = \frac{mn}{N} + \frac{m(m-1)n(n-1)}{2N(N-1)},$$

so

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{mn}{N} + \frac{m(m-1)n(n-1)}{N(N-1)} - \frac{m^2 n^2}{N^2} = \frac{mn\,(N-m)\,(N-n)}{N^2(N-1)}.$$

Example 20.5. Consider the hat check problem where n people drop off their hats at a reception desk and random hats are returned.

Letting $X_i = 1\{\text{person } i \text{ receives their original hat}\}$, we see that the number of people who get their hats back is $X = \sum_{i=1}^{n} X_i$.

Since $E[X_i] = \frac{1}{n}$, we have

$$E[X] = \sum_{i=1}^{n} E[X_i] = n \cdot \frac{1}{n} = 1$$

Similarly,

$$E[X_i X_j] = P(i \text{ and } j \text{ get hats back}) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)},$$

so

$$E[X^2] = \sum_{i=1}^{n} E[X_i] + \sum_{i \neq j} E[X_i X_j] = n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} = 2,$$

and thus

$$\text{Var}(X) = E[X^2] - E[X]^2 = 2 - 1^2 = 1.$$

21. Covariance

A recurring theme in our discussions of independence is some sort of factoring condition.

In terms of expectation, we have

Theorem 21.1. *If X and Y are independent random variables and g, h are nonnegative or integrable, then*

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Proof. We will prove the Theorem under the assumption that X and Y are continuous.
The discrete case is almost identical and the general case uses the same basic ideas.

Denoting the densities of X and Y in the usual manner, we have

$$\begin{aligned} E[g(X)h(Y)] &= \iint_{\mathbb{R}^2} g(x)h(y)f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} h(y)f_Y(y)\left(\int_{-\infty}^{\infty} g(x)f_X(x)dx\right)dy = E[g(X)]\int_{-\infty}^{\infty} h(y)f_Y(y)dy = E[g(X)]E[h(Y)]. \end{aligned}$$

□

One useful measure of certain types of dependence relations between two random variables is their covariance.

Definition. If X and Y are random variables with $E[X] = \mu_X$, $E[Y] = \mu_Y$, and $E[X^2], E[Y^2] < \infty$, then we define the *covariance* between X and Y as

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

There is a nice shortcut formula for the covariance that is useful in computations:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] = E[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y] \\ &= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X\mu_Y \\ &= E[XY] - E[Y]E[X] - E[X]E[Y\} + E[X]E[Y] = E[XY] - E[X]E[Y]. \end{aligned}$$

Thus just as the variance is the expectation of the square minus the square of the expectation, the covariance is the expectation of the product minus the product of the expectations.

Note that Theorem 21.1 shows that if X and Y are independent, then $E[XY] = E[X]E[Y]$, so $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$.

However, it is possible for random variables to have covariance zero (in which case we say that they are uncorrelated), but still be dependent.

For example, suppose that $X \sim \mathrm{Unif}(-1,1)$ and let $Y = X^2$. Then X and Y are clearly dependent, but since $E[X] = \frac{1}{2}\int_{-1}^{1} x dx = 0$ and $E[X^3] = \frac{1}{2}\int_{-1}^{1} x^3 dx = 0$, we have

$$\mathrm{Cov}(X,Y) = E[X\cdot X^2] - E[X]E[X^2] = E[X^3] - E[X]E[X^2] = 0 - 0E[X^2] = 0.$$

Basically, covariance is a measure of linear dependence: If $\mathrm{Cov}(X,Y) > 0$, then larger values of X imply larger values of Y, while if $\mathrm{Cov}(X,Y) < 0$, then larger values of X imply smaller values of Y.

In the above example, X and Y depended on one another nonlinearly.

Our next proposition contains some useful observations about covariance.

Proposition 21.1. *Suppose that $X, X_1, ..., X_m$ and $Y, Y_1, ..., Y_n$ have finite second moments, and let $a \in \mathbb{R}$. Then*

(1) $\text{Cov}(X, a) = 0$
(2) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
(3) $\text{Cov}(X, X) = \text{Var}(X)$
(4) $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
(5) $\text{Cov}\left(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} \text{Cov}(X_i, Y_j)$

Proof. We will only prove statement 5 since the rest are immediate consequences of the definition.

If $X_1, ..., X_m$ and $Y_1, ..., Y_n$ have finite second moments, then one can show that $\sum_{i=1}^{m} X_i$ and $\sum_{j=1}^{n} Y_j$ do as well.

Letting μ_{X_i} and μ_{Y_j} denote the means of the variates in question, it follows from the linearity of expectation that $E\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mu_{X_i}$ and $E\left[\sum_{j=1}^{n} Y_j\right] = \sum_{j=1}^{n} \mu_{Y_j}$.

Therefore,

$$\begin{aligned}
\text{Cov}\left(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j\right) &= E\left[\left(\sum_{i=1}^{m} X_i - \sum_{i=1}^{m} \mu_{X_i}\right)\left(\sum_{j=1}^{n} Y_j - \sum_{j=1}^{n} \mu_{Y_j}\right)\right] \\
&= E\left[\sum_{i=1}^{m} (X_i - \mu_{X_i}) \sum_{j=1}^{n} \left(Y_j - \mu_{Y_j}\right)\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} E\left[(X_i - \mu_{X_i})\left(Y_j - \mu_{Y_j}\right)\right] = \sum_{i=1}^{m}\sum_{j=1}^{n} \text{Cov}(X_i, Y_j).
\end{aligned}$$

□

Corollary 21.1. *Suppose that $X_1, ..., X_n$ have finite second moments. Then*

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j).$$

In particular, if the $X_i's$ are uncorrelated (as is the case for independent random variables), then

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

Proof. Statements 3, 5, and 2 of Proposition 21.1 give

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j).
\end{aligned}$$

The second statement follows directly from the first. □

Example 21.1. Suppose that $X_1, ..., X_n$ are i.i.d. with mean $E[X_1] = \mu$ and variance $\text{Var}(X_1) = \sigma^2$. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the sample mean and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ the sample variance.

By linearity

$$E\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot n\mu = \mu,$$

and by independence

$$\text{Var}\left(\overline{X}_n\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

At first glance, it may seem that we defined the sample variance with the wrong normalizing factor. The reason that we don't look at the arithmetic average of the squared deviations from the sample mean is that we want our estimator to be consistent in the sense that $E\left[S_n^2\right] = \sigma^2$.

To see that this is so, we first compute

$$\begin{aligned}
E\left[S_n^2\right] &= \frac{1}{n-1}\sum_{i=1}^{n} E\left[\left(X_i - \overline{X}_n\right)^2\right] \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} E\left[X_i^2\right] - 2\sum_{i=1}^{n} E\left[X_i\overline{X}_n\right] + \sum_{i=1}^{n} E\left[\overline{X}_n^2\right]\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} E\left[X_i^2\right] - 2E\left[\overline{X}_n\sum_{i=1}^{n} X_i\right] + nE\left[\overline{X}_n^2\right]\right) \\
&= \frac{1}{n-1}\left(nE\left[X_1^2\right] - 2E\left[\overline{X}_n\cdot n\overline{X}_n\right] + nE\left[\overline{X}_n^2\right]\right) \\
&= \frac{1}{n-1}\left(nE\left[X_1^2\right] - 2nE\left[\overline{X}_n^2\right] + nE\left[\overline{X}_n^2\right]\right) = \frac{1}{n-1}\left(nE\left[X_1^2\right] - nE\left[\overline{X}_n^2\right]\right).
\end{aligned}$$

Now the shortcut formula $\text{Var}(Y) = E[Y^2] - E[Y]^2$ shows that $E\left[X_1^2\right] = \sigma^2 + \mu^2$ and $E\left[\overline{X}_n^2\right] = \frac{\sigma^2}{n} + \mu^2$, so we see that

$$E\left[S_n^2\right] = \frac{1}{n-1}\left(nE\left[X_1^2\right] - nE\left[\overline{X}_n^2\right]\right) = \frac{n}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) = \frac{1}{n-1}\cdot(n-1)\sigma^2 = \sigma^2.$$

Example 21.2. When we discussed the negative binomial distribution, we noted that if $Y \sim \text{NegBin}(r, p)$, then $Y = \sum_{i=1}^{r} X_i$ with $X_1, ..., X_r$ i.i.d. Geometric(p) random variables. (Y is the time until the r^{th} success in i.i.d. Bernoulli trials and X_i is the time between the $(i-1)^{\text{st}}$ success and the i^{th}.)

Recalling that $\text{Var}(X_i) = \frac{1-p}{p^2}$, it is now elementary to give an honest derivation of the variance of Y:

$$\text{Var}(Y) = \text{Var}\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r}\text{Var}(X_i) = \frac{r(1-p)}{p^2}.$$

The relationship $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$ shows that the covariance between two random variables can be very high even if the two are not strongly linearly dependent just because one happens to be much larger than the other.

For example, if there is a moderate covariance between one's height and that of their second cousin when both are measured in meters, the covariance will be huge if one is measured in meters and the other in millimeters.

To account for such discrepancies in scale, one often considers the correlation instead.

Definition. The *correlation* between random variables X and Y with $\text{Var}(X) = \sigma_X^2, \text{Var}(Y) = \sigma_Y^2 \in (0, \infty)$ is defined as

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

Lemma 21.1. *If X is a random variable with $E|X| < \infty$, then $|E[X]| \leq E|X|$.*

Proof. Let $Y = |X| - X$. Then Y is nonnegative, so $0 \leq E[Y] = E|X| - E[X]$, and thus $E|X| \geq E[X]$. Similarly, $Z = |X| + X$ is nonnegative, so $0 \leq E[Z] = E|X| + E[X]$, hence $E|X| \geq -E[X]$ as well and the result follows. □

Lemma 21.2. *If X and Y are random variables with $E|X|, E|Y| < \infty$, then*

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

Proof. The first inequality follows from Lemma 21.1. For the second, we observe that for all $t \in \mathbb{R}$,

$$0 \leq E\left[(|X| + t|Y|)^2\right] = E\left[X^2\right] + 2tE|XY| + t^2 E\left[Y^2\right] = q(t),$$

so q has at most one real root and thus a nonpositive discriminant

$$(2E|XY|)^2 - 4E\left[Y^2\right] E\left[X^2\right] \leq 0.$$ □

Proposition 21.2. *If X and Y have finite nonzero variances, then $-1 \leq \rho(X, Y) \leq 1$.*

Proof. Lemma 21.2 gives

$$\text{Cov}(X, Y)^2 = |E[(X - \mu_X)(Y - \mu_Y)]|^2 \leq E\left[(X - \mu_X)^2\right] E\left[(Y - \mu_Y)^2\right] = \sigma_X^2 \sigma_Y^2,$$

hence

$$\rho(X, Y)^2 = \frac{\text{Cov}(X, Y)^2}{\sigma_X^2 \sigma_Y^2} \leq 1.$$ □

Correlation is thus covariance normalized to lie between -1 and 1. This is why we call random variables with zero covariance uncorrelated. Since the normalizing factor is positive, we can likewise talk about random variables having positive or negative correlation based on the sign of the covariance.

* Reasoning about discriminants as above, one can show that $|\rho(X, Y)| = 1$ if and only if $Y = aX + b$ for some constants a, b, so a correlation of ± 1 means a perfect linear relationship.

Example 21.3. Let 1_A and 1_B be indicators for the events A and B. Then

$$\mathrm{Cov}(1_A, 1_B) = E[1_A 1_B] - E[1_A]E[1_B] = P(A \cap B) - P(A)P(B) = P(B)\left(P(A\,|B\,) - P(A)\right).$$

Thus 1_A and 1_B are positively correlated if $P(A\,|B\,) > P(A)$, negatively correlated if $P(A\,|B\,) < P(A)$, and uncorrelated if A and B are independent.

In fact, for indicator random variables, independence and uncorrelatedness are synonymous.

It turns out that jointly normal random variables also have this property. (This is not true for normal random variables that are not jointly normal, though it is often mistakenly claimed to be.)

22. Conditional Expectation

Recall that if X and Y are discrete, then the conditional mass function for X given that $Y = y$ is given by

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

provided that $p_Y(y) > 0$.

$p_{X|Y}(\cdot\,|y)$ is clearly a p.m.f. for such y since $p_{X,Y}, p_Y$ are nonnegative and

$$\sum_x p_{X|Y}(x\,|y) = \frac{\sum_x p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_Y(y)}{p_Y(y)} = 1.$$

Accordingly, it makes sense to define the conditional expectation of X given that $Y = y$ by

$$E[X\,|Y = y] = \sum_x x p_{X|Y}(x\,|y).$$

Example 22.1. Suppose X and Y are independent Binomial(n,p) random variables. For $m \le 2n$, what is $E[X\,|X+Y=m]$?

We first observe that $X + Y \sim$ Binomial$(2n,p)$. This can be derived analytically or by representing X and Y as sums of independent Bernoulli random variables. Accordingly, for any $k \le \min\{m,n\}$,

$$\begin{aligned} P(X = k\,|X+Y=m) &= \frac{P(X=k, Y=m-k)}{P(X+Y=m)} = \frac{\binom{n}{k}p^k(1-p)^{n-k}\binom{n}{m-k}p^{m-k}(1-p)^{n-m+k}}{\binom{2n}{m}p^m(1-p)^{2n-m}} \\ &= \frac{p^k p^{m-k}}{p^m} \cdot \frac{(1-p)^{n-k}(1-p)^{n-m+k}}{(1-p)^{2n-m}} \cdot \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}} = \frac{\binom{n}{k}\binom{n}{m-k}}{\binom{2n}{m}}. \end{aligned}$$

In other words, conditional on $X + Y = m$, $X \sim$ Hypergeometric$(2n,n,m)$, and thus has mean $\dfrac{mn}{2n} = \dfrac{m}{2}$.

In exactly the same way, if X and Y are jointly continuous and $f_Y(y) > 0$, then

$$f_{X|Y}(x\,|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

is a density function since it is nonnegative and

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

Thus we can define the *conditional expectation* of X given that $Y = y$ as

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

Since $p_{X|Y}$ and $f_{X|Y}$, are mass functions and densities in the usual sense, all the familiar results about expectations apply.

For example, if (X,Y) is continuous, then $E[g(X)\,|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x\,|y)dx$ for suitable g.

Example 22.2. Suppose that the joint density of X and Y is given by

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

What is $E[X^3 \,|Y = y]$?

The marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^y \frac{e^{-y}}{y} dx = e^{-y}, \quad y > 0,$$

so

$$f_{X|Y}(x\,|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{y} 1_{(0,y)}(x).$$

In other words, conditional on $Y = y$, $X \sim \text{Unif}(0, y)$.

Even without this observation, we can compute,

$$E[X^3 \,|Y = y] = \int_{-\infty}^{\infty} x^3 f_{X|Y}(x\,|y)dx = \int_0^y \frac{x^3}{y} dx = \frac{x^4}{4y}\Big|_{x=0}^{y} = \frac{y^3}{4}.$$

It is useful (and actually less problematic from a theoretical standpoint) to think of conditional expectation as a random variable:

$$E[X\,|Y\,] = h(Y), \quad h(y) = E[X\,|Y = y].$$

That is, $E[X\,|Y\,](\omega)$ is the random variable which is equal to $E[X\,|Y = y]$ whenever $Y(\omega) = y$.

With this perspective, we can sometimes simplify computations of ordinary expectation by using conditional expectations as an intermediary.

Specifically, we have the law of total expectation

Proposition 22.1. $E[X] = E[E[X\,|Y\,]]$

We will not prove this in general, but include the argument in the discrete case to illustrate the meaning:

Proof. If X and Y are discrete, we have

$$\begin{aligned} E\left[E[X\,|Y\,]\right] &= \sum_y E[X\,|Y = y]P(Y = y) \\ &= \sum_y \left(\sum_x x \frac{P(X = x, Y = y)}{P(Y = y)} \right) P(Y = y) \\ &= \sum_y \sum_x x P(X = x, Y = y) \\ &= \sum_x x \sum_y P(X = x, Y = y) = \sum_x x P(X = x) = E[X]. \end{aligned}$$

□

Example 22.3. Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables with finite mean, and that N is an $\mathbb{N}$-valued random variable with finite mean which is independent of the X_i's. What can we say about the random sum $\sum_{i=1}^{N} X_i$?

Let $S = \sum_{i=1}^{N} X_i$. By Proposition 22.1 and the independence assumptions, we have

$$\begin{aligned} E[S] &= E\left[E[S\,|N]\right] = \sum_{n=1}^{\infty} E[S\,|N=n]P(N=n) \\ &= \sum_{n=1}^{\infty} E\left[\sum_{i=1}^{n} X_i\,|N=n\right] P(N=n) \\ &= \sum_{n=1}^{\infty} E\left[\sum_{i=1}^{n} X_i\right] P(N=n) \\ &= \sum_{n=1}^{\infty} nE[X_1]P(N=n) \\ &= E[X_1]\sum_{n=1}^{\infty} nP(N=n) = E[X_1]E[N]. \end{aligned}$$

In a similar manner, assuming that X_1 and N have finite variance, one can compute $E[S^2]$ and thus $\mathrm{Var}(S)$. The computation is fairly tedious, so we just state the result:

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = E[N]\mathrm{Var}(X_1) + E[X_1^2]\mathrm{Var}(N).$$

Since probabilities can be expressed as expectations of indicators via $P(A) = E[1_A]$, we can define the conditional probability of A given Y as the random variable $P(A\,|Y\,) = E[1_A\,|Y\,]$.

Proposition 22.1 then implies the law of total probability

$$P(A) = E\left[P(A\,|Y\,)\right].$$

When Y is discrete, this says that $P(A) = \sum_y P(A\,|Y=y)P(Y=y)$.

Of course, $P(A\,|Y=y)P(Y=y) = P(A, Y=y)$, so all we are really doing is partitioning the event A according to the value of Y.

Example 22.4. Suppose that you wish to hire the best of n candidates for a particular job. However, immediately after each interview you must decide whether you will hire that candidate or move on to the next interview. Once rejected, a candidate cannot be recalled. (Often this problem is stated in terms of proposing marriage after a series of dates. Spurned suitors will not accept later proposals.) When making a decision, your only information is the relative rankings of the candidates you have seen thus far. Assuming that the candidates can be ranked from best to worst and all $n!$ interview orders are equally likely, is there a strategy that gives a high probability of hiring the best candidate?

It turns out that the following strategy does a surprisingly good job:

For some fixed $1 \leq k \leq n$, reject the first k candidates and then choose the first candidate who is better than all those seen before.

Let X be the position of the best candidate (so X is uniformly distributed on $\{1, ..., n\}$) and let B_k be the event that the best candidate is hired.

Then,

$$P(B_k) = \sum_{i=1}^{n} P(B_k \,|X = i)P(X = i) = \frac{1}{n}\sum_{i=1}^{n} P(B_k \,|X = i).$$

If $i \leq k$, then $P(B_k \,|X = i) = 0$ since the first k candidates are automatically rejected.

If $i > k$, then the best candidate is selected if the best of the first $i-1$ candidates was one of the first k.

Since the relative orderings of the first $i-1$ candidates are equiprobable given that $X = i$, we have

$$P\left(\text{candidate } j \text{ is best of first } i-1 \,|X = i\right) = \frac{1}{i-1}$$

for all $1 \leq j \leq i-1$.

Accordingly, for $i > k$,

$$P(B_k \,|X = i) = P\left(\text{best of first } i-1 \text{ is among first } k \,|X = i\right) = \frac{k}{i-1}.$$

Thus the probability of choosing the best candidate is

$$\begin{aligned} P(B_k) &= \frac{1}{n}\sum_{i=1}^{n} P(B_k \,|X = i) = \frac{1}{n}\sum_{i=k+1}^{n} \frac{k}{i-1} \\ &= \frac{k}{n}\sum_{i=k}^{n-1} \frac{1}{i} \approx \frac{k}{n}\int_{k}^{n} \frac{dx}{x} = \frac{k}{n}\log\left(\frac{n}{k}\right). \end{aligned}$$

(The integral is approximated by the left Riemann sum for a partition of $[k, n]$ into intervals of length 1. We could get a lower bound by viewing it as a right Riemann sum for the integral from $k-1$ to $n-1$, but the answer wouldn't be quite as elegant...)

To find the optimal k, we differentiate $g(x) = \frac{x}{n}\log\left(\frac{n}{x}\right)$ and solve

$$0 = g'(x) = \frac{1}{n}\log\left(\frac{n}{x}\right) + \frac{x}{n}\left(-\frac{n}{x^2}\right)\frac{x}{n} = \frac{1}{n}\left(\log\left(\frac{n}{x}\right) - 1\right)$$

to get $x = \dfrac{n}{e}$. (This is a maximum since $g'(x) = \frac{1}{n}\left(\log(n) - \log(x) - 1\right)$ is clearly decreasing in x.)

Taking k to be the nearest integer to $\dfrac{n}{e}$ shows that our probability of choosing the best candidate is approximately

$$\frac{\frac{n}{e}}{n}\log\left(\frac{n}{\frac{n}{e}}\right) = \frac{1}{e} \approx 0.368.$$

To recap, if we reject the first $\left[\frac{n}{e}\right]$ candidates and then hire the first who is better than those previously interviewed, then we will select the top candidate about 37% of the time. Note that this result does not depend on the size of n (except for the accuracy of the approximation used in the argument).

Example 22.5. Suppose X and Y are independent with densities f_X and f_Y.

By conditioning on the value of Y, we obtain

$$\begin{aligned} P(X<Y) &= \int_{-\infty}^{\infty} P(X<Y\,|Y=y\,)f_Y(y)dy \\ &= \int_{-\infty}^{\infty} P(X<y\,|Y=y\,)f_Y(y)dy \\ &= \int_{-\infty}^{\infty} P(X<y)f_Y(y)dy \\ &= \int_{-\infty}^{\infty} F_X(y)f_Y(y)dy \end{aligned}$$

where $F_X(y)=\int_{-\infty}^{y} f_X(x)dx$.

Of course this can also be obtained by

$$\begin{aligned} P(X<Y) &= \iint_{x<y} f_{X,Y}(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{y} f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} f_Y(y)\left(\int_{-\infty}^{y} f_X(x)dx\right)dy = \int_{-\infty}^{\infty} F_X(y)f_Y(y)dy. \end{aligned}$$

23. Laws of Large Numbers

For the remainder of the course, we will concentrate on asymptotic properties of sums of i.i.d. random variables.

Our first set of results are the weak and strong laws of large numbers, which show that the sample mean $\frac{1}{n}\sum_{i=1}^n X_i$ converges (in various senses) to the population mean $E[X_1]$ as $n \to \infty$. These give support to the interpretation of probabilities representing long term relative frequencies:

Suppose that a sequence of independent trials is performed, and let X_i be the indicator that the event A occurred in trial i. Then $S_n = \sum_{i=1}^n X_i$ is the number of times that A occurred in the first n trials, and the laws of large numbers show that the frequency with which A occurs, $\frac{1}{n}S_n$, converges to $E[X_1] = P(A)$.

(The weak law says that $\frac{1}{n}S_n$ will be close to $P(A)$ with high probability for large n, and the strong law says that almost all particular realizations of this sequence of experiments will result in an empirical probability which is close to $P(A)$ for large n.)

We begin with some useful results on tail probabilities called *Chebychev bounds*.

Theorem 23.1. *Suppose that X is a nonnegative random variable. Then for any $a > 0$,*

$$P(X \geq a) \leq \frac{E[X]}{a}.$$

Proof. Let $I = 1\{X \geq a\}$. Then $\dfrac{X}{a} \geq I$, so

$$P(X \geq a) = E[I] \leq \frac{E[X]}{a}. \qquad \square$$

Corollary 23.1. *Suppose that X is a random variable with finite mean $\mu = E[X]$ and variance $\sigma^2 = E[(X-\mu)^2]$. Then for any $a > 0$,*

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

Proof. Since $(X-\mu)^2$ is a nonnegative random variable, Theorem 23.1 gives

$$P(|X-\mu| \geq a) = P\left((X-\mu)^2 \geq a^2\right) \leq \frac{E\left[(X-\mu)^2\right]}{a^2} = \frac{\sigma^2}{a^2}. \qquad \square$$

The preceding results allow us to control the probability that a random variable with finite mean is large and that a random variable with finite variance is far from its mean.

In many cases, the tails decay much quicker than the polynomial rate given by Chebychev bounds. For example, if $Z \sim \mathcal{N}(0,1)$, then

$$\begin{aligned} P(|Z| > a) &= \frac{2}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{x^2}{2}} dx \leq \frac{2}{\sqrt{2\pi}} \int_a^\infty \frac{x}{a} e^{-\frac{x^2}{2}} dx \\ &= \sqrt{\frac{2}{\pi a^2}} \int_a^\infty x e^{-\frac{x^2}{2}} dx = \sqrt{\frac{2}{\pi a^2}} \int_{\frac{a^2}{2}}^\infty e^{-u} du = \sqrt{\frac{2}{\pi a^2}} e^{-\frac{a^2}{2}}. \end{aligned}$$

On the other hand, if X has p.m.f. $p(1) = p(-1) = \frac{1}{2k}$ and $p(0) = 1 - \frac{1}{k}$, then $E[X] = 0$ and $\mathrm{Var}(X) = E[X^2] = \frac{1}{k}$, so

$$P(|X - E[X]| \geq 1) = P(X = \pm 1) = \frac{1}{k} = \frac{\mathrm{Var}(X)}{1^2},$$

hence there are infinitely many examples where Chebychev is sharp.

Definition. A sequence of random variables $X_1, X_2, \ldots$ is said to *converge in probability* to X if for every $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left(|X_n - X| > \varepsilon\right) = 0.$$

In this case, we write $X_n \to_p X$.

In words, $X_n \to_p X$ if the probability that X_n differs from X by any given amount can be made arbitrarily small by choosing n sufficiently large.

The *weak law of large numbers* says that under very general conditions, the arithmetic mean of i.i.d. random variables, $\frac{1}{n}\sum_{i=1}^n X_i$, converges in probability to their common expectation, $E[X_1]$.

Theorem 23.2. *Suppose that $X_1, X_2, \ldots$ are i.i.d. with $E|X_1| < \infty$. Then $\dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i \to_p E[X_1]$.*

We will not prove Theorem 23.2 in its full strength, but we can show that the sample mean converges to $\mu = E[X_1]$ in probability if we tack on the small additional assumption that $\sigma^2 = \mathrm{Var}(X_1) < \infty$. In this case, we can weaken the i.i.d. assumption as well.

Theorem 23.3. *Suppose that $X_1, X_2, \ldots$ are pairwise uncorrelated random variables with common mean $\mu = E[X_i]$ and uniformly bounded variances $\mathrm{Var}(X_i) \leq \sigma^2 < \infty$. Then $\dfrac{1}{n}\displaystyle\sum_{i=1}^n X_i \to_p \mu$.*

Proof. By linearity,

$$E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{1}{n} \cdot n\mu = \mu$$

and by uncorrelatedness,

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \leq \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Thus for any $\varepsilon > 0$, Chebychev gives

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)}{\varepsilon^2} \leq \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty. \qquad \square$$

The reason for the "weak" in the WLLN is that convergence in probability is a relatively weak mode of convergence.

Definition. We say that a sequence of random variables $X_1, X_2, ...$ *converges almost surely* to X if for all $\varepsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| > \varepsilon\right) = 0.$$

In this case, we write $X_n \to X$ a.s.

Thus almost sure convergence means that there is an $E \subseteq \Omega$ with $P(E) = 1$ such that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all $\omega \in E$. (With probability 1, X_n converges pointwise to X.)

If $X_n \to X$ a.s., then $P(|X_n - X| > \varepsilon \text{ a finite number of times}) = 1$, but this is not necessarily true if $X_n \to_p X$. For convergence in probability, the outcomes which are mapped near their image under X may vary significantly for different values of n.

Example 23.1. Let P be the uniform distribution on $(0, 1)$ and define

$$X_1 = 1_{(0,\frac{1}{2})},\ X_2 = 1_{[\frac{1}{2},1)}\ X_3 = 1_{[0,\frac{1}{4})},\ X_4 = 1_{[\frac{1}{4},\frac{1}{2})}, ...,\ X_{2^n+k} = 1_{(\frac{k}{2^n},\frac{k+1}{2^n}]}, ...$$

Then $X_n \to_p 0$ since for any $\varepsilon \in (0, 1)$,

$$P(|X_{2^n+k}| > \varepsilon) = P\left(\left(\frac{k}{2^n}, \frac{k+1}{2^n}\right]\right) = \frac{1}{2^n} \to 0.$$

However, $X_n(\omega)$ does not converge to 0 (or anything else) for any $\omega \in (0, 1)$ since there are infinitely many values of n with $X_n(\omega) = 1$ and infinitely many with $X_n(\omega) = 0$.

Conversely, it is not hard to show that almost sure convergence implies convergence in probability.

Indeed, given any $\varepsilon > 0$, let $A_n(\varepsilon) = \bigcup_{m\geq n} \{|X_m - X| > \varepsilon\}$. Then $A_n(\varepsilon) \searrow A(\varepsilon) = \bigcap_{n=1}^{\infty} A_n(\varepsilon)$.

Since $X_n(\omega) \nrightarrow X$ for all $\omega \in A(\varepsilon)$, $X_n \to X$ a.s. implies that $P(A) = 0$.

Thus, by continuity from above, we have

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) \leq \lim_{n\to\infty} P(A_n(\varepsilon)) = P(A(\varepsilon)) = 0,$$

so almost sure convergence is strictly stronger than convergence in probability.

The *strong law of large numbers* is an upgrade of the WLLN to almost sure convergence.

Theorem 23.4. *Suppose that $X_1, X_2, ...$ are i.i.d. with $E|X_1| < \infty$. Then $\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_1]$ a.s.*

We will prove the SLLN under the additional assumption that the X_i's have finite fourth moment.

To do so, we observe that $X_n \to c$ a.s. if and only if for every $\varepsilon > 0$, $P(|X_n - c| > \varepsilon \text{ i.o.}) = 0$.

This is because $a_n \to a$ if and only if for every $\varepsilon > 0$, $|a_n - a| > \varepsilon$ for only finitely many n, and

$$\{|X_n - c| > \varepsilon \text{ i.o.}\} = \{|X_n - c| > \varepsilon \text{ for only finitely many } n\}^C.$$

When phrased in this language, we are reminded of the first Borel-Cantelli Lemma, which we proved as Proposition 6.4:

Proposition. *Let $A_1, A_2, \ldots$ be a sequence of events. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.*

Thus to prove that $X_n \to c$ a.s., it suffices to show that for every $\varepsilon > 0$, $\sum_{n=1}^{\infty} P(|X_n - c| > \varepsilon) < \infty$.

Theorem 23.5. *Suppose that $X_1, X_2, \ldots$ are i.i.d. with $E\left[X_1^4\right] < \infty$. Then $\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X_1]$ a.s.*

Proof. By taking $Y_i = X_i - E[X_1]$, we may suppose without loss of generality that $E[X_i] = 0$.

Writing $S_n = \sum_{i=1}^{n} X_i$, we have

$$E[S_n^4] = E\left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right)\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{l=1}^{n} X_l\right)\right] = \sum_{i,j,k,l} E\left[X_i X_j X_k X_l\right].$$

Since $E[X_i] = 0$, independence implies that all terms of the form $E[X_i X_j^3]$, $E[X_i X_j X_k^2]$, and $E[X_i X_j X_k X_l]$ with i, j, k, l distinct are zero (as the expectation of the product is the product of the expectations).

Thus we only need to consider the n terms of the form $E[X_i^4]$ and the $3n(n-1)$ terms of the form $E[X_i^2 X_j^2]$ (determined by the $\binom{n}{2}$ ways of picking the indices and the $\binom{4}{2}$ ways of picking which two of the four sums gave rise to the smaller index).

Now

$$E[X_i^2 X_j^2] = E[X_i^2]E[X_j^2] = E[X_i^2]^2 \leq E[X_i^4]$$

(since $0 \leq \mathrm{Var}(X_i^2) = E[X_i^4] - E[X_i^2]^2$), so we have

$$E[S_n^4] = \sum_{i,j,k,l} E\left[X_i X_j X_k X_l\right] = nE[X_1^4] + 3n(n-1)E[X_1^2 X_2^2] \leq nE[X_1^4] + 3n(n-1)E[X_1^4] = Cn^2$$

where $C = 3E[X_1^4] < \infty$.

Accordingly, Chebychev's inequality shows that

$$P\left(\left|\frac{1}{n}S_n\right| > \varepsilon\right) = P\left(S_n^4 > n^4\varepsilon^4\right) \leq \frac{E[S_n^4]}{n^4\varepsilon^4} \leq \frac{C}{\varepsilon^4 n^2},$$

so

$$\sum_{n=1}^{\infty} P\left(\left|\frac{1}{n}S_n\right| > \varepsilon\right) \leq \sum_{n=1}^{\infty} \frac{C}{n^2} < \infty$$

and the result follows from Borel-Cantelli.

□

Example 23.2. Suppose that we wished to compute an integral of the form $\int_a^b f(x)dx$ for an arbitrary function f. In general, there will not be a nice closed form solution, so one would have to use numerical approximations like the trapezoid rule. The strong law justifies an alternative approach known as *Monte Carlo integration*.

First one generates a "random" sample, $u_1, u_2, ..., u_n$, from the $U(0,1)$ distribution using MATLAB or some such program.

Then set $x_i = (b-a)u_i + a$ so that $x_1, x_2, ..., x_n$ is a random sample from $U(a,b)$.

Finally, let $y_i = f(x_i)$ and approximate the integral as $\frac{b-a}{n}\sum_{i=1}^{n} y_i$.

The reason this works is that if $X_1, ..., X_n$ are i.i.d. $U(a,b)$, then $f(X_1), ..., f(X_n)$ are i.i.d. with $E[f(X_i)] = \frac{1}{b-a}\int_a^b f(x)dx$.

The strong law of large numbers implies that $\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to E[f(X_1)]$ a.s., so

$$\int_a^b f(x)dx = (b-a)E[f(X_1)] \approx \frac{b-a}{n}\sum_{i=1}^{n} f(X_i)$$

for large n.

This technique is applicable for integrals over very arbitrary regions in any dimension provided that one can generate points uniformly over the domain of integration. Thus one often sees Monte Carlo methods in high dimensional integrals where there are fewer known numerical techniques.

24. MGFs and the CLT

We conclude the course with one of the crowning jewels of probability, the *central limit theorem*, which establishes the central role of the normal distribution in statistics.

Theorem 24.1. *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $\mu = E[X_1]$ and finite variance $\sigma^2 = \mathrm{Var}(X_1) \in (0, \infty)$. Then for every $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt.$$

In words, the cumulative distribution function of $R_n = \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \dfrac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}}$ converges pointwise to the standard normal c.d.f.
(We say that R_n *converges in distribution* to $Z \sim \mathcal{N}(0,1)$ and write $R_n \Rightarrow Z$.)

It is hard to overemphasize how amazing this result is. If you sum up a bunch of i.i.d. random variables from essentially any distribution, then you get a random variable whose distribution is approximately normal!

It is not possible to give an honest, self-contained proof of the CLT here, but we will give a sketch of the argument as an excuse to discuss moment generating functions.

Definition. The *moment generating function* of a random variable X is defined by $M_X(t) = E[e^{tX}]$ for all $t \in \mathbb{R}$ for which the expectation is finite.

One way to think about moment generating functions is that the Taylor series expansion for the exponential gives $e^{tX} = \sum_{k=0}^{\infty} \frac{t^k}{k!} X^k$ pointwise. Assuming that we can interchange expectation and summation (which requires Fubini/Tonelli conditions) gives $M_X(t) = E[e^{tX}] = \sum_{k=0}^{\infty} \mu_k \frac{t^k}{k!}$ where $\mu_k = E[X^k]$ is the k^{th} moment of X.

In a similar spirit, formally differentiating $M_X(t) = E[e^{tX}]$ (which requires DCT conditions) gives $M_X'(t) = E[Xe^{tX}]$, $M_X''(t) = E[X^2 e^{tX}]$, and, in general, $M_X^{(k)}(t) = E[X^k e^{tX}]$. Accordingly, $M_X^{(k)}(0) = E[X^k] = \mu_k$.

Example 24.1. Suppose that $X \sim \mathrm{Binom}(n, p)$. Then

$$M_X(t) = E[e^{tX}] = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{kt} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} e^{kt} = \left[1 - p + pe^t\right]^n.$$

Note that $M_X'(t) = n\left[1 - p + pe^t\right]^{n-1} pe^t$, so $E[X] = M_X'(0) = n\left[1 - p + p\right]^{n-1} p = np$.
Higher moments can be computed similarly.

Example 24.2. Suppose that $X \sim \mathrm{Poisson}(\lambda)$. Then

$$M_X(t) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} e^{kt} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

Example 24.3. If $X \sim \text{Exp}(\lambda)$, then

$$M_X(t) = E[e^{tX}] = \int_0^\infty \lambda e^{-\lambda x} e^{tx} dx = \lambda \int_0^\infty e^{(t-\lambda)x} dx = \left. \frac{\lambda}{t-\lambda} e^{(t-\lambda)x} \right|_{x=0}^\infty .$$

Thus $M_X(t)$ is only defined for $t < \lambda$, in which case $M_X(t) = \frac{\lambda}{\lambda - t}$.

Example 24.4. If $X \sim \mathcal{N}(0,1)$, then

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{x^2}{2}} e^{tx} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}(x^2 - 2tx)} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}[(x-t)^2 - t^2]} dx = e^{\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{(x-t)^2}{2}} dx = e^{\frac{t^2}{2}} \end{aligned}$$

(where the final equality is because the $\mathcal{N}(t,1)$ density integrates to 1).

In order to generalize the last example to arbitrary μ and σ^2, we observe

Proposition 24.1. *If X is a random variable with mgf $M_X(t)$ and $a, b \in \mathbb{R}$, then $aX + b$ has mgf $M_{aX+b}(t) = e^{bt} M_X(at)$.*

Proof.

$$M_{aX+b}(t) = E[e^{t(aX+b)}] = E[e^{t(aX)} e^{bt}] = e^{bt} E[e^{(at)X}] = e^{bt} M_X(at).$$

□

Example 24.5. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $X =_d \sigma Z + \mu$ with $Z \sim \mathcal{N}(0,1)$, so Proposition 24.1 and Example 24.4 show that

$$M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\frac{\sigma^2 t^2}{2} + \mu t}.$$

Another useful property of mgfs is

Proposition 24.2. *If $X_1, X_2, ..., X_n$ are independent with $M_{X_1}(t), ..., M_{X_n}(t)$ defined for all $t \in I$, then*

$$M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t)$$

for all $t \in I$.

Proof. For $t \in I$,

$$M_{\sum_{i=1}^n X_i}(t) = E[e^{t \sum_{i=1}^n X_i}] = E\left[\prod_{i=1}^n e^{tX_i}\right] = \prod_{i=1}^n E[e^{tX_i}] = \prod_{i=1}^n M_{X_i}(t)$$

since independence implies that the expectation of the product is the product of the expectations. □

One can show that (when they exist) mgfs uniquely determine distributions, so the above result often enables us to find the distribution of sums of independent random variables more efficiently than using convolution arguments.

It is perhaps worth observing that $M_X(t)$ is kind of like the Laplace transform of the distribution of X. To see this, suppose that X is nonnegative with density f.
Then $M_X(t) = \int_0^\infty e^{tx} f(x)dx$, whereas $\mathcal{L}(f)(s) = \int_0^\infty e^{-sx} f(x)dx$.

A major shortcoming of mgfs is that they do not always exist. For this reason, people usually work with *characteristic functions*, defined by $\varphi_X(t) = E[e^{itX}]$ (which is like a Fourier transform). Since $\left|e^{itx}\right| = 1$ for all x, t, ch.f.s always exist, and many of results involving mgfs (including the proof of the CLT) carry over pretty directly.

The general idea behind the central limit theorem is to prove that the MGFs (or ch.f.s) of $\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$ converge to the mgf (ch.f.) of a standard normal. The reason this works is the following "continuity theorem" (which is nontrivial and will not be proved here):

Theorem 24.2. *Suppose that $Y_1, Y_2, \ldots$ have mgfs $M_{Y_i}(t)$ and Y has mgf $M_Y(t)$ which is continuous at 0. If $M_{Y_n}(t) \to M_Y(t)$ for all t in a neighborhood of 0, then $Y_n \Rightarrow Y$.*

We will also need the following calculus lemma.

Lemma 24.1. *Suppose that $\{a_n\}_{n=1}^\infty$ is a sequence of real numbers with $a_n \to a$. Then*

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Proof. We first observe that

$$\log\left(\left(1 + \frac{a_n}{n}\right)^n\right) = n \log\left(1 + a_n n^{-1}\right) = a_n \frac{\log\left(1 + a_n n^{-1}\right)}{a_n n^{-1}}.$$

Since $a_n \to a$, we have $a_n n^{-1} \to 0$, so using

$$\lim_{x\to 0} \frac{\log(1+x)}{x} = \lim_{x\to 0} \frac{(1+x)^{-1}}{1} = 1$$

(by L'Hospital's rule), we see that

$$\lim_{n\to\infty} \log\left(\left(1 + \frac{a_n}{n}\right)^n\right) = \lim_{n\to\infty} a_n \frac{\log\left(1 + a_n n^{-1}\right)}{a_n n^{-1}} = a.$$

Since the exponential function is continuous, we have

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = \exp\left(\log\left(\left(1 + \frac{a_n}{n}\right)^n\right)\right) = e^a. \qquad \square$$

With these results in hand, we can sketch out a proof of a weakened form of the CLT:

Assume that $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$. Assume also that the mgf of the $X_i's$ exists in a neighborhood of 0.
Set $Y_i = \frac{X_i - \mu}{\sigma}$ (so the $Y_i's$ are independent with $E[Y_i] = 0$ and $\mathrm{Var}(Y_i) = 1$), and let $M_Y(t)$ denote the common mgf of the $Y_i's$.
Then $R_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$ has MGF $M_{R_n}(t) = M_Y\left(\frac{t}{\sqrt{n}}\right)^n$ by Proposition 24.2.

A second order Taylor expansion of M_Y about the origin gives

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = M_Y(0) + M_Y'(0)\frac{t}{\sqrt{n}} + \frac{1}{2}M_Y''(0)\frac{t^2}{n} + r_2\left(\frac{t}{\sqrt{n}}\right)$$

where the remainder term satisfies $\lim_{n\to\infty} \frac{r_2\left(\frac{t}{\sqrt{n}}\right)}{\left(\frac{t}{\sqrt{n}}\right)^2} = 0$ for any t in a sufficiently small neighborhood of 0.

Since $M_Y(0) = 1$, $M_Y'(0) = E[Y] = 0$, and $M_Y''(0) = E[Y^2] = \text{Var}(Y) = 1$, we have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Taking $a_n = \frac{t^2}{2} + no\left(\frac{t^2}{n}\right) = t^2\left(\frac{1}{2} + \frac{o\left(\frac{t^2}{n}\right)}{\frac{t^2}{n}}\right) \to \frac{t^2}{2}$ in Lemma 24.1 shows that

$$M_{R_n}(t) = M_Y\left(\frac{t}{\sqrt{n}}\right)^n = \left[1 + \frac{1}{n}\left(\frac{t^2}{2} + no\left(\frac{t^2}{n}\right)\right)\right]^n \to e^{\frac{t^2}{2}}.$$

Example 24.4 shows that $Z \sim \mathcal{N}(0,1)$ has mgf $M_Z(t) = e^{\frac{t^2}{2}}$, so it follows from Theorem 24.2 that $R_n \Rightarrow Z$.